Document Title: Exploring Soil Bacterial Communities for Forensic Applications: A Genomics Approach

Author: Bo U. Pietraszkiewicz

Document No.: 237984

Date Received: March 2012

Award Number: 2009-IJ-CX-0021

Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Exploring Soil Bacterial Communities for Forensic Applications: A Genomics Approach

Bo U. Pietraszkiewicz

B.S., Central Connecticut State University, 2003

M.S., University of Connecticut, 2009

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2010

Exploring Soil Bacterial Communities for Forensic Applications: A Genomics Approach

Bo U. Pietraszkiewicz, Ph.D.

University of Connecticut, 2010

Abstract

It is accepted that soil evidence can be used in forensic investigations, where bacteria in soil are used to generate DNA profiles. The research presented in this thesis investigates how soil can be best used for forensic applications. Although bacterial profiles can be generated using several molecular methods, terminal restriction fragment length polymorphism (T-RFLP) analysis has been used most frequently to produce forensically relevant profiles. The second chapter proposes an alternative to T-RFLP analysis: comprehensive restriction fragment length polymorphism analysis (C-RFLP). This alternate typing method utilizes high performance liquid chromatography (HPLC) to separate and visualize unlabeled DNA fragments. However, neither method readily allows forensic scientists to extrapolate which types of bacteria are present in the soil sample in question. Knowing the molecular identity of a peak in a profile (i.e. which bacterial group is responsible for the presence of observed peaks) provides an additional layer of potentially informative information. In chapter three, 454 high throughput sequencing was used to survey fourteen soil samples, cataloging the major and minor components to soil bacterial communities. From these extensive DNA libraries, five bacterial groups were selected as candidates for group-specific bacterial typing. The main goal of chapter four was to determine the forensic potential of using such targeted analysis. DNA from soils was amplified using group-specific primers, digested with a restriction enzyme, and resolved using HPLC. HPLC was used because of its potential shown in chapter two and also to demonstrate that fragments could be collected and identified by sequence. The data show that group-specific profiles can be generated and used for forensic comparison due to the sufficient genetic variability within groups tested. This suggests that targeted molecular analysis of bacteria has great potential as a forensic soil typing tool and should be explored further. Ultimately, research on group-specific typing will aid in the development of a multiplex kit to be used in crime labs nationwide.

APPROVAL PAGE

Doctor of Philosophy Dissertation

Exploring Soil Bacterial Communities for Forensic
Applications: A Genomics Approach

Presented by

Bo U. Pietraszkiewicz, B.S., M.S.

Major Advisor ____

Linda D. Strausbaugh

Associate Advisor ____

Rachel J. O’Neill

Associate Advisor ____

Joerg Graf

University of Connecticut

2010

Table of Contents

List of Figures

List of Tables

Chapter 1 – Introduction to the forensic analysis of soils

I. Nonhuman DNA evidence in forensic science

In the rapidly evolving field of forensic science, the discovery, application, and validation of new genetic techniques is crucial for forensic evidence to remain a powerful tool in the courtroom. Human DNA typing has influenced the forensic community greatly by acting as a catalyst for other forensic applications. The forensic community recognizes that human DNA typing by short tandem repeats (STRs) is a very powerful tool because of its strong foundations in science and statistics. A recent report by the National Academy of Sciences, as sited by *The New York Times* article, “Science Found Wanting in Nations Crime Labs”, stressed the need for crime laboratories to incorporate more science into the services they offer [Moore, 2009]. This is not a trivial task, as the human genome has been the focus of genetic and population studies for decades. The vast amount of information known about the human genome has contributed to the development of a widely-accepted, comprehensive DNA typing protocol. However, not all crime scenes will contain human DNA evidence. As a result, the forensic community has recognized that other types of nonhuman DNA evidence should be used when appropriate.

Plant, animal, and soil materials all contain diverse genetic information, and can potentially be included as valuable pieces of evidence in a forensic case [Halverson and Basten, 2005; Horswell *et al.*, 2002; Menotti-Raymond *et al.*, 1997; Miller Coyle *et al.*, 2001; Yoon *et al.*, 1993]. However, with human DNA acting as the catalyst for forensic science technology in the late 1980s, it comes as no surprise that most of the nation’s forensic laboratories are primarily set up for human DNA analysis. This is also true for financial reasons, as human DNA evidence is routinely submitted to labs for testing. While it is unreasonable to expect that nonhuman DNA evidence will someday surpass human DNA in terms of volume, the potential information lying within these samples is not trivial, especially in cases where human DNA evidence cannot be used or is not available.

Two examples of criminal cases where nonhuman DNA evidence was used are the Palo Verde murder case (Arizona) and a murder case involving a cat named, Snowball (Prince Edward Island) [Menotti-Raymond *et al.*, 1997; Yoon *et al.*, 1993]. In 1992, a woman's body was found in an Arizona desert. Next to the body was a Palo Verde tree. During the course of the investigation, police had questioned a man who they later discovered owned a truck that contained Palo Verde seed pods in the truck bed. These seed pods became a key piece of evidence which ultimately linked the suspect to the crime scene. Generating DNA profiles from the genetic information in these seeds pods was novel to forensic investigations. Scientists used a molecular method called randomly amplified polymorphic DNA (RAPD) to generate DNA profiles from the tree pods at the crime scene, the evidentiary pods from the truck, as well as control trees from the area. It was shown that there was enough genetic variation within the Palo Verde tree population to distinguish single trees. This powerful nonhuman DNA evidence was successfully used to link the suspect to the crime scene, and ultimately lead to a conviction.

In 1995, a woman's body was discovered on Prince Edward Island, Canada. Before the woman's body was actually found, a coat was located in the woods close to her home. This coat was stained with blood from the victim and contained an additional type of evidence – white cat hair. During the course of the investigation, police visited the home of her estranged husband and noticed he had a white cat, Snowball. Investigators DNA typed the cat hair found on the coat to see if they could link it back to Snowball. Using short tandem repeats, researchers from the Laboratory for Genomic Diversity at the National Cancer Institute in Frederick, Maryland generated a DNA profile from 10 feline loci. Snowball's DNA profile was compared to the individual cat hairs found on the coat and a match was concluded at all 10 loci. The likelihood of another cat being the source of the DNA profile was also determined with a small population study of local and non-local cats. This use of feline DNA evidence was the first of its kind in Canada as well as the United States, and has since provided a great example for its acceptance in court.

The success of human STR typing certainly had a positive impact on the use of STR typing for domestic felines and canines. During the mid-1990s, canine and feline population studies increased, where researchers were documenting not only STR allele frequencies in breeds but also mitochondrial DNA haplotypes [Halverson and Basten, 2005]. Commercially available typing kits also surfaced at this time, making the integration of animal DNA testing into the repertoire of forensic labs that much simpler. Unfortunately, all nonhuman DNA typing protocols are not this straightforward. Plant DNA typing, for example, can be approached in many different ways. As seen in the Palo Verde case, RAPD was used. Amplified fragment length polymorphism (AFLP) and STR analysis can also be used to generate DNA profiles from plants. Plants come in a variety of species, each with its own genomic content and extraction challenges.

Because of this, there may be one technique that works well with one type of plant but not with another. Specialized analysis, like marijuana typing using AFLP, has been identified as a useful forensic typing method [Miller Coyle *et al.*, 2001].

There is another potentially informative type of nonhuman DNA evidence that can be found at crime scenes. Soil evidence is very different from plant and animal evidence in the sense that both plant and animal DNA profiles are generated from the genome of one species. Even if there is a mixture of multiple pet hairs or leaves, these items can be separated. This is not the case with soil, where the most common way to generate a DNA profile from soil is from bacteria. Soil is probably one of the most diverse microbial ecosystems on the planet [Torsvik and Ovreas, 2002]. Their abundance and diversity make bacteria an excellent molecular target for soil analysis. However, the early use of soil in forensic investigations did not include any molecular typing methods.

II. The progression of soil analysis methods

In 1935, the Unites States Federal Bureau of Investigation began analyzing soil samples based on physical properties [Finley *et al.*, 2004; Morgan and Bull, 2007]. Physical classification of soil based on color, mineral composition, and texture can provide valuable points of comparison due to wide variation in each of these classifications. Soil color is often determined by comparing dried sample to a color reference, most frequently the Munsell color chart. Mineral composition classification is a valuable characteristic as well. Most soils contain a combination of organic materials and minerals. However, the percent compositions and types of minerals differ from soil to soil. Soil particle size and texture can also be used to physically classify soils. Particle size is determined by sieving the soil through a matrix and thereby classified as: sand, silt or clay. When mineral composition and particle size information are combined, a more detailed soil textural designation is achieved. It is important to note that two soils can have the same textural classification without having the same bacterial populations [Tate *et al.*, 2000]. Before the use of molecular typing methods for bacteria in soil this was not known. If forensic soil analysis was needed in the 1950s, for example, two samples would have to be compared using physical classification. A study published by Sugita and Marumo in 1996 suggested that color classification could be used to forensically differentiate soils. However, when combined with other classification techniques, the power of discrimination will increase [Miller Coyle *et al.*, 2008; Sugita and Marumo, 1996].

When analyzing physical characteristics of soil, the potential for subjective interpretation must be recognized. Determining the color of a soil sample based on comparison to a chart can be prone to error, especially if the analyst is a novice. If soil classification must be done, it would be wise to have the data interpreted by more than one person. An ideal situation for a crime lab would be to have a soil expert on hand. However, this is an unlikely scenario given the financial burden to maintain such a position. For these and other reasons, forensic soil analysis has evolved to take a molecular approach. Adopting a DNA typing test makes use of any DNA scientists who are already on staff, especially if final methods are similar to human typing.

III. Molecular methods for the DNA typing of soil

Many fields of forensic science are built upon strong foundations in biology, chemistry and physics principles. The forensic analysis of soil is no exception. Successful strategies of soil analysis were based on microbiology and molecular biology research. Advancements beyond physical classification began with culture-dependent techniques. A soil sample was suspended in buffer solution and spread onto various agar plates. Different nutrients would be used to selectively grow certain bacteria where the presence and absence of growth identifies the bacteria found in the soil sample. This technique had its limitations, specifically with contamination. Later, scientists discovered that culturable organisms only comprised 1% of the total bacteria present [Kirk *et al.*, 2004]. From a forensics perspective, missing information about 99% of any population inhibits the ability to accurately compare samples. Soil samples are now routinely profiled using PCR-based, culture-independent molecular techniques targeting bacteria, allowing for a more objective analysis.

In-depth molecular analysis of bacterial communities in soil first requires an extraction technique that efficiently removes DNA from the soil matrix. Currently, extraction protocols and kits are available that utilize hot detergent lysis and/or mechanical bead beating [Martin-Laurent *et al.*, 2001; Roose-Amsaleg *et al.*, 2001; Yeates *et al.*, 1997]. Depending on the extraction technique used, as well as the amount of starting material, DNA quantity and quality will vary [Feinstein *et al.*, 2009; Martin-Laurent *et al.*, 2001; Roose-Amsaleg *et al.*, 2001]. For example, an extraction protocol that does not efficiently break gram positive cells will not produce nucleic acids from those cells. Conversely, if an extraction procedure is too aggressive on the cells, the DNA will be sheared. Also, the amount of starting material can affect final DNA yield. Most commercially available kits are used in conjunction with table top microcentrifuges, limiting the maximum amount of starting material to approximately 2 grams. Non-kit based extraction methods like the one published by Yeates *et al.* accommodates up to 100 grams [Yeates *et al.*, 1997]. Each extraction protocol has strengths and weaknesses, and the availability of protocols for both large and small starting amounts is valuable to forensics.

Another common problem with DNA extraction from soil is the co-extraction of humic substances. Humic acid, fulvic acid and humin are humic substances normally found in soil. These compounds accumulate in soil because of plant and animal decomposition [Zipper *et al.*, 2003]. It has been reported that as little as 1 nanogram of humic substances can inhibit PCR amplification [Menking *et al.*, 1999]. PCR inhibition is caused by the large molecule’s affinity for ionic substances, which in a PCR reaction leads to magnesium being sequestered from *Taq* polymerase [Roose-Amsaleg *et al.*, 2001; Zipper *et al.*, 2003]. The amount of humic substances vary in soil so not all nucleic acid extracts will contain the same amount of humic contamination.

There are ways to minimize the impact of contamination with PCR inhibitors. One would be to dilute the extraction stock so that the inhibitor is also diluted [Roose-Amsaleg *et al.*, 2001]. Also, there are reagents that can be added to the PCR reaction, like bovine serum albumin, to sequester humic substances. GeneReleaser ™ is a commercially available product that sequesters PCR inhibitors as well [Yeates *et al.*, 1997]. If contamination is very high, these simple measures may not be enough to minimize inhibition. Purification protocols are available to reduce the amount of contaminants in the stock extraction, including the use of cesium chloride density gradient ultracentrifugation, chromatography separation, or gel electrophoresis [Roose-Amsaleg *et al.*, 2001]. It is unreasonable to expect that purification protocols will remove all inhibitors. However, a combination of any of these procedures should help to generate an efficient PCR amplification.

To create a DNA profile from bacteria in soil, a universal genetic target is most often chosen for PCR amplification. The bacterial ribosomal operon is a region of the bacterial genome that is used for molecular analysis; specifically, the 16S ribosomal RNA (rRNA) gene has been used most frequently. There are three genes within the ribosomal operon (5S, 16S and 23S); the 16S gene has been the focus of molecular studies because of its manageable size and informative content. The 16S gene is composed of conserved and hypervariable regions. There are nine differently sized variable regions spread throughout the gene. As bacteria evolved, mutations in hypervariable regions that were not detrimental to the production of the 16S ribosomal protein were maintained. The combination of these mutations taxonomically differentiate bacteria. Molecular methods take advantage of the conserved regions of the 16S gene using primers that anneal to them to produce genetically variable amplicons. These amplicons represent both culturable and non-culturable bacteria, and the genetic information present can be translated into a bacterial DNA profile.

There are a multitude of analysis methods that the forensic community can use to generate DNA profiles from soil, although none were specifically created with forensics in mind. Therefore, the forensic community must choose a method that best suits its specialized applications. Some of the PCR-based analysis techniques include denaturing gradient gel electrophoresis (DGGE), temperature gradient gel electrophoresis (TGGE), and terminal restriction fragment length polymorphism (T-RFLP) analysis [Hill *et al.*, 2008; Janssen *et al.*, 2006; Muyzer *et al.*, 1993; Lerner *et al.*, 2006; Liu *et al.*, 1997]. Both DGGE and TGGE utilize either chemical or temperature gradients, respectively, within a polyacrylamide gel to separate PCR amplicons based on sequence. There is a direct correlation with low denaturing speed and high G+C content; amplicons that contain a higher G+C content will denature last among amplicons moving the slowest on the gel. Results from this separation are often faint and fuzzy, causing interpretation to be subjective and time consuming. While this technique is widely used by microbiologists, transition into a forensics lab is not ideal primarily because the equipment needed to run these experiments is not normally found in standard crime labs [Miller Coyle *et al.*, 2008]. But lack of equipment does not mean it cannot be useful. A study published by Lerner *et al.* (2006) explored the use of DGGE to type soil samples collected during a murder investigation. Although DGGE is not likely to be a routine analysis in crime labs, it is important to know that there are methods capable of forensically differentiating soils.

The T-RFLP method, first introduced by Liu *et al.* in 1997, has been accepted as a quick and reliable method for generating bacterial profiles. T-RFLP analysis begins with PCR amplification of bacterial DNA from the extracted soil sample. Universal primers tagged with a fluorophore on the terminal end target a specific region of the bacterial 16S gene generating heterogeneous amplicons each containing a fluorophore tag. It is also possible to use two different fluorophores on either end of the amplicon. Next, a restriction enzyme is chosen to digest the amplicons, producing fragments of DNA that vary in length. Only the labeled terminal ends are visualized on a DNA sequencing platform, with resolution of fragments based on length polymorphisms. The result is an electropherogram that depicts the length variants as peaks. This analysis method is especially promising for forensics because the DNA fragments are separated on instrumentation that most crime labs already have.

Although T-RFLP continues to be the most widely used technique because of its accuracy and reproducibility, it can be affected by biases introduced during PCR. In general, all PCR-based analysis techniques are affected in some way by primer design, extraction method, the *Taq* polymerase used for amplification, and the number of cycles in the PCR reaction [Egert and Friedrich, 2003; Martin-Laurent *et al.*, 2001; Suzuki and Giovannoni, 1996; Wintzingerode *et al.*, 1997]. PCR-based analysis methods are also influenced by the composition of bacterial genomes. Different species of bacteria have genomes that contain different copy numbers of the 16S gene [Farrelley *et al.*, 1995; Klappenbach *et al.*, 2000]. A bacterial species that contains 14 copies of the ribosomal operon will be amplified more efficiently than a species that only has 1 copy in its genome. This ultimately will lead to a biased ratio of PCR products towards species with more operon copies, even though there may be an equal amount of total cells. A fundamental understanding of each of these inherent biases allows researchers to modify extraction and amplification protocols to minimize most biases. While it is unreasonable to expect a complete suppression of bias, in order for bacterial community analysis to carry any validity a general acceptance of these biases operating uniformly is needed [Martin-Laurent *et al.*, 2001]. Any forensic DNA typing protocol must outline the exact steps and reagents needed for nucleic acid extraction and PCR amplification to ensure reproducibility and consistency.

Ecological and environmental biology research has provided the framework for the successful application of techniques, like T-RFLP, to forensic soil analysis [Heath and Saunders, 2006; Horswell *et al.*, 2002]. In 2002, Horswell *et al.* demonstrated that DNA profiles could be generated from soil using T-RFLP, and used to differentiate soil samples. Although these results have great potential, forensic scientists still need to consider the potential for other genetic targets, different analysis methods, and the impact that environmental variables have on the meaning of a match. Exploring alternatives may discover cost-efficient, quicker methods that are more amenable to forensic applications.

IV. Research synopsis

The research presented herein casts a wide net around basic soil diversity measures pertaining to how soil can be best used for forensic applications. Exploring forensically relevant questions required the use of techniques and equipment that are not intergraded into most crime labs, like nucleic acid HPLC and 454 pyrosequencing. Many of the experiments presented adopt a proof-of-principle approach, demonstrating that soil analysis can be feasible using a variety of methods.

Sophisticated methods that extract as much information on bacterial communities as possible will better inform us of what makes soil samples the same or different. For soil analysis to have any forensic feasibility, we must be able to demonstrate the possibility to differentiate many different soils types. This was the goal of the first set of experiments, where a novel typing method was developed and compared to the established soil typing method, T-RFLP. One of the limitations to forensic soil analysis is the lack of standardized match criteria, so experiments were designed as a first pass at establishing them. While these experiments were successful, the complexity of the results from universal bacterial typing suggested that this approach was not ideal for forensic use, leading us to ask the question whether less generic typing schemes would offer improvement.

Next generation 454 pyrosequencing (chapter 3) was used to build in-depth surveys on soil communities to provide rationale for group-specific analysis. By uncovering the native diversity in several soil samples, similarities and differences among soils could be more accurately assessed. 454 data cataloged an immense amount of inter- and intra-bacterial group diversity, leading to and providing rationale for the identification of several potential group-specific targets.

The last set of experiments (chapter 4) also took a proof-of-principle approach and resulted in the design and pilot application of group-specific assays to differentiate soil. Many of the forensically relevant questions addressed in chapter 2 were revisited, exploring how geography, ecosystem, time and meteorological events impact forensic soil analysis.

The data presented herein offers a broad first pass view into the realm of forensic soil analysis. This broad approach allowed for many questions to be addressed, with the results prompting focus on the next set of more narrow questions. The research presented helped shape the way to think about forensic soil analysis. The long-term goals of this extensive basic research are to establish feasibility and parameters for forensic applications and to ultimately aid in developing forensic kits that are both comprehensive and widely accepted.

Chapter 2 – Assessing the potential of a novel bacterial typing method in the forensic analysis of soils.

I. Introduction

Forensic science has played a critical role in civil and criminal investigations for decades. Throughout this period, advancements in scientific technology have allowed investigators to not only broaden the scope of what is forensically relevant evidence, but also provide greater scientific support for that evidence in court. During this time, soil became recognized for its potential value in forensic investigations [Heath and Saunders, 2006; Horswell *et al.*, 2002; Lerner *et al.*, 2006]. Given the various living components of soil ecosystems, a genetic profile of soil can be generated using different organisms as molecular targets [Bridge and Spooner, 2001; Hill *et al.*, 2008; Yeates *et al.*, 2003]. However, bacteria are used most often because of their high quantity and rich diversity in soil [Hill *et al.*, 2008; Torsvik and Ovreas, 2002]. By representing the total genetic diversity of bacterial communities in a DNA profile, soil samples can be objectively compared. Soil can be valuable to forensic investigations in two ways. First, it can serve as associative evidence that links a reference sample to an evidentiary sample. Second, soil evidence may provide investigative leads in cases where reference samples cannot be collected because crime scene locations are unknown.

Implementation of PCR-based methods to generate bacterial DNA profiles from soil allows for objective analysis of potentially highly informative forensic evidence. With thousands of different species of bacteria estimated to be found in one gram of soil, the goal of forensic soil analysis is to use a profiling method that is sensitive enough to detect differences in bacterial communities [Torsvik and Ovreas, 2002]. Such detection should allow investigators to accurately determine the relatedness of two samples without over-reaching interpretation. The current gold standard for generating forensically relevant bacterial profiles from soil includes the use of terminal restriction fragment length polymorphism analysis (T-RFLP) [Heath and Saunders, 2006; Hill *et al.*, 2008; Horswell *et al.*, 2002].

T-RFLP is an analytical technique that resolves flurophore-labeled DNA fragments created from a restriction enzyme digest of PCR amplicons. The 16S ribosomal RNA gene (16S rRNA) is often the target of PCR amplification, using universal bacterial specific primers to amplify variable regions of this gene. In 2002, Horswell *et al.* demonstrated potential for the use of T-RFLP as a way to generate forensically relevant bacterial profiles from soil [Horswell *et al.*, 2002]. In 2008, Meyers and Foran characterized some environmental challenges associated with this typing method [Meyers and Foran, 2008]. Although this approach has been proven valuable by both studies, it is important to investigate additional methods for DNA fragment visualization and separation. New methods may prove more amenable to forensic applications, particularly with respect to reproducibility, resolution, and cost. The data presented in this chapter investigates the use of high performance liquid chromatography as a means to resolve and analyze digested DNA fragments.

Denaturing HPLC (DHPLC) analysis has previously been used to study microbial communities in the human intestine [Goldenberg *et al.*, 2007] and marine samples [Barlaan *et al.*, 2005]. It has been utilized to track microbial infections in humans [Domann *et al.*, 2003], as well as identify specific bacterial species [Hurtle *et al.*, 2002].

Specialized HPLC systems (such as the Transgenomic WAVE® Nucleic Acid Detection system) are designed to separate DNA fragments by length by elution from a DNASep™ column. Samples can be analyzed under denaturing or non-denaturing conditions. For this research, HPLC soil analysis begins by universally amplifying bacterial 16S ribosomal DNA, followed by restriction enzyme digestion of the products. The subsequent, comprehensive pools of DNA fragments are separated by HPLC, detected by ultraviolet light absorption at 260nm, and are represented by peaks in a resulting chromatogram. Therefore, a DNA profile from soil can be easily generated without the use of a fluorophore, an important advantage given the high cost of purchasing fluorophore-labeled primers. The HPLC chromatogram reflects the genetic variability among soil bacterial communities. The data output from HPLC software is easy to read, highly reproducible, and automatically generates several peak attributes. A desirable feature of the Transgenomic WAVE® system is that individual fragments can be collected and subjected to post-run analysis, such as DNA sequencing. The ability to further characterize peaks by sequence can provide additional layers of discrimination not easily accomplished with standard T-RFLP analysis. This feature will be discussed in chapter four.

The data presented in this chapter centers around the introduction of a novel bacterial soil profiling method called comprehensive restriction fragment length polymorphism analysis (C-RFLP). Through implementation of HPLC, we have developed an alternative way to represent the genetic variability of bacterial communities in soil. The variability is easily translated into a DNA profile that has been used to compare soil samples in this study. Additionally, C-RFLP has been compared to T-RFLP analysis to determine which method shows the most forensic potential.

II. Results

II.a. Design of sample collection

This research utilizes a set of soil samples designed to represent bacterial communities from both presumed similar and radically different ecosystem. The set allows for three major classifications to be studied: (1) soils that share a general ecosystem and local geography, (2) predicted radically different soil ecosystems, and (3) soils that only share a common ecosystem ("biological replicates"). All sampling locations visited for this research are listed in Table 1.

Soil cores were collected from the first 2 inches beneath the horizon (excluding the freshwater sediment and sewage sludge samples). Five soil cores were taken from each sampling site. The site of the first core was chosen then the remaining four cores were taken two feet in each cardinal direction. Compact soil cores were placed into a plastic zip top bag then homogenized by hand.

Table 1 – Soil Sample Classification and Ecosystem Information

Soil Sample Name	Ecosystem	Collection Location
AG Farm	Agricultural Corn Plot	Mansfield-Storrs, CT
Swan Lake	Maintained Lawn adjacent to lake	Storrs, CT – Main Campus UConn
Mirror Lake	Maintained Lawn adjacent to lake	Storrs, CT – Main Campus UConn
Great Lawn	Maintained Lawn	Storrs, CT – Main Campus UConn
Cemetery	Maintained Lawn	Storrs, CT – Main Campus UConn
Field	Maintained Lawn	Middletown, CT
River	Freshwater River Sediment	Portland, CT
Sewage Sludge	2°Sewage Treatment Sludge	Middletown, CT
Lawn 1	Maintained Lawn	Wolcott Hill Park, West Hartford, CT
Lawn 3	Maintained Lawn	Batterson Park, New Britain, CT
Lawn 4	Maintained Lawn	Batterson Park, New Britain, CT
Lawn 5	Maintained Lawn	AW Stanley Park, New Britain, CT
Lawn 6	Maintained Lawn	AW Stanley Park, New Britain, CT
Lawn 7	Maintained Lawn	Stanley Quarter Park, New Britain, CT
Lawn 8	Maintained Lawn	Stanley Quarter Park, New Britain, CT
Lawn 9	Maintained Lawn	Falcons Field, New Britain, CT
Lawn 10	Maintained Lawn	Falcons Field, New Britain, CT
Lawn 11	Maintained Lawn	Walnut Hill Park, New Britain, CT
Lawn 12	Maintained Lawn	Walnut Hill Park, New Britain, CT
Lawn 13	Maintained Lawn	Martha Hart Park, New Britain, CT
Lawn 14	Maintained Lawn	Martha Hart Park, New Britain, CT
Lawn 16	Maintained Lawn	Washington Park, New Britain, CT
Lawn 17	Maintained Lawn	Skinner Road School, Ellington, CT
Lawn 19	Maintained Lawn	Windermere School, Ellington, CT

II.b. Validation of C-RFLP method

A novel way to generate DNA profiles representing bacterial communities in soil was created for this research, called C-RFLP. In C-RFLP analysis, the universal amplification of the bacterial 16S rRNA gene and restriction enzyme digestion of resulting amplicons are carried out using well-established molecular techniques. However, the use of HPLC to separate and detect DNA fragments for the generation of potentially forensically relevant DNA profiles from soil has not been previously described. To ensure that HPLC separation is reliable and reproducible, a set of validation experiments were done. The goal of validation was to ensure that soil samples could be profiled, and that data points collected from profiling the same soil sample multiple times were consistent for each analysis. Given the high sensitivity of the WAVE® system, it is expected that the DNA fragments will be precisely separated each time a soil sample is profiled.

To establish the reproducibility of profiles, DNA from 8 soil samples from the University of Connecticut (Storrs, CT) and Middletown, CT were each profiled on three separate instrument runs (technical replicates). For the validation trials, each technical replicate began with initial PCR amplification. On the WAVE® system, the smallest fragments in the sample are detected first, beginning approximately 5 minutes post-injection. All fragments are detected by ultraviolet light which allows for a constant measure of absorbance at 260nm over a run time of 28 minutes. As DNA fragments are detected, their identity is represented by a peak. Each peak is characterized by height (measured absorbance, millivolts), and the time that the fragment elutes from the column (retention time, minutes). The largest fragments in each sample are the last to elute from the column, approximately 26 minutes post-injection.

Figure 1 shows technical replicates from the AG Farm and Great Lawn. The chromatograms have not been cropped in order to illustrate the data from a complete 28 minute run. Peaks detected during the first 3-4 minutes of a run are attributed to excess primers from the PCR reaction, and are not informative in analysis. A side-by-side comparison of replicate profiles demonstrates that not only are the presence and absence of peaks reproduced, but the unique morphology of peaks is replicated. Each technical replicate was carried out individually, with all three trials (Trials A, B, and C) occurring on three separate days.

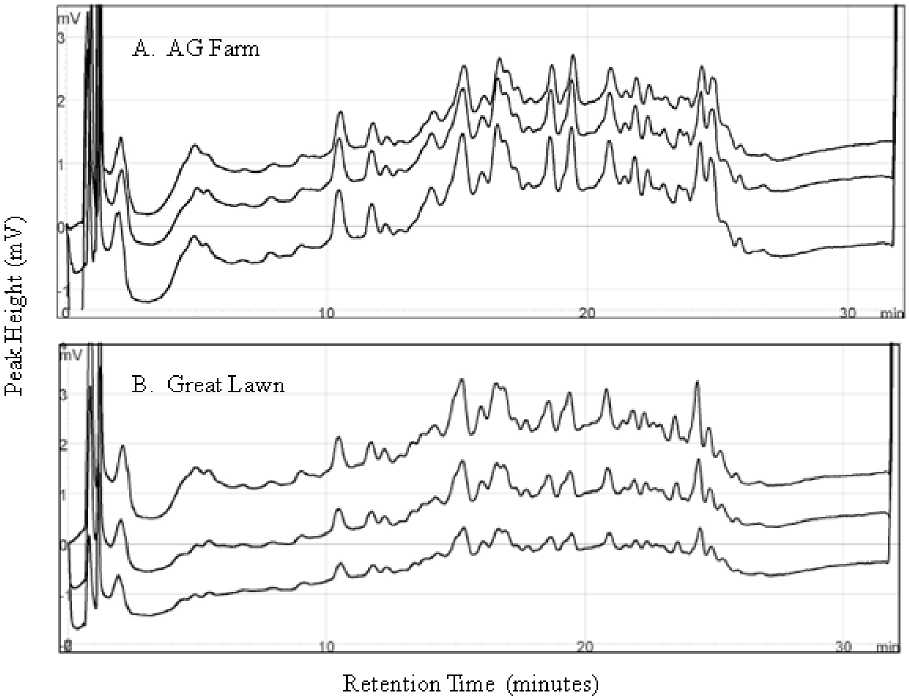


Figure 1 – Consolidated C-RFLP profiles of *AluI* digested 16S rRNA gene amplicons from soil bacterial community.

Profiles shown from soil collected from AG Farm (panel A), and soil from Great Lawn (panel B). Independently run traces exhibit high similarity, based on examination of peak patterns. DNA fragments unique to each sampling location begin elution off of the column beginning (approximately) 5 minutes into injection. Presence of DNA is represented as a peak in the chromatogram. Peak height is noted along the y-axis, reported in millivolts (mV). Retention time is noted along the x-axis, and is reported in minutes.

In order to better illustrate the reproducibility of each C-RFLP profile, chromatograms were cropped and profiles were expanded to focus on amplicon fragments generated from *AluI* digestion (Figure 2). Five of the highest peaks that span the full elution run in each profile were selected to objectively evaluate reproducibility of fragment elution times. The sensitivity of the WAVE® detection system allows for retention time to be reported to the thousandth of one minute, from which seconds can be calculated. Figure 2 shows individual technical replicates for the AG Farm (Trials A-C).

Peaks chosen for analysis are labeled 1-5. Table 2A lists the retention times in minutes for the 5 selected peaks. The results indicate that the select peaks (fragments) are eluting off the column at nearly identical times in each run. The variation in the retention times of the 5 replicate peaks in AG Farm range from 0.78 second (variation between trials A and B, peak 4 is 0.013 minute) to 4.8 seconds (variation between trials A and C, peak 1 is 0.08 minute). Analysis of all 8 soil samples used in this validation experiment showed that the greatest shift in peak retention was no greater than 6 seconds (or 0.1 minute rounded time; see Chapter 8, Figures S27A – S27H). Considering the entire run length, 6 seconds of a 28 minute run accounts for only 0.36% of total time. The minor shifts observed among replicate peaks are expected given the high sensitivity of this instrument. Fragment separation can be influenced by the number of injections that have run through the column, freshness of solutions A and B, as well as the purity (cleanliness) of the column. Although individual analysis of peaks demonstrates high reproducibility, it is important to determine whether these variables were affecting the elution of all fragments equally. HPLC separation of DNA fragments would not be a reliable profiling method if all fragments contributing to a profile pattern are not equally affected by these variables.

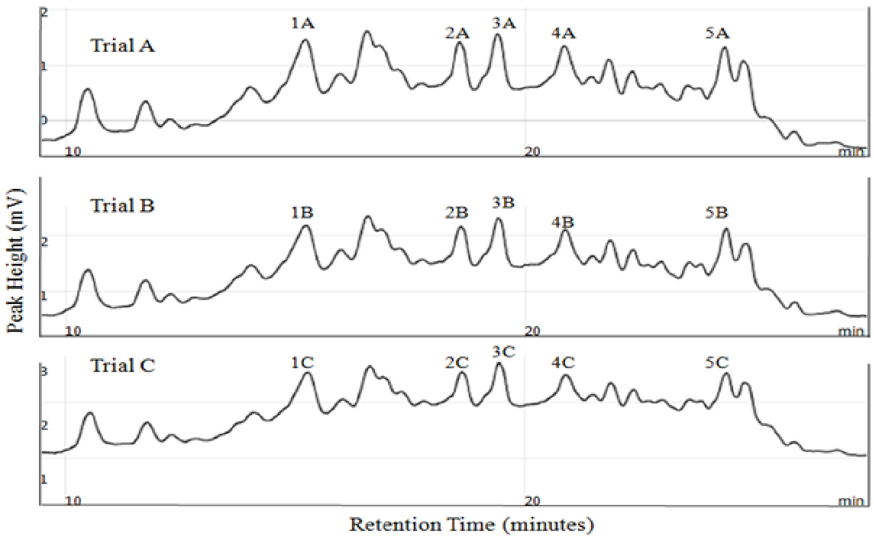
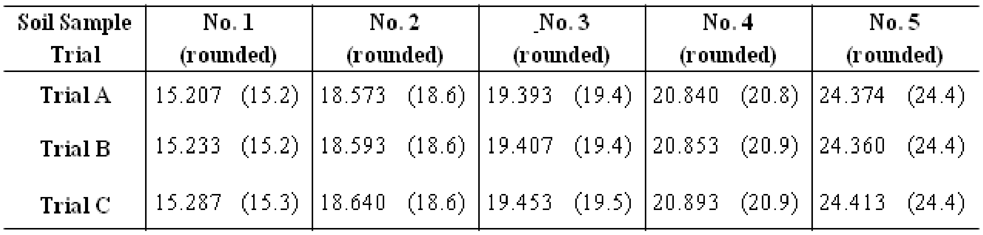


Figure 2 – Analysis of C-RFLP peak retention times from Agricultural Farm soil

Trials A, B, and C represent independently run replicate C-RFLP bacterial community profiles from Agricultural Farm soil. Profiles have been cropped to show all fragments eluted between (approximately) 10 and 28 minutes. Five of the largest peaks were selected, and their respective retention times were compared for reproducibility (Tables 2A and 2B). Note: Y-axis scales (Peak Height – mV) vary between each trial.

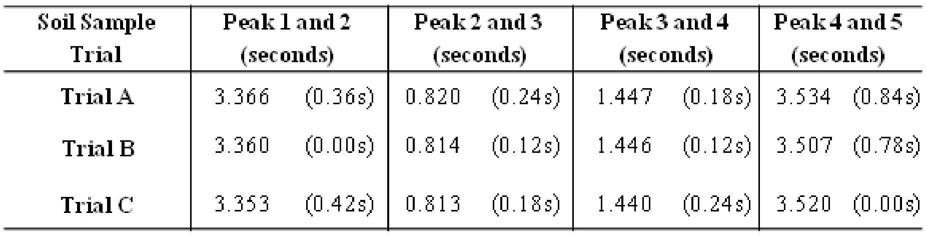
Table 2A – Peak retention times for individual fragments: Fig. 2



Soil Sample Trial	No. 1 (rounded)	No. 2 (rounded)	_No. 3 (rounded)	No. 4 (rounded)	No. 5 (rounded)
Trial A	15.207 (15.2)	18.573 (18.6)	19.393 (19.4)	20.840 (20.8)	24.374 (24.4)
Trial B	15.233 (15.2)	18.593 (18.6)	19.407 (19.4)	20.853 (20.9)	24.360 (24.4)
Trial C	15.287 (15.3)	18.640 (18.6)	19.453 (19.5)	20.893 (20.9)	24.413 (24.4)

Retention times are reported in minutes, as determined by Navigator ® Software. Each peak's retention time has also been rounded to the nearest tenth of one minute (listed in parentheses).

Table 2B – Elapsed elution time between select fragments: Fig. 2



Soil Sample Trial	Peak 1 and 2 (seconds)	Peak 2 and 3 (seconds)	Peak 3 and 4 (seconds)	Peak 4 and 5 (seconds)
Trial A	3.366 (0.36s)	0.820 (0.24s)	1.447 (0.18s)	3.534 (0.84s)
Trial B	3.360 (0.00s)	0.814 (0.12s)	1.446 (0.12s)	3.507 (0.78s)
Trial C	3.353 (0.42s)	0.813 (0.18s)	1.440 (0.24s)	3.520 (0.00s)

The calculated difference in time between peak elution within each trial run is given in seconds (listed in parentheses).

Any variation in injection conditions should affect all fragments equally, thus shifting the pattern as a whole. Table 2B compares the elapsed time (in minutes) between the elution of select fragments. The results show that the integrity of the AG Farm profile is virtually unaffected between replicates. Using the average elapsed time between fragments as a standard, the variation in time is calculated in seconds (shown in parentheses). The data for the AG Farm shows that the spacing between fragments varies no more than 0.84 seconds from the average time. The data for the remaining validation trials shows the same spacing calculations to be no greater than 2.82 seconds for all trials (Chapter 8, Figures S27A – S27H). Based on the validation experiments done in this study, we are confident that HPLC fragment analysis generates reliable data that is highly reproducible, providing an objective method for comparing bacterial communities in soil.

Bacterial C-RFLP and T-RFLP profiles from all soil/sediment samples were compared using a relatedness calculation. The Sorensen similarity index determines the percent relatedness of two samples based on the number of peaks shared between two samples. This index is calculated by the formula: 2(number of peaks shared between two profiles) / (the sum total of all peaks detected in both profiles) [Meyers and Foran, 2008]. The index has values from 0 (no similarity) to 1.0 (100% identical). When comparing any two C-RFLP profiles, we define shared peaks as such if retention times are within a range of +/- 6 seconds (0.1 minute) of one another. This definitive range was determined by reproducibility trials since known replicate peaks did not vary more than +/- 0.1 minute. For ease of interpretation, the retention time of peaks was rounded to the nearest tenth of one minute. A shared peak between any two T-RFLP profiles is defined as two fragments having exactly the same base pair length when the lengths are rounded to the nearest whole number. GeneScan ™ software assigns each fragment a size which includes tenths or hundredths of a base. For ease of interpretation, all computed fragment lengths were converted to whole numbers.

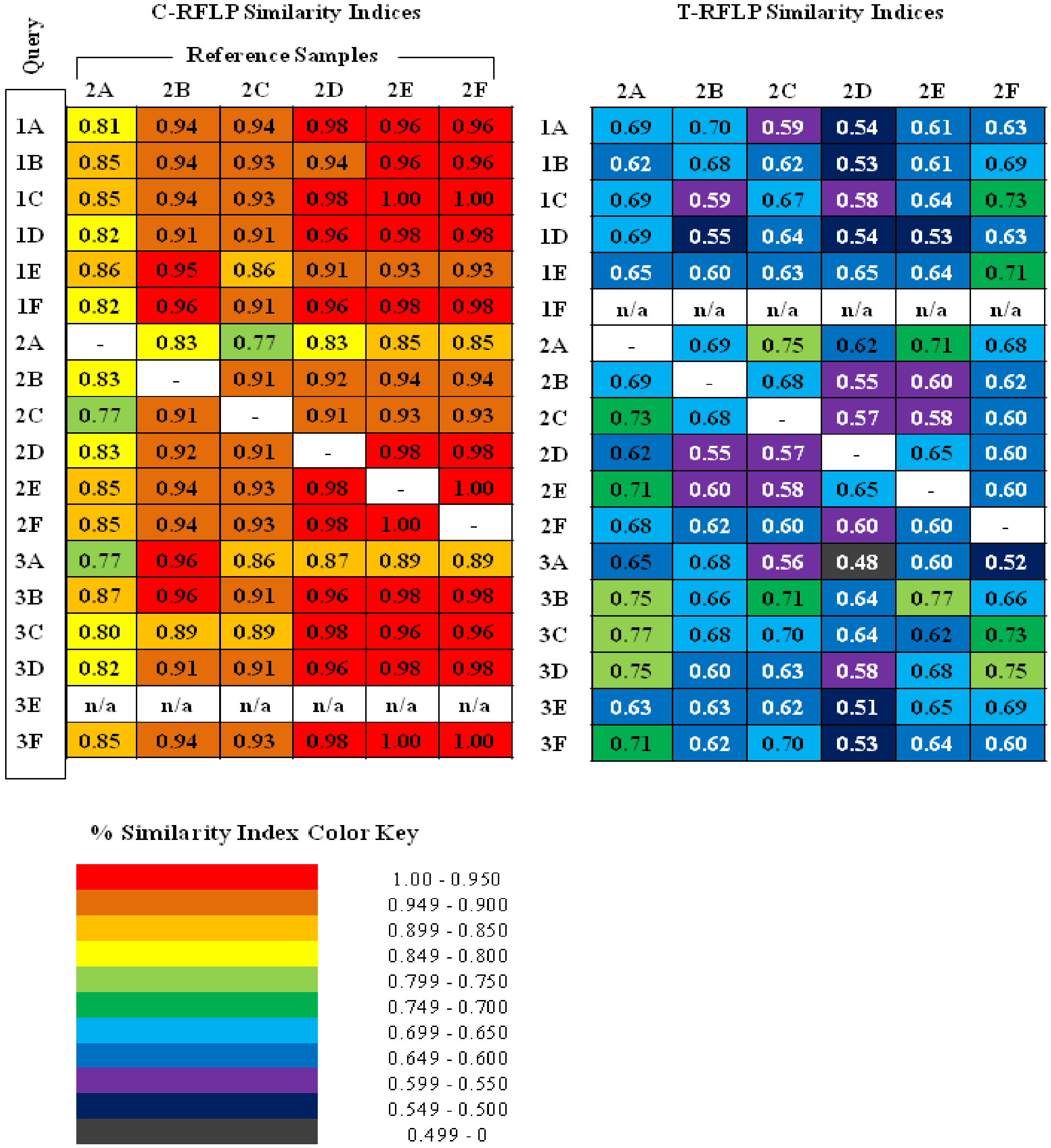
II.c. Grid Collection

A forensically relevant typing method should be able to successfully interpret soil samples as similar if they are collected from the same uninterrupted area. For example, multiple soil samples collected from a continuous section of maintained lawn should produce C-RFLP and T-RFLP profiles that show high relatedness between all of the samples. Unfortunately, one of the short comings to the use of soil as forensic evidence is lack of an established criterion for determining not only what a 'high' similarity is, but whether two soil samples are the same or different (match). Although bacterial communities in soil have been shown to be heterogeneously dispersed within a single area [Ettema and Wardle, 2002; Girvan *et al.*, 2003], a forensically relevant typing method must not be too sensitive as to falsely conclude that two known soil samples did not originate from one location. Moreover, the meaning of a sample match must also be addressed. For example, does a high similarity index between two samples always support the conclusion that two soils *definitely* came from the *same* location, or is it more appropriate to conclude that high similarity *only suggests* two samples *could have come from* the same location? By collecting multiple samples from an uninterrupted maintained lawn, two points will be addressed. First, which typing method generates data more closely resembling the data we expect to see from an uninterrupted lawn (i.e.

high similarity index values). Second, if it is practical to use the similarity index values to establish a criterion for sample matching ("match threshold"). Samples collected for grid profiling should have the highest similarity indices among all sample comparisons done in this study. If this is not the case, then the expectations for what soil evidence can tell us in a forensic context must be fine tuned.

A 75' (width) by 150' (length) portion of the Great Lawn was sectioned into three rows. Each row (1-3) contained 6 sampling locations (A-F). Each grid sample was comprised of 5 soil cores taken within a 4' diameter (central core with remaining cores taken 2' in each cardinal direction). Each grid was spaced 25' apart. At the time of collection, the Great Lawn's landscape contained thick grass, clover patches, and sandy areas where grass was not growing. Care was taken to ensure that grid samples were primarily taken from thick lawn areas.

C-RFLP and T-RFLP profiles were generated for each grid. Similarity indices were calculated for all samples, using each successfully profiled grid as a reference for all others so as to ensure outlier references were not chosen. Figure 3 depicts a side by side comparison of the results obtained from using each grid in row 2 as a reference sample. Row 2 data is representative of the results generated from rows 1 and 3 for both typing methods. Query samples are listed down the left side of each panel. Values have been color coded for ease of interpretation. Some soil grids were not able to be successfully profiled ("n/a").



C-RFLP Similarity Indices

Reference Samples

Query	2A	2B	2C	2D	2E	2F
1A	0.81	0.94	0.94	0.98	0.96	0.96
1B	0.85	0.94	0.93	0.94	0.96	0.96
1C	0.85	0.94	0.93	0.98	1.00	1.00
1D	0.82	0.91	0.91	0.96	0.98	0.98
1E	0.86	0.95	0.86	0.91	0.93	0.93
1F	0.82	0.96	0.91	0.96	0.98	0.98
2A	-	0.83	0.77	0.83	0.85	0.85
2B	0.83	-	0.91	0.92	0.94	0.94
2C	0.77	0.91	-	0.91	0.93	0.93
2D	0.83	0.92	0.91	-	0.98	0.98
2E	0.85	0.94	0.93	0.98	-	1.00
2F	0.85	0.94	0.93	0.98	1.00	-
3A	0.77	0.96	0.86	0.87	0.89	0.89
3B	0.87	0.96	0.91	0.96	0.98	0.98
3C	0.80	0.89	0.89	0.98	0.96	0.96
3D	0.82	0.91	0.91	0.96	0.98	0.98
3E	n/a	n/a	n/a	n/a	n/a	n/a
3F	0.85	0.94	0.93	0.98	1.00	1.00

T-RFLP Similarity Indices

	2A	2B	2C	2D	2E	2F
1A	0.69	0.70	0.59	0.54	0.61	0.63
1B	0.62	0.68	0.62	0.53	0.61	0.69
1C	0.69	0.59	0.67	0.58	0.64	0.73
1D	0.69	0.55	0.64	0.54	0.53	0.63
1E	0.65	0.60	0.63	0.65	0.64	0.71
1F	n/a	n/a	n/a	n/a	n/a	n/a
2A	-	0.69	0.75	0.62	0.71	0.68
2B	0.69	-	0.68	0.55	0.60	0.62
2C	0.73	0.68	-	0.57	0.58	0.60
2D	0.62	0.55	0.57	-	0.65	0.60
2E	0.71	0.60	0.58	0.65	-	0.60
2F	0.68	0.62	0.60	0.60	0.60	-
3A	0.65	0.68	0.56	0.48	0.60	0.52
3B	0.75	0.66	0.71	0.64	0.77	0.66
3C	0.77	0.68	0.70	0.64	0.62	0.73
3D	0.75	0.60	0.63	0.58	0.68	0.75
3E	0.63	0.63	0.62	0.51	0.65	0.69
3F	0.71	0.62	0.70	0.53	0.64	0.60

% Similarity Index Color Key

🟥	1.00 - 0.950
🟧	0.949 - 0.900
🟧	0.899 - 0.850
🟨	0.849 - 0.800
🟩	0.799 - 0.750
🟩	0.749 - 0.700
🟦	0.699 - 0.650
🟦	0.649 - 0.600
🟪	0.599 - 0.550
🟦	0.549 - 0.500
⬛	0.499 - 0

Figure 3 – C-RFLP and T-RFLP similarity index heat map for Great Lawn collection grid.

Similarity indices for respective profiles from each grid location on Great Lawn. Panel the left lists C-RFLP data; T-RFLP data shown on the right. Figure shows comparisons using each sample collected from row 2. Reference samples are listed across top of color grids. Query samples are listed down the left-hand side of each panel. Similarity values are color coded according to the ranges indicated in the color key. While only data from row 2 is shown, the similarity ranges shown are representative of all comparisons. "n/a" indicates a profile was not able to be generated.

Similarity indices of soil samples analyzed by C-RFLP show that this method is better suited for replicate testing. Based on the data shown in Figure 3, only C-RFLP profiling gives an expected measure of relatedness of multiple samples collected from a single ecosystem and geography. All samples considered, bacterial community relatedness fell within a range of 0.77 – 1.00 for C-RFLP, while T-RFLP indices fell within the range of 0.48 – 0.77 (Figure 4).

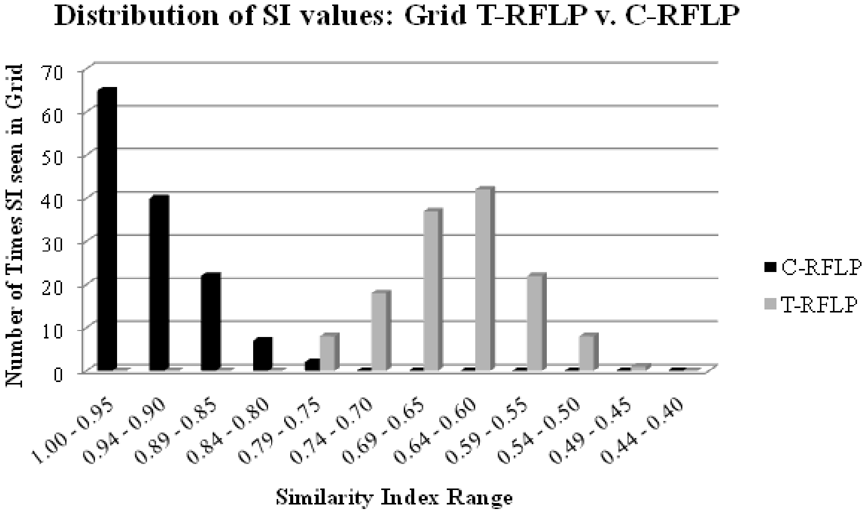


Figure 4 – Distribution of similarity index values in grid experiment: T-RFLP versus C-RFLP.

Distribution of similarity index values observed in grid analysis. Height of bar graph indicates how many times the corresponding similarity index value was seen in the grid. The average SI for T-RFLP analysis is 0.64. The average SI for C-RFLP analysis is 0.93.

A majority of values clustered around 0.95 – 1.00, and 0.60 – 0.69 for C-RFLP and T-RFLP, respectively. This distribution supports the C-RFLP profile for grid 2A as being an outlier (Figure 3). The C-RFLP profile for 2A was not as robust as the others, perhaps due to PCR inhibition or inefficient nucleic acid extraction. Taking into account all 136 comparisons, the average similarity index value within the grid sampling was 0.931 for

C-RFLP analysis; T-RFLP analysis produced an average similarity index of 0.639. Although grid 2A (C-RFLP) can be considered an outlier, to provide a conservative threshold this data has been included in the average. If the data were excluded, the average C-RFLP value rises to 0.944. From this grid data, it is proposed that a match threshold of at least 0.93 similarity be used to establish that two soil samples *likely came* from the same location when using the C-RFLP typing method. Additionally, it is proposed that a match threshold of 0.64 be used to establish the same relatedness of soil samples when using the T-RFLP typing method. The remaining samples collected in this research will be used to assess the reliability and accuracy of these thresholds on known, unrelated soils.

II.d. Assessment of C-RFLP in soil individualization

The data show that C-RFLP is a reliable method for DNA fragment separation, generating profiles that are reproducible and easily interpreted. The C-RFLP method has also performed well in grid analysis, consistently generating profiles of very high similarity from multiple samples taken from one area. Next, the ability of the C-RFLP method to differentiate soil samples from various locations and ecosystems is assessed.

In this research the term ‘ecosystem’ is defined as a biogeographical location that can be characterized by its natural environment. Ecosystems that share a specific type of vegetation, for example, may also share bacterial groups that prefer the nutrient conditions provided by that environment [Girvan *et al.*, 2003]. It can be hypothesized that the more ecologically diverse two ecosystems are, the more diverse the bacterial communities native to them will be. A forensically valuable method for soil profiling should successfully differentiate soil samples independent of the range of similarity between the bacterial communities. In order to thoroughly evaluate C-RFLP's discriminating potential, soils were examined belonging to three major classifications: (1) soils that share a general ecosystem and local geography, (2) predicted radically different soil ecosystems, and (3) "biological replicate" ecosystems. The T-RFLP method was also used to profile all samples.

Easily accessible locations around the University of Connecticut campus (Storrs, CT) were chosen to represent soils sharing a local geography within a common ecosystem. Soil from the agricultural farm was sampled from a corn plot maintained for research purposes. The Great Lawn is a maintained lawn between two buildings characterized by high foot traffic. Soils were collected from locations adjacent to small lakes on campus. The Swan Lake location is well-shaded and surrounded by plants and trees, while the sampling location at Mirror Lake is characterized by more open space and less vegetation. Swan Lake soil was collected 3 feet from water, while Mirror Lake soil was collected 30 feet from water. Soil collected from a cemetery on campus represented a lawn surrounded by trees (low foot traffic area).

Samples collected from the Middletown, CT area provided three ecologically diverse and radically different environmental samples: maintained lawn, freshwater sediment, and sewage treatment sludge. Soil collected from a maintained lawn ("Field" sample) bordered by trees near the Snow Elementary School was subject to moderate foot traffic. Secondary sewage sludge was obtained from a waste treatment plant. Sediment from the Connecticut River was obtained from the river's edge, approximately two inches below the water surface.

Samples to serve as "biological replicate" soils were collected from maintained lawns in 10 community parks/recreational areas located in Hartford and Tolland counties (samples named as "Lawn #" - Table 1). Maintained lawns are prevalent in these locations, providing an excellent option for studying soils that can be superficially classified as presumed biological replicate ecosystems. All sampling locations were confined to lawn areas that were 10 feet away from tree/shrub borders. Two soil samples were collected from each park, at opposite ends of the area.

The first sample set used to evaluate the ability of C-RFLP to differentiate soils are samples that represent very different ecosystems. The samples collected in the Middletown area each represent unique ecosystems (field soil, river sediment, and sewage sludge). Figure 5 shows the C-RFLP profiles for these samples. Calculated indices confirm low similarity between these three soils. The field and river share 27% of peaks (0.27); the field and sludge share 43% of peaks (0.43); the river and sludge share 42% of peaks (0.42). These data support our initial assumption that soils from radically different ecosystems are characterized by very different bacterial communities. Furthermore, the data show that HPLC was able to successfully detect these presumed differences and represent them in a chromatogram that provides unambiguous data points for analysis. Based on the match criteria set forth in the grid analysis section, all similarity index comparisons here fall well below the 0.93 match threshold, providing empirical support that these samples are not from the same location.

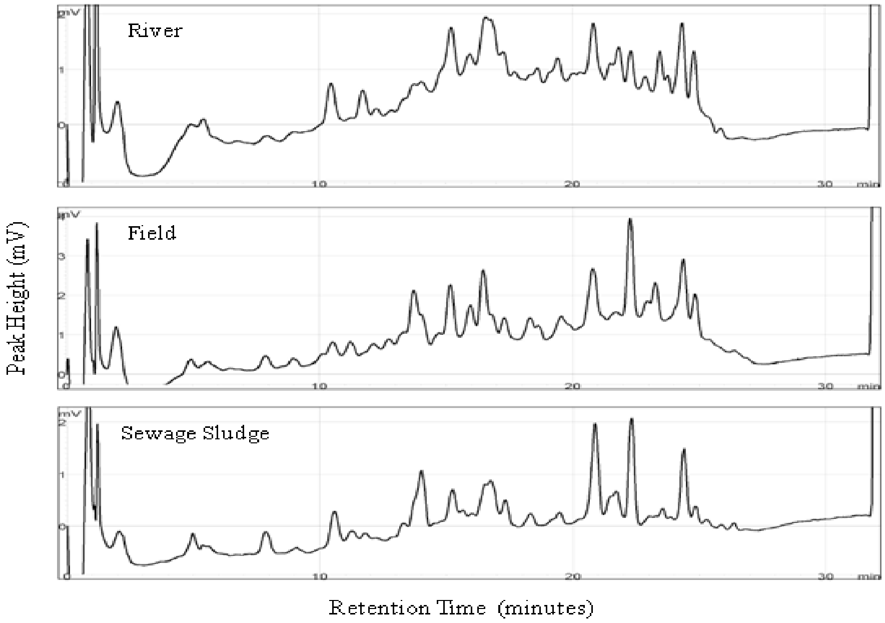


Figure 5 – Bacterial community C-RFLP profiles from three ecologically diverse sampling locations (Middletown, CT)

Bacterial community C-RFLP profiles generated from universally amplified, digested 16S rRNA amplicons. While all data from time 0 – 30 minutes is shown, digested bacterial DNA fragments are detected beginning approximately 8 minutes post-injection. Note: Y-axis scales (Peak Height – mV) vary between trials.

The next set used to evaluate the potential use of C-RFLP as a forensic typing method is soils that share a general ecosystem and local geography. Soil samples collected from the University of Connecticut are geographically localized, having been collected within an approximately 2 mile radius. It can be hypothesized that these profiles will have higher similarity indices (as compared to those in the first set) due to the proximity of sampling locations, as well as the presence of environmental characteristics common to these ecosystems (see Table 1) [Horner-Devine *et al.*, 2004]. Figure 6 aligns these C-RFLP profiles. The chromatograms in Figure 6 share similarities in peak distribution and morphology. This suggests that the bacterial communities native to each sampling site at the University of Connecticut may share some similarities in structure and composition.

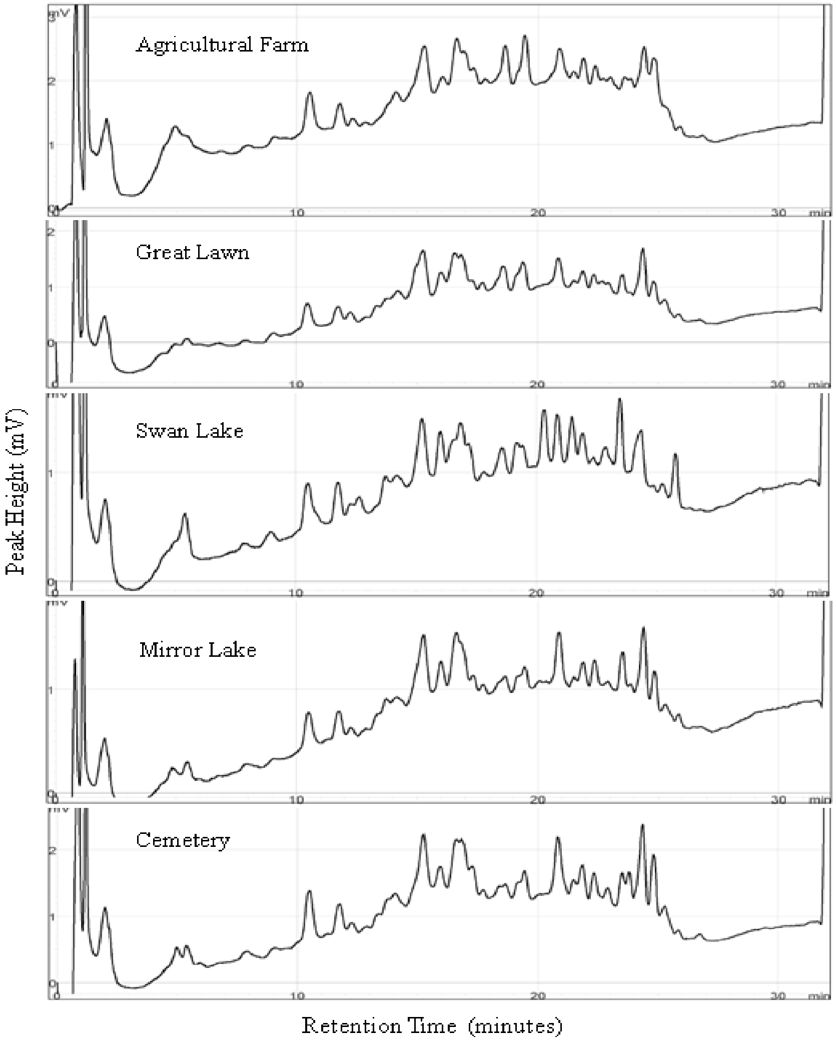


Figure 6 – Bacterial community C-RFLP profiles from geographically localized samples (University of Connecticut)

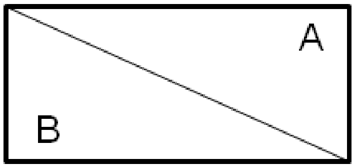
Bacterial community C-RFLP profiles generated from universally amplified, digested 16S rRNA amplicons. Samples represent soils that share a general ecosystem and local geography. While all data from time 0 – 30 minutes is shown, digested bacterial DNA fragments are detected beginning approximately 8 minutes post-injection. Note: Y-axis scales (Peak Height – mV) vary between trials.

Using the Great Lawn profile as the reference, the following similarity indices were calculated: AG Farm, 0.82; Swan Lake, 0.56; Mirror Lake, 0.69; Cemetery, 0.71. These similarity indices support the hypothesis that bacterial communities in soils collected from similar ecosystems will have higher similarity indices than those calculated from profiles generated from diverse ecosystem soils. The soil with the lowest similarity to Great Lawn was Swan Lake. Although Swan Lake was geographically localized to all the others, the soil at this location was unique. Swan Lake soil was collected 3 feet from water, and contained a noticeable amount of fibrous materials in addition to organic soil. All other samples were solely organic soils collected from maintained lawns. Given these characteristics, the bacterial community in the Swan Lake soil likely contained species fit for survival in this micro-environment that are not present in the other locations. With respect to the match threshold, all samples profiled in this second set also fell below 0.93, further validating this value as potential match criteria.

Lawn samples were analyzed as a group to determine the extent of bacterial community sharing between soils collected from presumed biological replicate locations. From a forensics perspective, it is important to determine whether soils from locations that all look the same (superficially) produce profiles that are distinguishable. It can be hypothesized that since these locations all have a single environmental ecosystem in common, there may be a set of bacterial groups that are indigenous to soils found in these lawns. As a result, bacterial community profiles may demonstrate higher than expected similarity indices when compared.

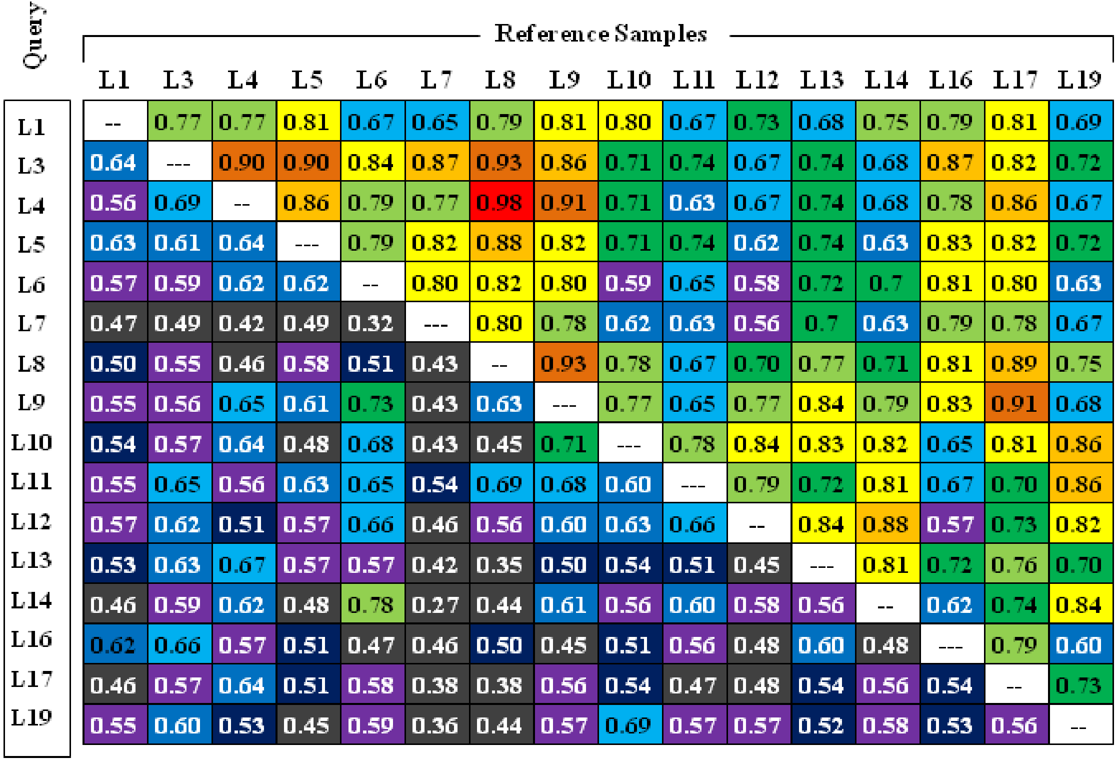
Figure 7 is a hybrid heat map of similarity indices for all C-RFLP and T-RFLP data generated for these Lawn samples (T-RFLP data is discussed in the next section).

The top right section (A) represents C-RFLP data. Each profile was compared against all others to ensure that outlier references were not chosen. The average similarity index for all Lawn comparisons is 0.758. Profiles generated from soils collected within the same park resulted in similarity indices that increased to an average of 0.810 [L3/4, 0.90; L5/6, 0.79; L7/8, 0.80; L9/10, 0.77; L11/12, 0.79; L13/14, 0.81]. As a general observation, the difference between 0.758 and 0.810 does not appear to be significant. Given the wide range of similarity values, C-RFLP profiles from these presumed biological replicate sites cannot be characterized by a specific percentage of relatedness. While these biological replicate sites are likely characterized by many of the same bacterial species, overall their soil profiles are distinguishable.



Section A – Similarity indices for C-RFLP

Section B – Similarity indices for T-RFLP



Reference Samples

Query	L1	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12	L13	L14	L16	L17	L19
L1	--	0.77	0.77	0.81	0.67	0.65	0.79	0.81	0.80	0.67	0.73	0.68	0.75	0.79	0.81	0.69
L3	0.64	---	0.90	0.90	0.84	0.87	0.93	0.86	0.71	0.74	0.67	0.74	0.68	0.87	0.82	0.72
L4	0.56	0.69	--	0.86	0.79	0.77	0.98	0.91	0.71	0.63	0.67	0.74	0.68	0.78	0.86	0.67
L5	0.63	0.61	0.64	---	0.79	0.82	0.88	0.82	0.71	0.74	0.62	0.74	0.63	0.83	0.82	0.72
L6	0.57	0.59	0.62	0.62	--	0.80	0.82	0.80	0.59	0.65	0.58	0.72	0.7	0.81	0.80	0.63
L7	0.47	0.49	0.42	0.49	0.32	---	0.80	0.78	0.62	0.63	0.56	0.7	0.63	0.79	0.78	0.67
L8	0.50	0.55	0.46	0.58	0.51	0.43	--	0.93	0.78	0.67	0.70	0.77	0.71	0.81	0.89	0.75
L9	0.55	0.56	0.65	0.61	0.73	0.43	0.63	---	0.77	0.65	0.77	0.84	0.79	0.83	0.91	0.68
L10	0.54	0.57	0.64	0.48	0.68	0.43	0.45	0.71	---	0.78	0.84	0.83	0.82	0.65	0.81	0.86
L11	0.55	0.65	0.56	0.63	0.65	0.54	0.69	0.68	0.60	---	0.79	0.72	0.81	0.67	0.70	0.86
L12	0.57	0.62	0.51	0.57	0.66	0.46	0.56	0.60	0.63	0.66	--	0.84	0.88	0.57	0.73	0.82
L13	0.53	0.63	0.67	0.57	0.57	0.42	0.35	0.50	0.54	0.51	0.45	---	0.81	0.72	0.76	0.70
L14	0.46	0.59	0.62	0.48	0.78	0.27	0.44	0.61	0.56	0.60	0.58	0.56	--	0.62	0.74	0.84
L16	0.62	0.66	0.57	0.51	0.47	0.46	0.50	0.45	0.51	0.56	0.48	0.60	0.48	---	0.79	0.60
L17	0.46	0.57	0.64	0.51	0.58	0.38	0.38	0.56	0.54	0.47	0.48	0.54	0.56	0.54	--	0.73
L19	0.55	0.60	0.53	0.45	0.59	0.36	0.44	0.57	0.69	0.57	0.57	0.52	0.58	0.53	0.56	--

% Similarity Index Color Key

🟥	1.00 - 0.950
🟧	0.949 - 0.900
🟧	0.899 - 0.850
🟨	0.849 - 0.800
🟩	0.799 - 0.750
🟩	0.749 - 0.700
🟦	0.699 - 0.650
🟦	0.649 - 0.600
🟪	0.599 - 0.550
⬛	0.549 - 0.500
⬛	0.499 - 0

Figure 7 – Hybrid C-RFLP and T-RFLP SI heat map for all Lawn (L) soil samples.

SI data for universal bacterial profiling. Section A – Similarity indices for C-RFLP; Section B – Similarity indices for T-RFLP. Similarity values are color coded according to the ranges indicated in the color key.

However, there are some exceptions to this conclusion, specifically those similarity indices that are 0.93 and greater. If we were to use our 0.93 match threshold to determine the likelihood of two soil samples originating from the same location, 3 (out of 120) sample comparisons would meet that criteria (L3 and L8; L4 and L8, and L8 and L9). Knowing the locations of these sites, these conclusions are incorrect. Situations like this in forensic investigations would erroneously lead an investigator to believe that samples likely came from the same place. These data emphasizes the value of collecting replicate samples from any area so as not to base interpretation on one sample that may not be representative of the entire area.

As to our 0.93 match threshold, our data supports the use of a very conservative interpretation where results must emphasize that a similarity of greater than 0.93 *only suggests* that the soils in question *possibly* originated from the same location. From this data we also show that soil profiles cannot be used to definitively affirm a single origination of a sample. Additionally, this data reinforces the need to define a location in a forensic context. "Being from the same location" is a broad characterization, as two reference points can be spatially distributed in a variety of ways within small parks, or even in confined areas like the grid. At this point, a forensic definition for what constitutes a single location cannot be determined.

II.e. Comparison to T-RFLP analysis

The data support C-RFLP as an alternative bacterial profiling method for forensic purposes. The C-RFLP method provides suitable resolution of differences in bacterial communities, important for use in a forensic context. These differences are represented in easy to read chromatograms containing data points that are used to calculate similarity. The set of soil samples in this research have tested the ability of the C-RFLP method to differentiate soils collected from both diverse and similar ecosystems. It is necessary, however, to compare these results with T-RFLP data.

Continuing with the Lawn biological replicate testing, T-RFLP similarity indices are shown in the bottom right section (B) in Figure 7. As stated earlier, a match threshold of 0.64 will be used to establish the likelihood of two soil samples originating from the same location based on T-RFLP data. The Lawn similarity index values shown in Figure 7 have an average of 0.549. Like C-RFLP data, there is a wide range of values seen (0.27 – 0.78). Also similar to C-RFLP, T-RFLP profiles from these presumed biological replicate sites cannot be characterized by a specific percent relatedness. Based on the 0.64 match threshold, there are 14 Lawn comparisons that would be classified as likely to have originated from the same place. L3 and L4, L9 and L10, and L11 and L12 soils did come from different areas of the same park, so their greater than 0.64 similarities are correctly interpreted. However, there are 11 other comparisons that would be incorrectly interpreted.

A point worth noting is the difference in similarity values calculated for the same sample using C-RFLP and T-RFLP. Generally, T-RFLP similarity indices are lower than their C-RFLP counterparts. This has no direct correlation with the usefulness of either method, rather it is a product of the amount of data points that each method generates, and the sensitivity of the instruments used. T-RFLP profiling produces 3-4 times more data points than C-RFLP. T-RFLP analysis detects terminal fragments that differ by one nucleotide, resulting in a significantly larger data set. Ultimately, however, the relevant question is which of the two methods is best for use as a forensic tool.

Table 3 compares T-RFLP and C-RFLP similarity indices for the soil samples discussed in figures 5, 6 and 7.

Table 3 – Similarity indices for soil profiles (as compared to Beach Hall Great Lawn soil, 2007 [BCH 2007])

C-RFLP Similarity Index	Soil Comparison	T-RFLP Similarity Index
0.82	AG Farm 2007	0.68
0.29	Swan Lake 2006 *	0.39
0.69	Mirror Lake 2007	0.50
0.71	Cemetery 2007	0.50
0.72	Field 2007	0.57
0.69	CT River 2007	0.52
0.72	Sludge 2007	0.40
0.63	Lawn 1	0.46
0.65	Lawn 3	0.53
0.60	Lawn 4	0.48
0.60	Lawn 5	0.47
0.44	Lawn 6	0.44
0.54	Lawn 7	0.34
0.63	Lawn 8	0.49
0.62	Lawn 9	0.41
0.56	Lawn 10	0.41
0.56	Lawn 11	0.53
0.45	Lawn 12	0.45
0.49	Lawn 13	0.41
0.56	Lawn 14	0.38
0.64	Lawn 16	0.39
0.57	Lawn 17	0.43
0.59	Lawn 19	0.40

All soil profiles were generated following the universal bacterial typing protocol. (*) Swan Lake 2006 was used as the sample comparison instead of 2007 because the 2007 soil sample was not able to be profiled by the T-RFLP method. Note: All Lawn soil samples were collected in November 2008.

Indices calculated were based on comparison to the Great Lawn sample. There does not appear to be any observable trend regarding ecosystem type and similarity index. For example, T-RFLP data shows Mirror Lake soil and River Sediment both differing by only 2% when compared to Great Lawn. The same comparison using C-RFLP analysis reveals identical similarity to the Great Lawn for both samples (0.69). This data emphasizes the fact that the actual bacterial diversity present in the samples cannot be extrapolated by comparing the calculated SI values. Both C-RFLP and T-RFLP only provide graphical representation of the different types and quantities of bacterial DNA in soil. These match thresholds can only be used to determine the likelihood that any two samples potentially originated from the same location. Using T-RFLP analysis, the 0.68 similarity of AG Farm to Great Lawn soil would be interpreted as soils likely originating from the same location. This conclusion would be incorrect. None of the comparisons facilitated by C-RFLP analysis generated similarity indices that could be incorrectly interpreted.

The forensic implication of this data to the use of T-RFLP as a typing method is significant. In order for soil to be reliably used in a forensic context, all interpretations of sample relatedness must be supported by empirical data. The T-RFLP grid data seen in Figure 4 depicts the range of similarity index values seen within one location, and 52.2% of the values listed in Table 3 fall within this range. In contrast, only 1 (4.3%) sample fell within the C-RFLP grid range. Using T-RFLP profile data to forensically establish the relationship between two soil samples may lead to improper interpretation more frequently than if using C-RFLP analysis. C-RFLP highlights enough differences between samples, yet is not so sensitive that it prompts samples collected from the same area to be interpreted as belonging to unrelated locations. While T-RFLP's sensitivity may be desirable for other applications, the data show its application in this capacity seems problematic.

It is essential to address the differences we see in similarity index values between C-RFLP and T-RFLP. That is, are C-RFLP and T-RFLP measuring the same thing? If we were to rank the sample comparisons for both methods by similarity index values, we would see that the corresponding rankings would not match. This tells us that the way DNA fragments are separated and visualized by both methods creates two very different representations of the genetic information. While both representations are accurate, the question is which method provides the best forensically relevant data.

II.f. Exploring major and minor T-RFLP peak variation

Simply looking at a universal T-RFLP electropherogram, one can clearly see that there are smaller and larger peaks (based on rfu height) (Figure 8). The intensity of a peak can be attributed to fragment quantity. Based on the T-RFLP profiles generated in this research, we can infer that there are major and minor components to bacterial communities in soil. The forensically relevant question here is what can the variation in similarity index values be attributed to: the minor peaks or the major peaks? In order to answer this question, we return to the grid analysis experiment.

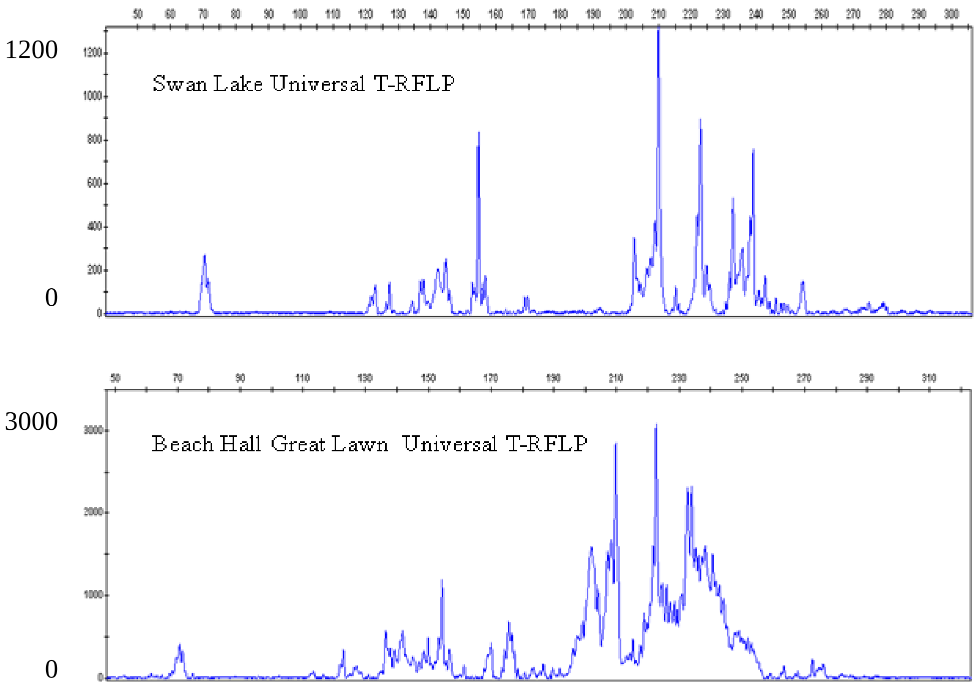
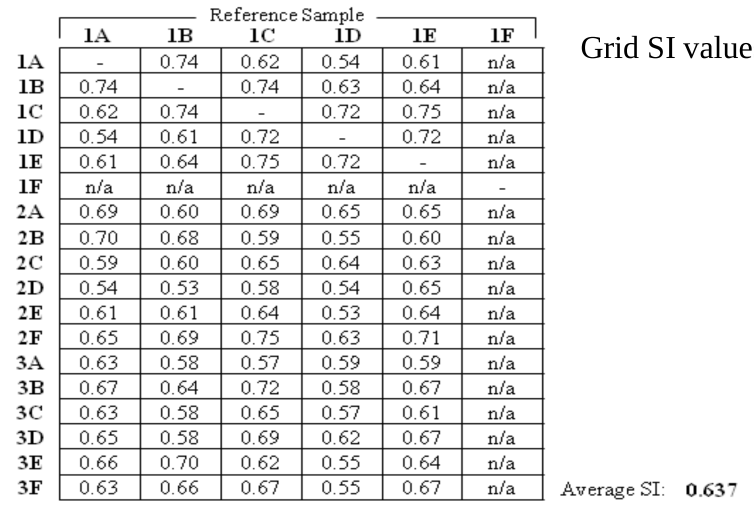


Figure 8 – T-RFLP electropherogram illustrating minor peaks.

Two T-RFLP electropherograms are shown that contain major (high rfu) and minor (low rfu) peaks. Each panel's minor rfu threshold is different based on the overall intensity of the profile. Swan Lake has a minor peak threshold of approximately 300rfu. Beach Hall has a minor peak threshold of approximately 750 rfu.

As shown in Figure 8, the two electropherograms shown have different y-axes (rfu). When loading fluorophore-labeled DNA fragments onto a capillary sequencer it is difficult to standardize the amount of terminal fragments that are loaded. This results in each sample having its own maximum y-axis range (tallest major peak). As a result, the rfu range for the smaller minor peaks will vary. For this experiment, each grid T-RFLP profile was individually examined and a minor peak threshold was chosen for each profile. Choosing each electropherogram's minor peak threshold was subjective; the goal of this threshold was to eliminate a majority of the smaller peaks. Once the smaller (minor) peaks were eliminated from analysis, new similarity index values could be calculated. If similarity index values increased, then we can conclude that the reason for such a wide variation was due to the minor peaks.

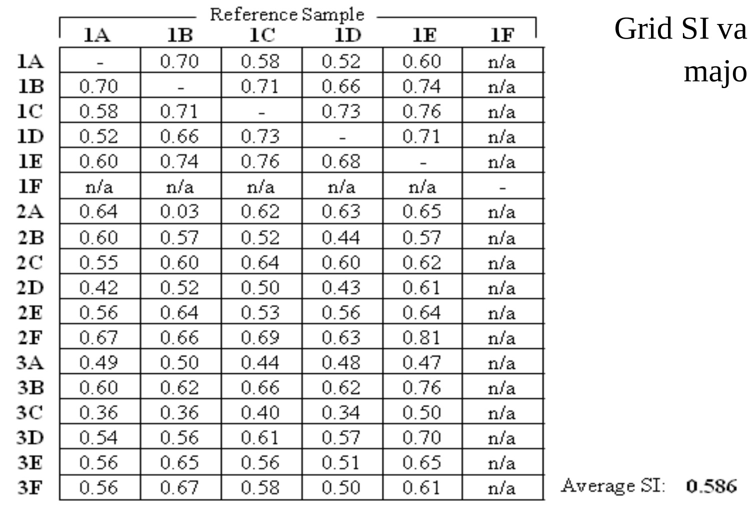
Figure 9 compares similarity index data from T-RFLP grid analysis for all peaks (top panel) and major peaks only (bottom panel). Only data using row 1 as a reference is shown. The data for rows 2 and 3 are consistent with row 1. When all peaks are considered for similarity index calculation, there are an average number of 49.8 peaks. When minor peaks are removed this number drops to 35.2. The data in the bottom panel show that minor peaks are not the source of the variation that we see from grid to grid. The average similarity index value actually decreases by 5.1% when minor peaks are not included. This suggests that some of the minor peaks are shared between the T-RFLP profiles.



	Reference Sample					
	1A	1B	1C	1D	1E	1F
1A	-	0.74	0.62	0.54	0.61	n/a
1B	0.74	-	0.74	0.63	0.64	n/a
1C	0.62	0.74	-	0.72	0.75	n/a
1D	0.54	0.61	0.72	-	0.72	n/a
1E	0.61	0.64	0.75	0.72	-	n/a
1F	n/a	n/a	n/a	n/a	n/a	-
2A	0.69	0.60	0.69	0.65	0.65	n/a
2B	0.70	0.68	0.59	0.55	0.60	n/a
2C	0.59	0.60	0.65	0.64	0.63	n/a
2D	0.54	0.53	0.58	0.54	0.65	n/a
2E	0.61	0.61	0.64	0.53	0.64	n/a
2F	0.65	0.69	0.75	0.63	0.71	n/a
3A	0.63	0.58	0.57	0.59	0.59	n/a
3B	0.67	0.64	0.72	0.58	0.67	n/a
3C	0.63	0.58	0.65	0.57	0.61	n/a
3D	0.65	0.58	0.69	0.62	0.67	n/a
3E	0.66	0.70	0.62	0.55	0.64	n/a
3F	0.63	0.66	0.67	0.55	0.67	n/a

Average SI: **0.637**

Grid SI values for all peaks



	Reference Sample					
	1A	1B	1C	1D	1E	1F
1A	-	0.70	0.58	0.52	0.60	n/a
1B	0.70	-	0.71	0.66	0.74	n/a
1C	0.58	0.71	-	0.73	0.76	n/a
1D	0.52	0.66	0.73	-	0.71	n/a
1E	0.60	0.74	0.76	0.68	-	n/a
1F	n/a	n/a	n/a	n/a	n/a	-
2A	0.64	0.03	0.62	0.63	0.65	n/a
2B	0.60	0.57	0.52	0.44	0.57	n/a
2C	0.55	0.60	0.64	0.60	0.62	n/a
2D	0.42	0.52	0.50	0.43	0.61	n/a
2E	0.56	0.64	0.53	0.56	0.64	n/a
2F	0.67	0.66	0.69	0.63	0.81	n/a
3A	0.49	0.50	0.44	0.48	0.47	n/a
3B	0.60	0.62	0.66	0.62	0.76	n/a
3C	0.36	0.36	0.40	0.34	0.50	n/a
3D	0.54	0.56	0.61	0.57	0.70	n/a
3E	0.56	0.65	0.56	0.51	0.65	n/a
3F	0.56	0.67	0.58	0.50	0.61	n/a

Average SI: **0.586**

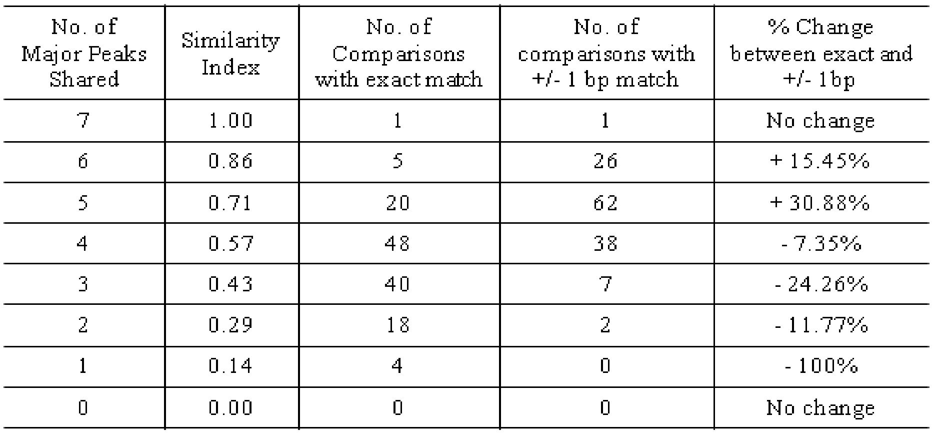
Grid SI values for only major peaks

Figure 9 – Impact of minor T-RFLP peaks on similarity index values for grid analysis.

Universal T-RFLP analysis of bacterial profiles from Beach Hall grid. Only row 1 references are shown. Reference results for rows 2 and 3 were similar. Table compares the difference in similarity index values when all peaks are used for comparison (top) to when only major peaks are used for comparison (bottom). Example of where minor peak rfu threshold would be is described in Figure 8. When all peaks are considered (average of 49.8 peaks), the total average SI for row 1 is 0.637. When only major peaks are considered (average of 35.2 peaks), the total average SI for row 1 is 0.586.

Another important point would be to determine the similarity of major peaks between T-RFLP profiles from the grid. The top 7 tallest peaks were identified in each of the T-RFLP grid electropherograms. Table 4 shows data for all grid comparisons. Column 2 lists the resulting similarity index value for all possible matches (i.e. 6 out of 7 peaks matching would have a 0.86 similarity value). Column 3 lists the total number of comparisons falling in that category. An overwhelming majority of the comparisons had either a 0.43 or 0.57 similarity. This tells us that there is variation among the major peaks as well the minor peaks.

Table 4 – Grid Analysis: Universal T-RFLP analysis of top 7 tallest (major) peaks



No. of Major Peaks Shared	Similarity Index	No. of Comparisons with exact match	No. of comparisons with +/- 1 bp match	% Change between exact and +/- 1bp
7	1.00	1	1	No change
6	0.86	5	26	+ 15.45%
5	0.71	20	62	+ 30.88%
4	0.57	48	38	- 7.35%
3	0.43	40	7	- 24.26%
2	0.29	18	2	- 11.77%
1	0.14	4	0	- 100%
0	0.00	0	0	No change

Table presents grid data for universal T-RFLP analysis. Top 7 tallest peaks are based on rfu (peak height). Columns 3 and 4 in the table compare data for exact match numbers versus peak matches that were +/- 1 base pair. The last column shows if there was an increase or decrease in the number of matches when a +/- 1 base pair criteria is used.

These results can be explained in one of two ways: either the variation we see is real, or the variation is due to a technical artifact. Two of these technical artifacts are incomplete enzyme digestion and incomplete +A addition. It is possible that either one or both of these technical artifacts are present here. Column 4 in Table 4 lists the number of peak matches when a match criterion of +/- 1 base pair is used. If there is incomplete +A addition in some of the major peaks, then making this match criteria more lenient would increase the number of matches. In fact, this is the case. When a +/- 1 base pair window is used, nearly half of all grid comparisons have a similarity index of 0.71 for the top 7 major peaks. There is a 15.45% increase in the amount of 6 out of 7 matches as well. The number of 7 out of 7 matches did not increase.

The data in this experiment show that both major and minor peaks contribute to the variation seen in the grid collection. For the remainder of this chapter, all T-RFLP data will include all peaks for analysis. The variation that is seen within the grid samples is likely due to a combination of true variation as well as technical artifact. Care was taken to minimize incomplete +A addition for all T-RFLP experiments performed for this research by including a 15 minute final extension step at the end of the PCR amplification. Also, amplicons were digested overnight to ensure complete digestion.

II.g. Impact of time on bacterial populations

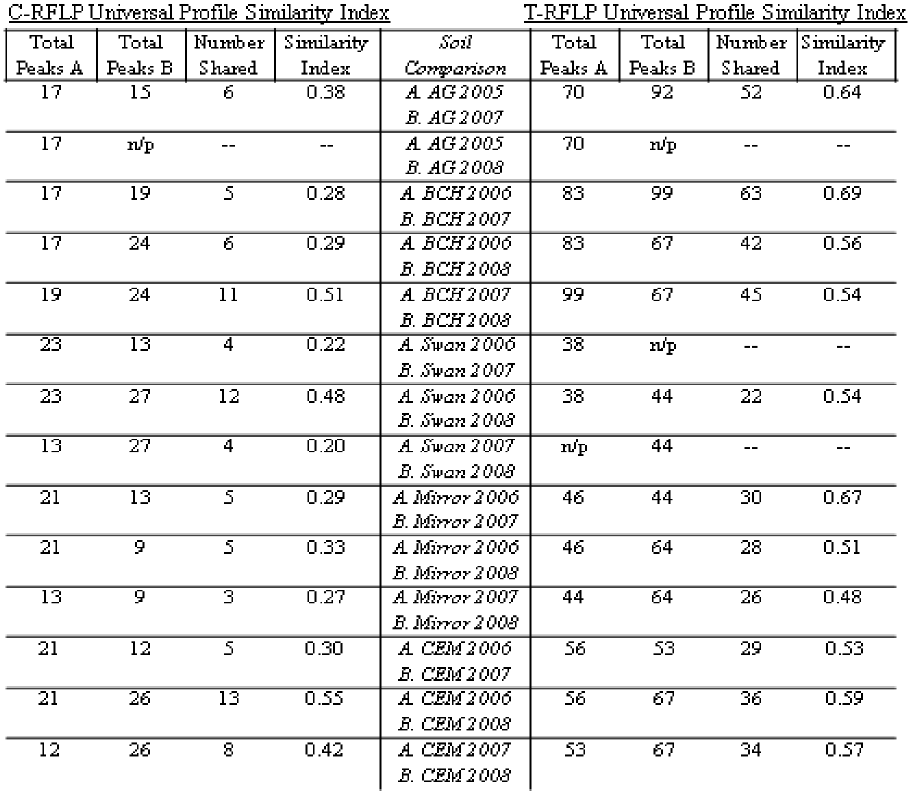
The environment's influence over bacterial community structure cannot be controlled. However, in order for soil to be of use to forensic investigations we must attempt to gauge how much this influence affects DNA profiles. Specifically we must investigate the potential for samples collected from the same area at different time points to falsely be interpreted as originating from unrelated locations.

The success of forensic soil typing is predicated on the fact that one can use the genomic content of a bacterial community to establish a connection between two soil samples. This process assumes that there is enough variability within bacterial communities to differentiate unrelated samples. By and large, the data presented thus far has demonstrated that this is the case for both C-RFLP and T-RFLP. However, if two soil samples were taken from the same location but not at the same time would the ability to successfully associate the two samples be compromised? The goal of this next experiment was to determine if C-RFLP and T-RFLP universal profiles remain consistent over the course of 1 year. If bacterial communities change drastically, similarity index values would be lower than the match criteria. The data presented here will shed light on how time can impact the interpretation of soil evidence.

Sampling locations around the University of Connecticut were visited during the month of July in 2006, 2007, and 2008. When soil samples were collected, care was taken to sample from nearly the exact same location each time. It is important to note that each location could be accessed by the public. Each location was also subject to lawn mowing and general landscaping. Both of these factors could potentially influence the bacterial community structure. Ultimately these factors were desirable, as they created realistic scenarios for testing.

Table 5 lists the similarity index data for both C-RFLP and T-RFLP universal profiling. Sample comparisons are listed in the middle of the table. None of the soil comparisons profiled by C-RFLP generated similarity index values greater than the 0.93 match threshold. The data suggests that the native bacterial communities to each location have changed within the course of 1 year. Without knowing that these soil samples did come from the same location, the analyst would reach an incorrect conclusion about the relatedness of soil samples.

Table 5 – Similarity index comparisons for C-RFLP and T-RFLP universal soil profiling: Year-to-year site monitoring



C-RFLP Universal Profile Similarity Index					T-RFLP Universal Profile Similarity Index			
Total Peaks A	Total Peaks B	Number Shared	Similarity Index	*Soil Comparison*	Total Peaks A	Total Peaks B	Number Shared	Similarity Index
17	15	6	0.38	*A. AG 2005* *B. AG 2007*	70	92	52	0.64
17	n/p	--	--	*A. AG 2005* *B. AG 2008*	70	n/p	--	--
17	19	5	0.28	*A. BCH 2006* *B. BCH 2007*	83	99	63	0.69
17	24	6	0.29	*A. BCH 2006* *B. BCH 2008*	83	67	42	0.56
19	24	11	0.51	*A. BCH 2007* *B. BCH 2008*	99	67	45	0.54
23	13	4	0.22	*A. Swan 2006* *B. Swan 2007*	38	n/p	--	--
23	27	12	0.48	*A. Swan 2006* *B. Swan 2008*	38	44	22	0.54
13	27	4	0.20	*A. Swan 2007* *B. Swan 2008*	n/p	44	--	--
21	13	5	0.29	*A. Mirror 2006* *B. Mirror 2007*	46	44	30	0.67
21	9	5	0.33	*A. Mirror 2006* *B. Mirror 2008*	46	64	28	0.51
13	9	3	0.27	*A. Mirror 2007* *B. Mirror 2008*	44	64	26	0.48
21	12	5	0.30	*A. CEM 2006* *B. CEM 2007*	56	53	29	0.53
21	26	13	0.55	*A. CEM 2006* *B. CEM 2008*	56	67	36	0.59
12	26	8	0.42	*A. CEM 2007* *B. CEM 2008*	53	67	34	0.57

Left side of table lists C-RFLP data; right side of table lists T-RFLP data. Each side contains data on the number of peaks in the designated profiles "A" and "B", the number of peak shared between two profiles and the resulting SI value. "n/p" designates no profile for comparison.

The similarity index results for T-RFLP analysis also demonstrate community change. However, there are two comparisons that gave a greater than 0.64 similarity index and one that was equal to the match criteria. Although this is a positive result for T-RFLP analysis, the remaining comparisons still suggest community fluctuation.

Results for C-RFLP and T-RFLP highlight an important limitation to the forensic analysis of soils. If reference samples are not collected within a timely fashion from when an evidentiary sample is received, there is a possibility that universal bacterial profiling will reach an incorrect conclusion about the relatedness of two soil samples.

Therefore, whenever possible, reference samples must be collected as soon as a location is known. Proper interpretation of soil data must then account for the possibility of time influencing results.

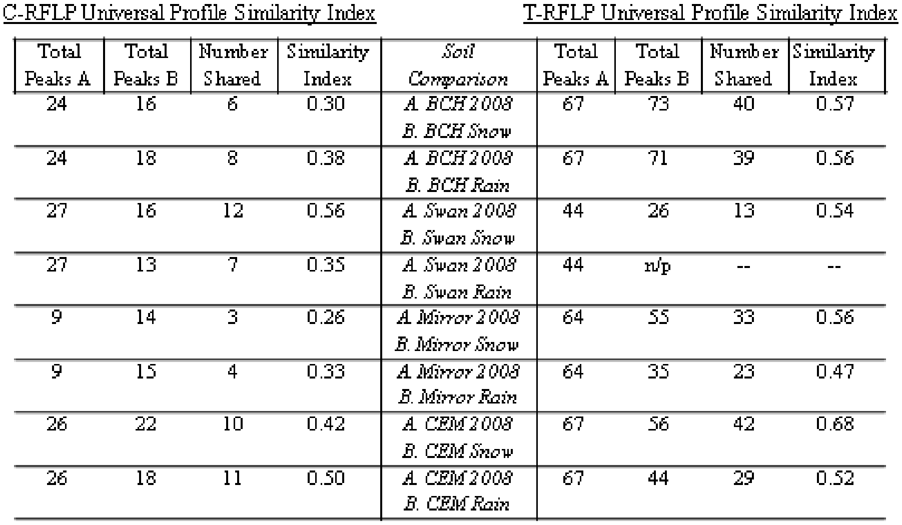
II.h. Impact of meteorological events on bacterial populations

Another forensically relevant factor that can influence soil bacterial communities are meteorological events. Sampling locations around the University of Connecticut were visited in order to evaluate whether universal bacterial profiles change after heavy rainfall and when the ground is covered in snow (as compared to a control sample).

During the month of March 2008, soil was collected from Beach Hall Great Lawn, Swan Lake, Mirror Lake, and Cemetery. This soil was collected on a day when the weather could be classified as "seasonable". Within a span of 3 weeks, there was one instance of heavy rainfall and one instance where the soil was covered by 1 inch of snow.

Soil samples were extracted and profiled using T-RFLP and C-RFLP. Table 6 shows similarity index data for all samples. The control samples (location, 2008) were each compared to the snow and rain samples. None of the C-RFLP similarity indices were greater than the 0.93 match criteria. Only one T-RFLP similarity index was greater than the 0.64 match criteria. The data suggests that bacterial communities do change as a result of meteorological events. This change happens quickly, as these 3 samples were collected within the same month. As discussed in the year-to-year section, the same interpretational limitation applies here. When collecting reference samples, it is important to make note of any meteorological events, as they may impact the native bacterial community structure.

Table 6 – Similarity index comparisons for C-RFLP and T-RFLP universal soil profiling: Meteorological event site monitoring



C-RFLP Universal Profile Similarity Index					T-RFLP Universal Profile Similarity Index			
Total Peaks A	Total Peaks B	Number Shared	Similarity Index	*Soil Comparison*	Total Peaks A	Total Peaks B	Number Shared	Similarity Index
24	16	6	0.30	*A. BCH 2008* *B. BCH Snow*	67	73	40	0.57
24	18	8	0.38	*A. BCH 2008* *B. BCH Rain*	67	71	39	0.56
27	16	12	0.56	*A. Swan 2008* *B. Swan Snow*	44	26	13	0.54
27	13	7	0.35	*A. Swan 2008* *B. Swan Rain*	44	n/p	--	--
9	14	3	0.26	*A. Mirror 2008* *B. Mirror Snow*	64	55	33	0.56
9	15	4	0.33	*A. Mirror 2008* *B. Mirror Rain*	64	35	23	0.47
26	22	10	0.42	*A. CEM 2008* *B. CEM Snow*	67	56	42	0.68
26	18	11	0.50	*A. CEM 2008* *B. CEM Rain*	67	44	29	0.52

Left side of table lists C-RFLP data; right side of table lists T-RFLP data. Each side contains data on the number of peaks in the designated profiles "A" and "B", the number of peak shared between two profiles and the resulting SI value. "n/p" designates no profile for comparison.

III. Discussion and Conclusions

The goal of this chapter was to investigate the potential for an alternative to T-RFLP analysis of bacterial communities in soil. While it is unlikely that soil evidence will be a part of every criminal investigation, it is important that the forensic community have the ability to use this evidence when needed in specific, high profile cases. Currently, T-RFLP is being investigated for potential use by the forensic community to objectively measure relatedness between samples. The main objective of the forensic comparison of soil samples is to determine if two soil samples could have come from the same location. The study published by Meyers and Foran provided data that suggests soil analysis is best suited for use as associative evidence [Meyers and Foran, 2008]. The data shown in this study supports this conclusion as well, with respect to both T-RFLP and C-RFLP analysis. As shown by year-to-year and meteorological event sampling, concrete conclusions about where unknown samples originate from may never be possible due to environmental variables that cannot be controlled. Environmental variables like temperature change and rain can influence bacterial community structure, thereby altering the "native" DNA profile [Lipson and Schmidt, 2004; Smit *et al.*, 2001]. This can falsely lead to samples being interpreted as unrelated. Therefore, it is important that evidentiary and reference samples are collected together and within a short time frame of one another.

The data presented demonstrates that C-RFLP bacterial community analysis is an additional way to represent bacterial variability in soil. Although C-RFLP analysis is subject to the same interpretational limitations as T-RFLP, the C-RFLP method seems more promising for forensic applications. First, a C-RFLP profile is easily interpretable because of a manageable number of data points. Second, the data points generated from fragment separation are highly reproducible. Third, the soil samples collected in this study were successfully individualized by C-RFLP. Although neither T-RFLP nor C-RFLP performed perfectly, C-RFLP analysis appears to be better suited for forensic applications based on this data. A more appropriate, broad statement would be that forensic soil analysis appears to be better paired with a molecular typing method less sensitive than T-RFLP.

When comparing C-RFLP and T-RFLP data using the Sorensen similarity index, the most obvious difference between the methods is their respective range values. The differences seen in similarity index values are attributed to the way DNA fragments are detected by HPLC and capillary electrophoresis. HPLC separation and visualization is achieved by measuring absorbance at 260nm. Because this measure is constant, peaks are often wide, spanning 0.1 – 0.2 minutes. Within these wider single peaks, there are likely multiple fragments being represented. Thus, HPLC separation consolidates closely sized fragments thereby reducing the number of peaks we expect to see. This is slightly counter intuitive, seeing as C-RFLP utilizes all fragments, while T-RFLP only resolves the terminally labeled one. In contrast, capillary separation resolves fragments that differ by one base pair. T-RFLP peaks are rarely wide and as a result the profiles contain many data points (3-4 times more data points than C-RFLP). Furthermore, the capillary electrophoresis instrument's ability to detect small amounts of fluorophore results in the potential to detect labeled fragments that are poorly represented in the sample. These minor fragments may go undetected using HPLC separation. Because the potential starting amounts of data points between each method are so different, the possible range of similarity index values will also be influenced. This may explain why T-RFLP analysis of all grid samples never generated similarity values higher than 0.77; T-RFLP analysis was too sensitive.

Grid analysis provided very valuable information regarding match criteria and sampling. In order for soil to be used as forensic evidence, multiple soil samples taken from a single, homogenous geography must be shown to have similarity indices higher than soils from unrelated locations so that related and unrelated samples can be objectively distinguished. This criterion is true for both C-RFLP and T-RFLP. Based on the T-RFLP data, samples taken from within a single location produced some similarity index values that were indistinguishable from samples being compared from unrelated locations. This may be a consequence of the extreme sensitivity of T-RFLP. The minor peaks T-RFLP generates may contribute to the significant amount of variation between samples taken from the same location. Also, our match threshold was established from a single grid experiment. It would be crucial to continue these sampling studies on different ecosystems. Increased data would strengthen the reliability of using a single similarity index value to determine relatedness. These experiments have also called attention to the value of collecting multiple samples from any soil. Given the heterogeneity of the soil matrix, a more complete analysis must include several samples so not to randomly choose one that may not accurately represent the entire location.

Establishing and validating match criteria will be a challenge to forensics. Also a challenge will be defining what a match means. For example, does a high similarity index between two samples always support the conclusion that two soils *definitely* came from the *same* location, or is it more appropriate to conclude that high similarity *only suggests* two samples *could have come from* the same location? Based on the data presented, a conservative interpretation is best. There is always the possibility for bacterial communities to change slightly in response to the environment and/or time. Also, there are some bacterial communities found in unrelated locations that happen to be very similar. The question that begs an answer is whether sample similarities are random chance events, or whether their high similarities can be explained biologically. An example of this was shown in Figure 7 where Lawn data was discussed. One explanation for high similarity may lie in the maintenance of the lawns. Perhaps similar (if not identical) fertilizer blends were used on all the lawns, thus normalizing bacterial populations towards groups that survive best under those conditions. At this point we can only speculate without further knowledge of lawn care, or other biological replicates to test.

More research needs to be done on forensically relevant questions exploring bacterial populations in soil. First, a very limited set of soil types was examined in this research. It is important to address how any method will perform on less homogenous areas of forensic interest, like forest samples. Second, nucleic acids were extracted from 1.0 gram of soil. It will be important to determine the minimum amount of soil needed to generate a reliable, forensic profile. Third, understanding how bacterial populations change in the natural environment will be crucial for forensic applications so that DNA profiles can be properly interpreted without overreaching analysis. Fourth, investigation into how statistics can be used to provide the most objective and powerful support is also needed. More grid studies would help establish whether there is a minimum threshold for concluding that two samples could originate from the same location. Additionally it would be of use to explore whether different ecosystems would require the use of different thresholds.

Any forensic method must demonstrate accuracy and reproducibility. The data show that universal T-RFLP and C-RFLP are not 100% accurate in reaching conclusions about relatedness. The question becomes, how can soil analysis be improved? Meyers and Foran (2008) suggested that targeting bacterial groups for analysis may be a better alternative to universal typing. The next chapter sets the stage for group-specific analysis by exploring soil bacterial communities using modern, high-throughput sequencing. 454 amplicon pyrosequencing provides a unique way to thoroughly catalog bacteria in soil, primarily because of the sequence depth achieved. Taking a detailed look at the major and minor components of the bacterial community will create a better understanding of what makes bacterial communities different in soil samples. These broad, in-depth DNA libraries will also focus attention on a select set of potentially informative bacterial groups.

Chapter 3 – Using next generation sequencing to survey bacterial communities in soil

I. Introduction

Obtaining an accurate and comprehensive picture of the structure and richness of bacterial communities in soil has been a long standing goal of microbial ecologists. With thousands of bacterial species estimated in 1 gram of soil, it comes as no surprise that achieving this goal is dependent on the use of an appropriate molecular method [Torsvik and Ovreas, 2002]. While there are many molecular technologies available, the consensus among microbiologists is that the best method will be PCR-based. Culture-dependent techniques are only capable of identifying a small percentage (1% or less) of the total bacterial community [Kirk *et al.*, 2004; Torsvik *et al.*, 1990]. Given the estimated richness of bacteria in soil, this small percentage cannot accurately represent the entire community. Thus, PCR amplification targeting the 16S rRNA gene is the most commonly used method for obtaining information about complex bacterial communities.

The 16S rRNA gene is composed of 9 hypervariable regions. Each region contains different degrees of polymorphisms, making some regions more informative than others for differentiating bacterial taxa [Baker *et al.*, 2004]. Conserved regions increase the potential for diverse bacteria to be amplified by providing flanking areas for primers to anneal [Baker *et al.*, 2004]. It is important to note that universal primers that truly amplify all bacteria do not exist. Although there are clusters of conserved nucleotides throughout portions of the 16S gene, ultra-conserved regions are not found in long enough stretches where primers of optimal length could anneal. Despite its limitations, the genetic information provided by 16S gene amplicons is sufficient for estimating bacterial richness [Sogin *et al.*, 2006].

Until recently, clone libraries were widely used to catalog bacterial species in complex samples [Janssen *et al.*, 2006; Liles *et al.*, 2003]. Although generating clone libraries is a tried and true molecular method, using this technique to assess actual bacterial diversity is problematic. First, only bacteria found in majority components of populations will be easily detected [Janssen *et al.*, 2006]. Rare contributors may never be detected if an insufficient number of clones are screened. Second, clone library analysis is labor-intensive with many duplicates sequenced. Ultimately, the question reverts back to, "How many clones must be sequenced to generate reliable data about the population composition?" Without prior knowledge about community richness, this question is difficult to answer.

Techniques like pyrosequencing have revolutionized the way scientists approach microbial surveys because the need for cloning is eliminated. Specifically, amplicon sequencing using the 454 GS FLX system has been used for many studies where in-depth bacterial surveys were necessary [Acosta-Martinez *et al.*, 2008; Humbolt and Guyot, 2009; Miller *et al.*, 2009; Sogin *et al.*, 2006]. Depending on the specific sequencing strategy used, 454 amplicon libraries can generate tens to hundreds of thousands of sequences per sample in a single run. Such depth of coverage provides a virtually unbiased representation of both major and minor components to bacterial communities.

Among the many fields that can benefit from 454 sequencing is forensic science. Although high throughput sequencing is not currently a common protocol in forensic labs, this type of analysis can advance forensic research. For example, Fierer *et al.*

described how 454 pyrosequencing was used to establish linkage between an object and individual based on the skin-associated bacteria left on the object by the individual [Fierer *et al.*, 2010]. The goal of this chapter was to use in-depth amplicon libraries from bacteria in soil to improve the underlying knowledge base for the potential use of soil as evidence. First, the sequence coverage required to accurately describe bacterial richness in soil was established. Based on this information, the bacterial community structures of diverse and similar ecosystems were screened to determine if sufficient differences exist to support the feasibility of forensic analysis. Soils collected after meteorological events and over time were screened to determine if there were measureable changes in community structure and to help establish parameters for collection in a forensic context. These data will also determine if specific ecosystems might be identified by unique bacterial signatures, providing possible investigative leads. The comprehensive nature of these results will help forensic scientists craft a DNA typing protocol for soil that utilizes bacteria to best advantage without overreaching interpretational limitations.

II. Results

II.a. Design of soil library sets

Amplicon libraries were generated from four subsets of the soil samples in Table 1. The first set represented four diverse soil samples from radically different ecosystems: Agricultural soil from a corn plot (AG Farm), soil from a maintained lawn area (Field), freshwater river sediment (CT River) and secondary waste treatment sludge (Sludge). Samples presumed to be radically different in terms of native bacterial communities delineate the taxa that contribute to these specific diverse communities and the levels of distinction between them. In-depth surveys will identify any bacterial groups specific to certain ecosystems.

The second set of samples was collected from a maintained lawn on the University of Connecticut campus (Beach Hall). At this location, soil was collected once in the month of July in the years 2006, 2007 and 2008. This set of samples revealed the extent of bacterial community change over the years.

The third set of samples was also collected from a maintained lawn on the University of Connecticut campus (Cemetery). At this location, soil was collected during the month of March, 2008 on three separate occasions. The specifics of sample collection are described elsewhere (Chapter 2, section II.h.). These samples revealed the extent of bacterial community change as a result of meteorological events and provide a second maintained lawn in close proximity to the first. Both year to year and meteorological event sampling will help guide collection strategies and determine the limitations to using soil for associative purposes.

The fourth set of soil samples were collected from different locations in Connecticut but were presumed biological replicate ecosystems of maintained lawns (Lawn 1, 9, 14 and 17). These locations were characterized by vegetation and landscape that looked the same on a superficial level. Maintained lawns in park areas around the state of Connecticut were sampled from a region of the lawn that was always 10’ away from a tree/shrub border. One purpose of studying unrelated locations that share a common ecosystem was to determine if bacterial communities provide “signatures” for the specific soil ecosystem. This would be valuable to forensics, especially if investigative leads were needed. Whether or not there were ecological signatures, this set of samples would also provide information on the possibility of distinguishing presumed biological replicate ecosystems from different locations.

The AG Farm, Field, CT River, Sludge, Beach Hall and Cemetery samples were all amplified using a V6 primer. At the time, this approximately 100 base pair amplicon was optimal for the manufacturer's standard chemistry emulsion PCR protocol. As the manufacturer optimized for product efficiency, a 100 base pair amplicon was outside the optimal size range for titanium chemistry emulsion PCR. For the remainder of the samples, all soils were amplified with primers targeting the V1 and V2 regions of the 16S gene, generating an approximately 400 base pair amplicon.

II.b. Measuring species diversity

One goal of this study was to use 454 amplicon sequencing to thoroughly catalog native bacterial communities from a variety of soil samples. An inherent problem in this approach was that actual community composition was not known, nor was depth of sequence coverage required to accurately determine species richness and abundance**.** The great advantage of 454 amplicon sequencing is that the large number of DNA sequences generated from one DNA library allows empirical evaluation of the effectiveness of coverage. First, the purpose of this section was to evaluate how well each sequencing run sampled community diversity. Second, using various statistical measures we evaluated each community's richness and structure.

One of the techniques used to evaluate sequence coverage involves construction of a rarefaction curve. A rarefaction curve is generated by plotting the number of operational taxonomic units (OTUs) observed against a subset of sequences from the original library. The morphology of the curve dictates if the sequence library has discovered a majority of the OTUs in the sample (flat line) or if many of the OTUs in the sample remain undiscovered (steep slope). Figure 10 shows 4 rarefaction curves for the DNA libraries generated in this study, grouped by sample set. A forensically relevant question answered with these curves is how many sequences are needed to accurately represent the bacterial community. Also, if there is a value that can be used for all sample types, or do certain ecosystems require a larger number of sequences than others? In the future, 454 amplicon sequencing is used to generate forensic bacterial community profiles from unknown soil ecosystems, it would be important to know the depth of coverage needed in order to accurately determine a possible location of origin.

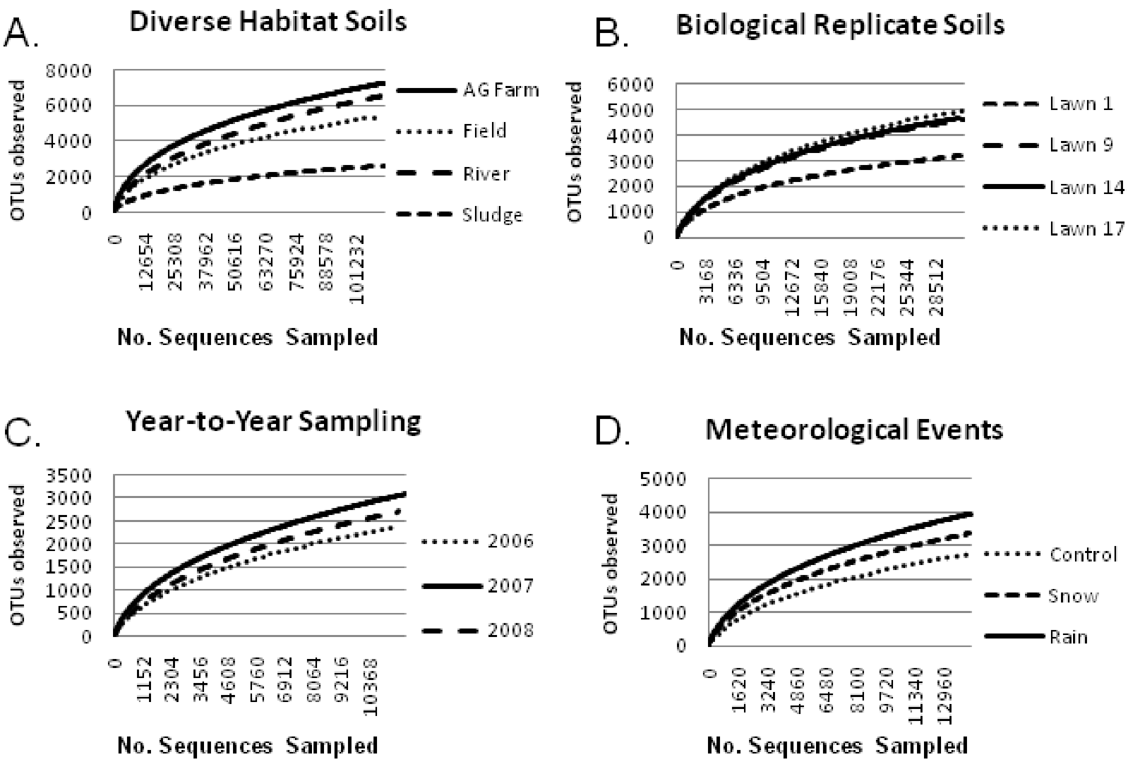


Figure 10 – Rarefaction curves for soil libraries

Rarefactions curves were generated for all amplicon DNA libraries in this study. Soil samples used to compare diverse ecosystems are shown in panel A. Biological replicate lawn soils are shown in panel B. Beach Hall soil collected in the years 2006, 2007 and 2008 are shown in panel C. Cemetery soils collected following meteorological events are shown in panel D. All rarefaction curves were plotted using a 95% sequence similarity. Sequence alignment and complete linkage clustering was performed using the Ribosomal Database Project's Pyrosequencing Pipeline (http://pyro.cme.msu.edu/).

There are differences in total DNA sequences from each library in Figure 10 (x-axis). One of the reasons for differences in final sequence counts is attributed to sample preparation. Positive DNA bead collection is never 100% efficient. Therefore, each sample will yield anywhere from 5% - 20% (or more) of the total bead input. The number of regions chosen for a PTP will also affect the number of possible sequences per run. The DNA libraries in panel A were run on a 4 region PTP which allowed for between 160,000 and 250,000 high quality reads per region (Titanium Sequencing Manual, Roche). The samples in panels B, C and D were run on an 8 region PTP which allowed for potential high quality reads of 80,000 to 120,000.

Panel A illustrates rarefaction curves for soils collected from diverse ecosystems. This sample set was used as a pilot deep-sequencing experiment that provided an estimate of how many sequences may be needed to describe very complex soils (i.e. agricultural farm soil) to the least complex sample of the group (i.e. sewage sludge). The curves show that each bacterial community contains a different amount of species diversity. The sewage sludge sample is the least diverse of the four samples, with its curve flattening out at approximately 38,000 sequences. There were 2,623 OTUs detected in the sludge sample. The agricultural farm soil has the most diverse bacterial community with 7,278 OTUs identified in its sequence library.

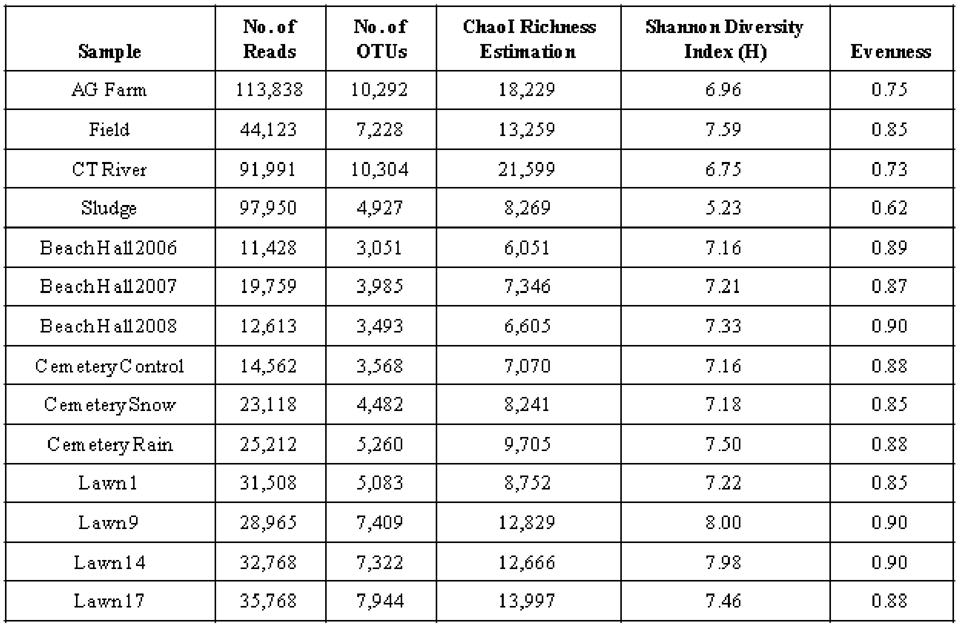
Although performing four region PTP runs does provide an in-depth survey of the bacterial community, such coverage might not be necessary to capture enough diversity so that comparisons can be made. For example, if 29,300 sequences were selected from the AG Farm and the Sewage sludge, 4,032 and 1,425 OTUs, respectively, would be detected in each sample. With respect to total OTUs detected in the complete library, each subset of sequences detects 55% of the total OTUs. The question becomes whether this 55% majority is sufficient for describing diversity. By increasing the PTP regions from four to eight we can determine whether smaller sequence libraries compromise sampling effectiveness.

The 454 data shown in panels B – D evaluated the effectiveness of using lower coverage runs to measure diversity. The average Lawn sequence library (Panel B) size was 32,252. In panel B, the curves for Lawn 9, 14 and 17 cluster tightly suggesting similar richness to their communities at the genus level (95% sequence similarity). Lawn 1 was less diverse at the genus level than the other three lawns. The plateaus seen in the curves for the Lawn libraries indicate that these lower coverage libraries are sufficient in detecting diversity.

Although the rarefaction curves in panels C and D all share the same morphology, the curves do not plateau as much as in panel B. This is likely a result of the fact that the Lawn sample libraries contained about 2.5 times more sequences. These data demonstrate that the more sequences you have the more OTUs you will detect, giving a more complete picture of diversity. The data shown in panels C and D were derived from even smaller libraries than the Lawns. Panel C (Beach Hall year-to-year) had the smallest average number of sequences at 14,600. Panel D (Cemetery meteorological event sampling) contained an average 20,964 sequences.

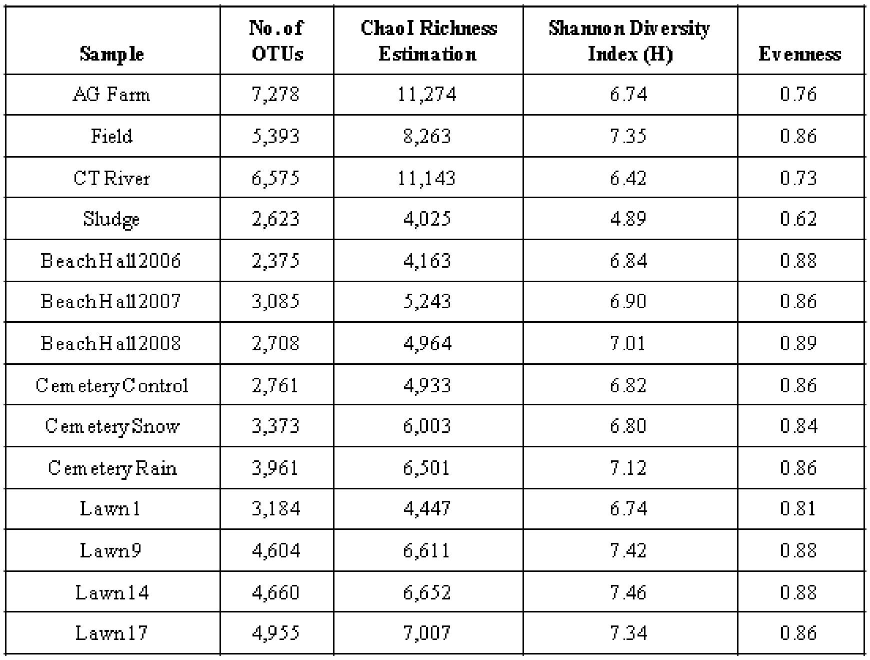
Tables 7A and 7B list various community diversity measures for all soil libraries. Table 7A includes data using a 97% sequence similarity. Table 7B includes data using a 95% sequence similarity. A 97% sequence similarity describes bacteria at the species level while a 95% sequence similarity categorizes at the genus level. The ChaoI richness estimation, Shannon diversity index (H) and Evenness values were calculated using the Ribosomal Database Project's (RDP) Pyrosequencing pipeline [Wang *et al.*, 2007]. The data show that both the AG Farm and CT River contain the largest number of OTUs at both the species and genus levels. The ChaoI richness measure predicts that there should be more OTUs in all the soil libraries we generated. In general, the libraries generated in this study identified at least 50% or more of the expected OTUs in all samples for both sequence similarities.

Table 7A – Bacterial community diversity comparisons: 97% sequence similarity for species identification



Sample	No. of Reads	No. of OTUs	ChaoI Richness Estimation	Shannon Diversity Index (H)	Evenness
AG Farm	113,838	10,292	18,229	6.96	0.75
Field	44,123	7,228	13,259	7.59	0.85
CT River	91,991	10,304	21,599	6.75	0.73
Sludge	97,950	4,927	8,269	5.23	0.62
BeachHall2006	11,428	3,051	6,051	7.16	0.89
BeachHall2007	19,759	3,985	7,346	7.21	0.87
BeachHall2008	12,613	3,493	6,605	7.33	0.90
CemeteryControl	14,562	3,568	7,070	7.16	0.88
CemeterySnow	23,118	4,482	8,241	7.18	0.85
CemeteryRain	25,212	5,260	9,705	7.50	0.88
Lawn1	31,508	5,083	8,752	7.22	0.85
Lawn9	28,965	7,409	12,829	8.00	0.90
Lawn14	32,768	7,322	12,666	7.98	0.90
Lawn17	35,768	7,944	13,997	7.46	0.88

Table 7B – Bacterial community diversity comparisons: 95% sequence similarity for genus identification



Sample	No. of OTUs	ChaoI Richness Estimation	Shannon Diversity Index (H)	Evenness
AG Farm	7,278	11,274	6.74	0.76
Field	5,393	8,263	7.35	0.86
CT River	6,575	11,143	6.42	0.73
Sludge	2,623	4,025	4.89	0.62
BeachHall2006	2,375	4,163	6.84	0.88
BeachHall2007	3,085	5,243	6.90	0.86
BeachHall2008	2,708	4,964	7.01	0.89
CemeteryControl	2,761	4,933	6.82	0.86
CemeterySnow	3,373	6,003	6.80	0.84
CemeteryRain	3,961	6,501	7.12	0.86
Lawn1	3,184	4,447	6.74	0.81
Lawn9	4,604	6,611	7.42	0.88
Lawn14	4,660	6,652	7.46	0.88
Lawn17	4,955	7,007	7.34	0.86

Tables list all soil libraries generated in this study. Total number of sequence reads are included. ChaoI richness estimation, Shannon diversity index (H) and Evenness calculations were generated by the Ribosomal Database Project's Pyrosequencing pipeline (http://pyro.cme.msu.edu/).

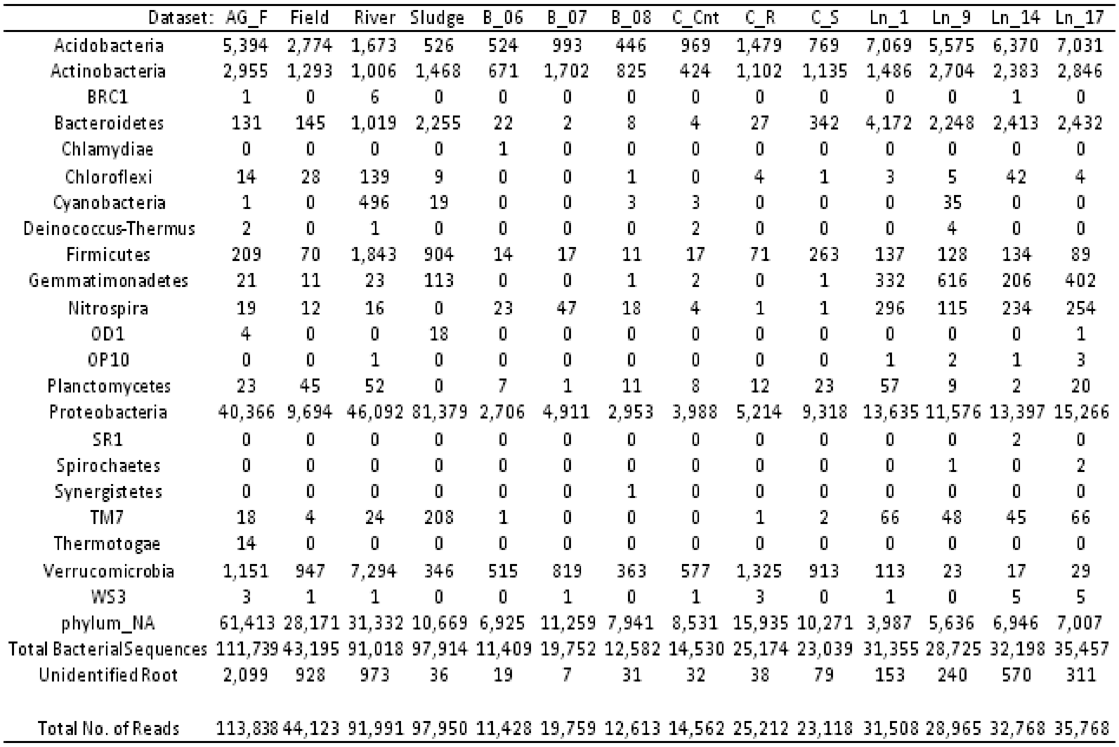
To fully describe bacterial diversity, two components are often addressed: species diversity (calculated by the Shannon Index) and Evenness. The Shannon Index (H) provides a measure of species diversity. The range of values for H will vary according to the sample size used for calculation. The larger the sample size, the greater the range of H. In general, as the H value increases diversity increases. However, H will be affected by evenness. If there are a few bacterial species/genera that dominate the community, H will decrease. Evenness is calculated from the H value, as it is ratio of the actual H value to the maximum H value. The values generated from the Evenness calculation range from 0 to 1, where a value closer to 1 means that the members of the bacterial community are more evenly distributed.

Of all soil libraries in this study, Lawn 9 contains the most evenly distributed and diverse bacterial community at the species level (0.03). In general, the Lawn samples contain the most evenly distributed communities. Furthermore, maintained lawns (all Lawn, Beach Hall, Cemetery soils as well as the Field sample) as a group have more diverse and evenly distributed bacterial communities than the AG Farm and CT River. While Tables 2A and 2B showed that the AG Farm and River contained the most OTUs, their H values likely dropped as a result of select species domination in the community. As expected, the sewage sludge was the least diverse sample in this study. Data for the genus level changes slightly, but the overall trends are still the same.

II.c. Taxonomic Classification of Soils

To determine the types and abundance of bacterial phyla represented in each soil library, all sequences were classified using the RDP. The taxonomic sequence classification results are shown in Table 8. Lawn soil libraries were created from V1-V2 amplicons. Because these amplicons were greater than 250 base pairs, the RDP recommended an 80% confidence threshold for classification. All other soil libraries consisted of V6 amplicons which were approximately 100 base pairs. At the time these data were analyzed, the RDP recommended using a 60% confidence threshold for classification. [Note: As of July 2010, the RDP updated this classification criterion to 50% based on a study by Claesson *et al.*, 2009]. Given the size difference of these amplicons, it is important to consider the adequacy of classification using libraries of different lengths. Liu *et al.* explored the use of short pyrosequencing reads to characterize bacterial communities [Liu *et al.*, 2007]. While it is true that using a larger segment of the 16S rRNA gene will resolve phylogenetic differences with greater accuracy, short reads of 100 base pairs can still perform as well as full-length 16S reads [Liu *et al.*, 2007].

Table 8 – Taxonomic sequence classification for amplicon libraries



Dataset:	AG_F	Field	River	Sludge	B_06	B_07	B_08	C_Cnt	C_R	C_S	Ln_1	Ln_9	Ln_14	Ln_17
Acidobacteria	5,394	2,774	1,673	526	524	993	446	969	1,479	769	7,069	5,575	6,370	7,031
Actinobacteria	2,955	1,293	1,006	1,468	671	1,702	825	424	1,102	1,135	1,486	2,704	2,383	2,846
BRC1	1	0	6	0	0	0	0	0	0	0	0	0	1	0
Bacteroidetes	131	145	1,019	2,255	22	2	8	4	27	342	4,172	2,248	2,413	2,432
Chlamydiae	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Chloroflexi	14	28	139	9	0	0	1	0	4	1	3	5	42	4
Cyanobacteria	1	0	496	19	0	0	3	3	0	0	0	35	0	0
Deinococcus-Thermus	2	0	1	0	0	0	0	2	0	0	0	4	0	0
Firmicutes	209	70	1,843	904	14	17	11	17	71	263	137	128	134	89
Gemmatimonadetes	21	11	23	113	0	0	1	2	0	1	332	616	206	402
Nitrospira	19	12	16	0	23	47	18	4	1	1	296	115	234	254
OD1	4	0	0	18	0	0	0	0	0	0	0	0	0	1
OP10	0	0	1	0	0	0	0	0	0	0	1	2	1	3
Planctomycetes	23	45	52	0	7	1	11	8	12	23	57	9	2	20
Proteobacteria	40,366	9,694	46,092	81,379	2,706	4,911	2,953	3,988	5,214	9,318	13,635	11,576	13,397	15,266
SR1	0	0	0	0	0	0	0	0	0	0	0	0	2	0
Spirochaetes	0	0	0	0	0	0	0	0	0	0	0	1	0	2
Synergistetes	0	0	0	0	0	0	1	0	0	0	0	0	0	0
TM7	18	4	24	208	1	0	0	0	1	2	66	48	45	66
Thermotogae	14	0	0	0	0	0	0	0	0	0	0	0	0	0
Verrucomicrobia	1,151	947	7,294	346	515	819	363	577	1,325	913	113	23	17	29
WS3	3	1	1	0	0	1	0	1	3	0	1	0	5	5
phylum_NA	61,413	28,171	31,332	10,669	6,925	11,259	7,941	8,531	15,935	10,271	3,987	5,636	6,946	7,007
Total Bacterial Sequences	111,739	43,195	91,018	97,914	11,409	19,752	12,582	14,530	25,174	23,039	31,355	28,725	32,198	35,457
Unidentified Root	2,099	928	973	36	19	7	31	32	38	79	153	240	570	311
Total No. of Reads	113,838	44,123	91,991	97,950	11,428	19,759	12,613	14,562	25,212	23,118	31,508	28,965	32,768	35,768

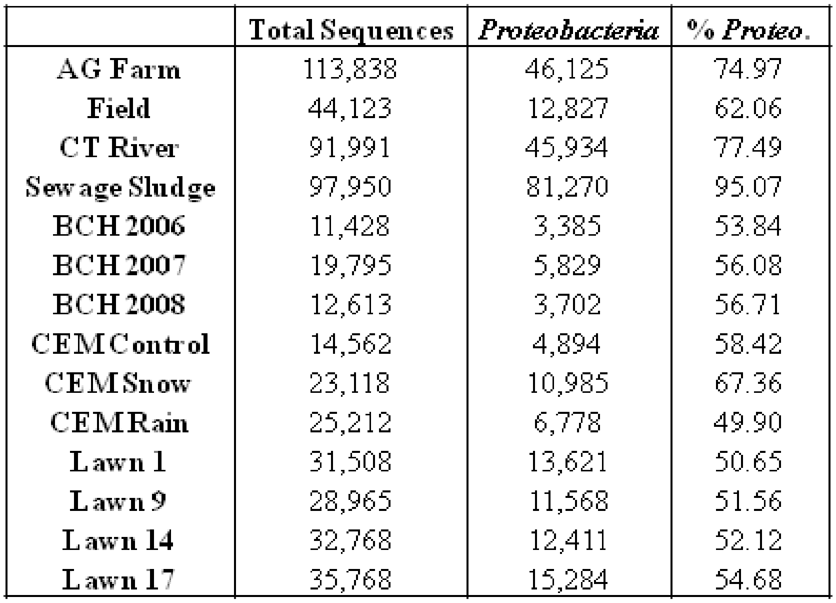
Major and minor bacterial phyla identified by sequence classification using the Ribosomal Database Project. Datasets are listed across the top: AG Farm (AG_F), Field, River, Sludge, Beach Hall 2006 (B_06), Beach Hall 2007 (B_07), Beach Hall 2008 (B_08), Cemetery Control (C_Cnt), Cemetery Rain (C_R), Cemetery Snow (C_S), Lawn 1 (Ln_1), Lawn 9 (Ln_9), Lawn 14 (Ln_14), and Lawn 17 (Ln_17). Row “phylum_NA” includes sequences that were assigned to a phylum but could not be further assigned. Sequences that could not be classified as bacteria, or were too short to classify are listed in the “Unclassified Root” row. The total number of sequence reads for all datasets are found at the bottom of each column.

Table 8 lists the abundance of bacterial sequences according to phylum. Not all known phyla are represented in the soil analyzed. A large portion (50.3% average) of the V6 sequence could not be classified further than either the bacterial kingdom or respective phylum (“phylum_NA” row of Table 8). The Lawn V1-V2 libraries averaged only 18.25% unclassified sequences. This is likely because the amplicon used to generate these libraries was larger than 250 base pairs allowing more accurate classification.

The purpose of such sequence classification is to identify common phyla found in soil. Based on the distribution in Table 8, there are major and minor components to the soil community. An example of a major phylum would be *Proteobacteria*, while a minor phylum would be the *Verrucomicrobia* group. There are other phyla that contain very low (less than 10) sequence matches. Such results can be misleading. For example, in the Beach Hall 2006 library a single *Chlamydiae* sequence is in 11,428 total sequences. The ambiguity of this result comes from the inability to demonstrate whether this match is real or coincidental. A sequencing error could have occurred leading to a coincidental match. Furthermore, no other soil library contained a *Chlamydiae* sequence. Because of this type of ambiguity, only major and minor phyla that are consistently represented in the soil libraries will be considered for analysis. These phyla include: *Proteobacteria, Acidobacteria, Actinobacteria, Bacteriodetes, Verrucomicrobia, Firmicutes, Nitrospira,* and *Cyanobacteria.*

The bacterial phylum with the most representation in all soil libraries is the *Proteobacteria* phylum. Table 9 lists the abundance of *Proteobacteria* sequences in all soils. The total number of *Proteobacteria* listed includes sequences that could only be identified as belonging to the phylum (i.e. no further taxonomic classification). The abundance of *Proteobacteria* in soils makes this phylum a good target for forensic analysis. The development of group-specific DNA typing will require bacterial groups that are common in soils, ensuring that interpretable data can be generated from a variety of soils. However, for specific groups to be forensically informative, there must be sufficient intra-group genetic diversity so that DNA profiles can be distinguishable between soils. Intra-group diversity will be discussed at the end of this chapter.

Table 9 – Abundance of *Proteobacteria* sequences in soil samples



	Total Sequences	*Proteobacteria*	% *Proteo*.
AG Farm	113,838	46,125	74.97
Field	44,123	12,827	62.06
CT River	91,991	45,934	77.49
Sewage Sludge	97,950	81,270	95.07
BCH 2006	11,428	3,385	53.84
BCH 2007	19,795	5,829	56.08
BCH 2008	12,613	3,702	56.71
CEM Control	14,562	4,894	58.42
CEM Snow	23,118	10,985	67.36
CEM Rain	25,212	6,778	49.90
Lawn 1	31,508	13,621	50.65
Lawn 9	28,965	11,568	51.56
Lawn 14	32,768	12,411	52.12
Lawn 17	35,768	15,284	54.68

Total number of high quality sequence reads per sample is listed. Both classified and unclassified *Proteobacteria* sequences are included in the total. Sequence classification was done using the RDP database. Members of the *Proteobacteria* phylum dominate all soils, with nearly 50% or greater representation in the community.

Proteobacteria were not the only identifiable members of the soil community. Figure 11 shows the relative abundance of the other major and minor bacterial phyla. The data show that phyla are not evenly distributed in each community, in agreement with the Evenness values generated in Tables 7A and 7B. Across samples, there are trends: after *Proteobacteria*, the next two largest phyla are either *Acidobacteria, Actinobacteria,* or *Bacteriodetes;* the *Verrucomicrobia, Firmicutes, Nitrospira,* and *Cyanobacteria* phyla each contribute to a lesser extent to the bacterial communities.

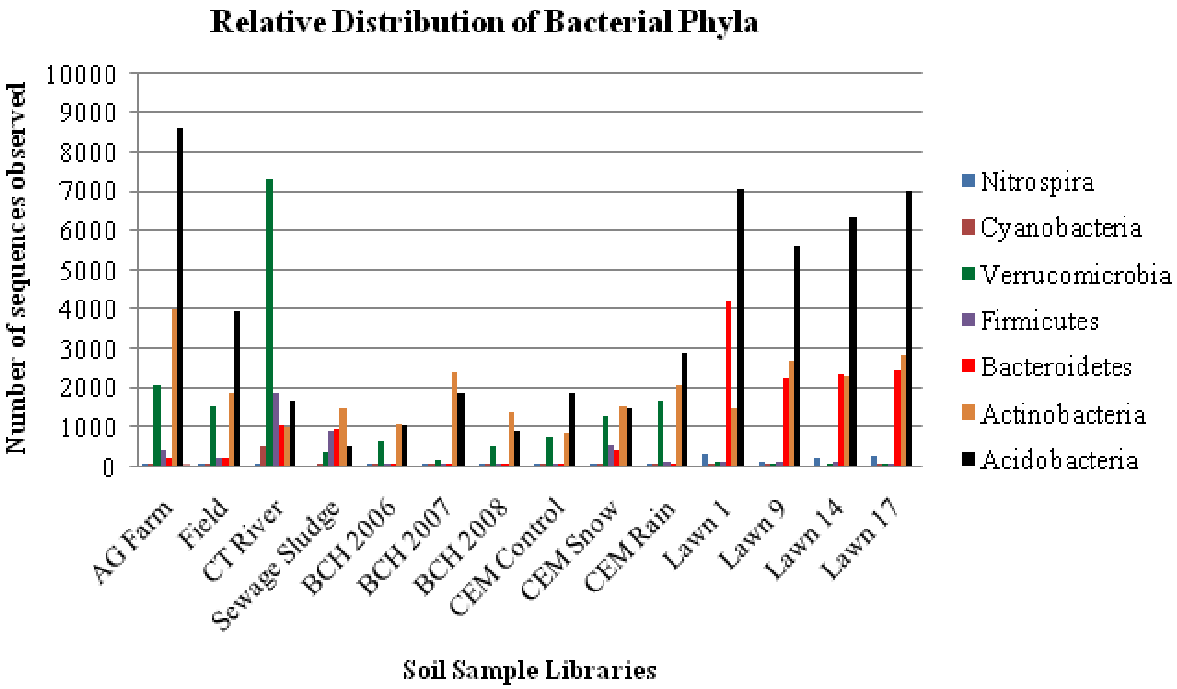


Figure 11 – Distribution of bacterial phyla in soil samples

Bar graph depicting the relative abundance of members of bacterial phyla. Sequences used include classified and unclassified matches. Sequences from *Proteobacteria* are not included so that all other phyla could be easily visualized. Sequence classification was done using the Ribosomal Database Project.

Since bacterial communities are biased towards only a few phyla, typing methods involving universal PCR amplification of the 16S rRNA gene must be analyzed with caution. PCR amplification is a competitive process (in terms of primer access to template), and universal DNA profiles will not be truly ‘universal’; rather they will tend to represent major components of bacterial communities. For forensic soil analysis, amplification of only major bacterial components may be sufficient in differentiating soils. This question will be addressed later on Chapter 4. Similarly, it is important to determine whether these minor components offer any additional potentially informative data. This will also be explored in Chapter 4.

To summarize thus far, 454 pyrosequencing has provided an in-depth look into the components of bacterial communities. Based on the data presented, soil communities are extremely diverse. The way that this diversity is structured within the community is biased toward a few major phyla. Of the twenty two bacterial taxa, eight have been selected as major and minor community components. These phyla will be used to measure the differences and similarities among diverse and similar ecosystems. These phyla will also be explored with regards to meteorological and time change in two soil locations.

II.d. Bacterial community structure in diverse ecosystems

The goal of this chapter is to utilize in-depth surveys to provide rationale for group-specific assay selection. A successful group-specific assay should work on a variety of soils from various ecosystems. Comparing diverse ecosystems will help to identify some potential group-specific targets. Figure 12 shows the relative abundance of the top 8 bacterial phyla (left side). The bar graph on the right side excludes *Proteobacteria* so that the other phyla can be more easily visualized. When the *Proteobacteria* are removed, we observe how the distribution of these phyla is unique to the CT River and Sewage Sludge. Although the AG Farm and Field are not from the same location they are both mineral soils. This may explain why the relative proportions of *Acidobacteria, Actinobacteria, Bacteriodetes* and *Verrucomicrobia* are similar. The relative abundance of *Cyanobacteria* is unique to the CT River sample. Also the CT River and Sewage Sludge samples contain a much high amount of *Bacteroidetes* than the mineral soils. In terms of forensic potential, *Bacteroidetes* and *Cyanobacteria* assays would likely not work with mineral soils.

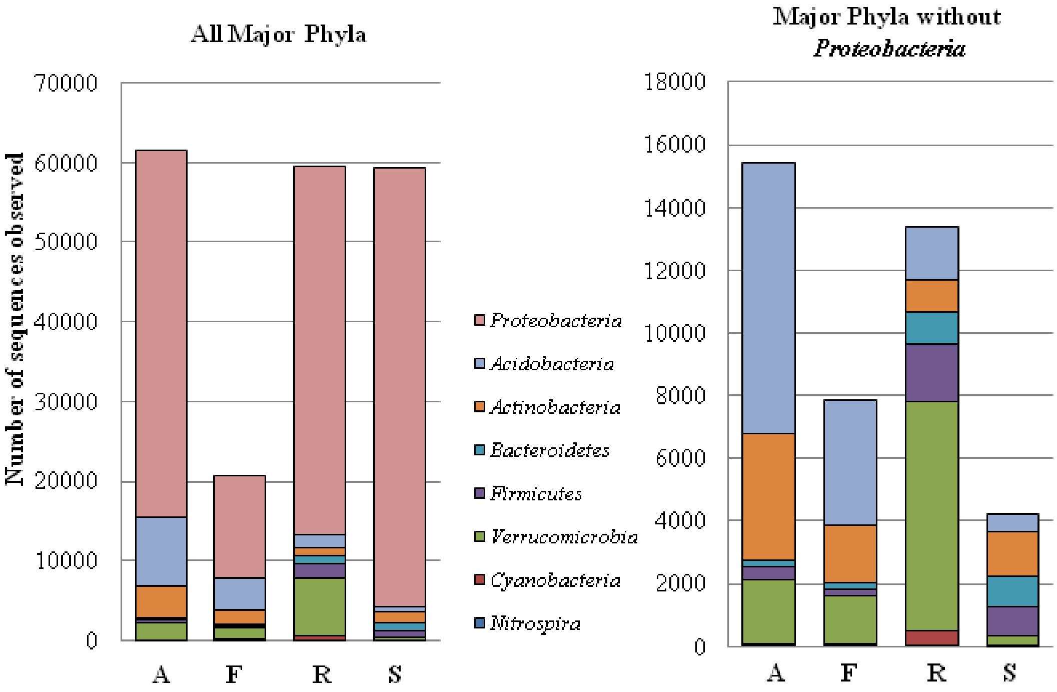


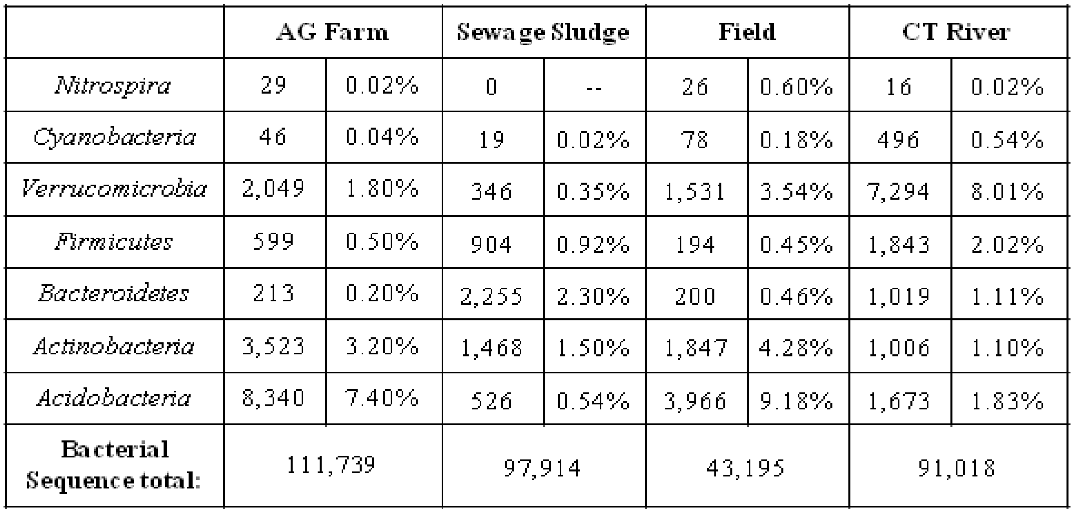
Figure 12 – Relative abundance of bacterial phyla in diverse soils/sediments

X-axis lists diverse soil samples: A – AG Farm; F – Field; R – CT River; S – Sewage Sludge. Bar graph on the left includes all major phyla observed following sequence classification with the Ribosomal Database Project's database. The bar graph on the right does not include members of the *Proteobacteria*.

Table 10 summarizes the percentages of major phyla in diverse ecosystems. *Proteobacteria* were not included so that other phyla could be better explored. While it was clear from the bar graph that there were quantitative differences among the datasets, Table 10 normalizes each phyla's contribution by percentage. Beginning with the AG Farm and Field soils, the distribution of phyla is very similar. One observed difference between these two libraries and the River and Sludge is the low abundance of

Bacteriodetes. There is approximately 10-fold less in these soils compared to Sludge, and approximately 5-fold less than in the River.

Table 10 – Percent composition of select phyla in diverse samples



	AG Farm		Sewage Sludge		Field		CT River	
Nitrospira	29	0.02%	0	--	26	0.60%	16	0.02%
Cyanobacteria	46	0.04%	19	0.02%	78	0.18%	496	0.54%
Verrucomicrobia	2,049	1.80%	346	0.35%	1,531	3.54%	7,294	8.01%
Firmicutes	599	0.50%	904	0.92%	194	0.45%	1,843	2.02%
Bacteroidetes	213	0.20%	2,255	2.30%	200	0.46%	1,019	1.11%
Actinobacteria	3,523	3.20%	1,468	1.50%	1,847	4.28%	1,006	1.10%
Acidobacteria	8,340	7.40%	526	0.54%	3,966	9.18%	1,673	1.83%
Bacterial Sequence total:	111,739		97,914		43,195		91,018	

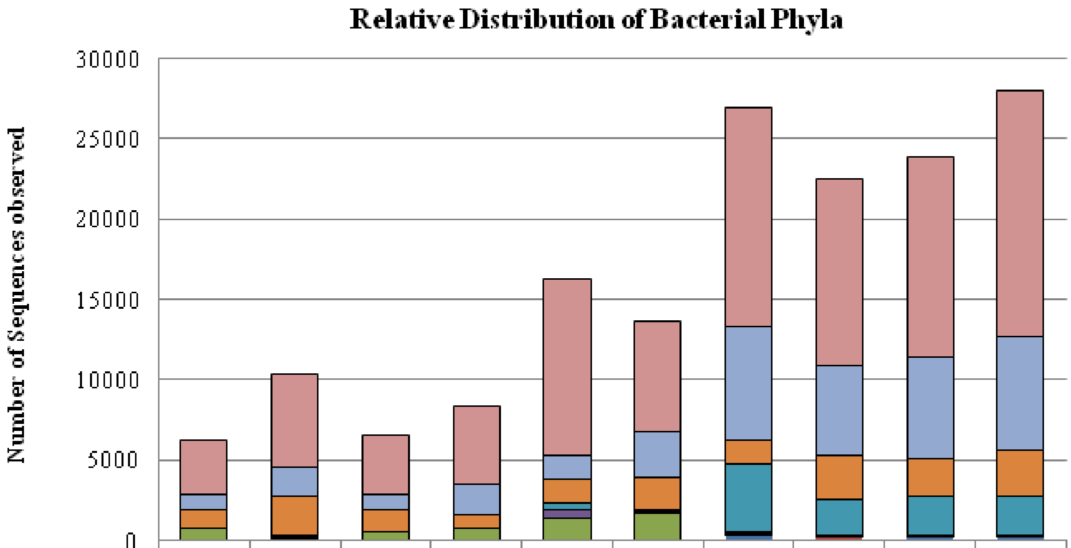
Total sequences and percent composition for select phyla. Sequences included were those both classified and unclassified. Members of the *Proteobacteria* phyla were not included in analysis so that other phyla could be highlighted.

However, if uncommon soil samples (like sludge and freshwater sediment) are submitted as forensic evidence, then there are phyla that can be used to differentiate them from mineral soils. The lack of representation from the *Acidobacteria*, *Verrucomicrobia* and *Nitrospira* phyla may be indicators of most types of Sewage Sludge. There was approximately 16-fold less *Acidobacteria* in the Sludge as compared to mineral soils. This is consistent with the biology of soils, as these three phyla are known habitants of soil. In the CT River sample there was high quantity of both the *Firmicutes* (3-fold greater than mineral soil) and *Cyanobacteria* (5-fold greater than mineral soil) which may be useful indicators of determining whether a sample likely originated from a freshwater ecosystem. Specifically, *Cyanobacteria* are not normally found in mineral soils, but know to be common in marine ecosystems.

II.e. Bacterial community structure in maintained lawns

The 454 data generated for diverse ecosystems demonstrated measureable, quantitative differences in each sample's respective bacterial community. Although *Proteobacteria* was the major contributor in all samples, the presence and absence of other phyla highlighted were noted as possible ecosystem-specific signatures. The question for possible forensic soil analysis is can soils that share more environmental characteristics be differentiated? The remainders of samples surveyed in this study were all classified as mineral soils from maintained lawn areas. The first task is to determine the relative abundance of phyla in each of these samples. This data will be able to show whether there are quantitative similarities among mineral soils that come from different locations. Also, these data will further establish possible candidate phyla for identifying mineral soils.

The bar graph in Figure 13 illustrates the relative distribution of the top 8 bacterial phyla identified from these similar ecosystems; below a table lists absolute sequence counts. The samples are clustered by set: Beach Hall samples, Cemetery samples, and Lawn samples. The data show that all soils are dominated by *Proteobacteria*, *Acidobacteria,* and *Actinobacteria,* supporting the data generated from the AG Farm and Field soil. These three phyla, specifically *Acidobacteria* and *Actinobacteria*, may be signature phyla for mineral soils. Fierer *at al.* also identified these phyla in forest, desert and prairie soils [Fierer *et al.*, 2005]. This has important forensic ramifications as this data shows promise for use of group-specific typing across many ecosystems.



	BCH 2006	BCH 2007	BCH 2008	CEM Control	CEM Snow	CEM Rain	Lawn 1	Lawn 9	Lawn 14	Lawn 17
■Proteobacteria	3385	5829	3702	4894	10985	6778	13621	11568	12411	15284
■Acidobacteria	1041	1852	870	1857	1489	2864	7068	5584	6336	7026
■Actinobacteria	1097	2387	1364	834	1520	2078	1482	2712	2312	2846
■Bacteroidetes	33	7	15	9	410	36	4174	2246	2373	2425
■Firmicutes	28	46	18	31	570	105	136	126	133	86
■Verrucomicrobia	670	180	521	727	1280	1680	114	23	16	28
■Cyanobacteria	1	2	4	13	45	2	3	63	0	3
■Nitrospira	32	91	34	12	8	41	296	116	230	254

Figure 13 – Distribution of major bacterial phyla in various maintained lawn soils

Bar graph depicts the relative abundance of phyla with absolute sequence counts below. Absolute sequence counts include classified and unclassified sequence reads. DNA libraries from all Beach Hall (BCH) and Cemetery (CEM) soils were created from V6 region amplicons. Using the RDP, a 60% confidence threshold was used for taxonomy classification for V6 libraries. DNA libraries from lawn samples were generated from V1 –V2 regions. An 80% confidence threshold was used to classify these sequences.

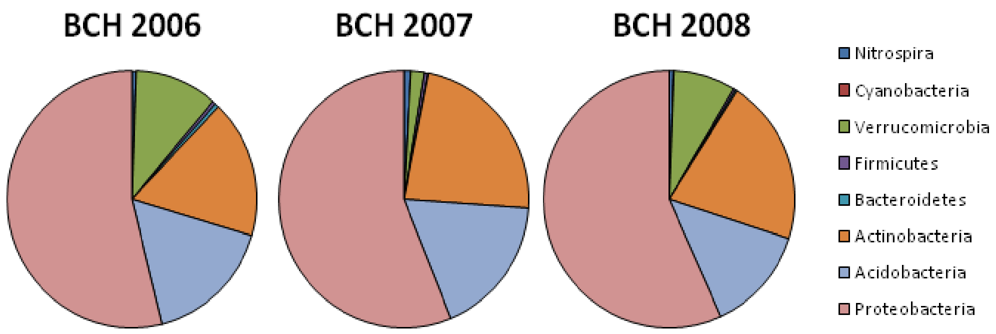
At a superficial level, there are libraries that appear to cluster; for example, the phyla abundance in both BCH 2006 and BCH 2008 are very similar. Also, the Lawn samples share similarity in abundance and display distinct composition compared to the other organic soils. Specifically, the *Bacteroidetes* phylum has strong representation in the Lawns. This data negates the previous statement that the *Bacteroidetes* phylum may not be a good candidate for mineral soils. A more appropriate statement would be that *Bacteroidetes* has the potential to differentiate some mineral soils. Since the soil libraries contain different amounts of sequence data, percent composition will be used to compare bacterial communities within the Beach Hall, Cemetery and Lawn sets. Any major outliers will be selected for further analysis.

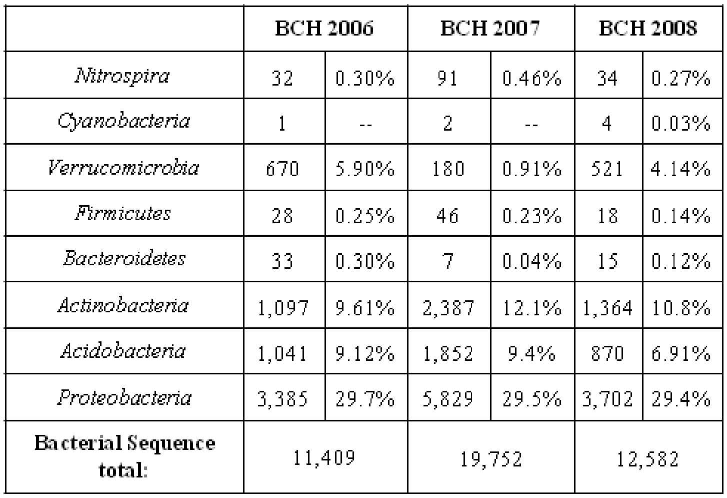
Thus far, 454 data has highlighted quantitative differences among a variety of soil samples. Common phyla to mineral soils have been identified, providing rationale for group-specific assay development. In the next section, annual community fluctuation is explored to determine if major phyla change over the course of time. These data will help to identify any phyla that are robust enough to withstand change. Bacterial phyla that remain consistent over time are desirable candidates for group-specific assays because data will be more consistent.

II.f. Bacterial community fluctuation from year-to-year

As discussed in chapter 2, time can influence bacterial community structure, as was evident from soils collected from the same location over two years. Although universal T-RFLP analysis performed better than C-RFLP by correctly identifying more soils collected from the same locations over the course of one year, neither method was consistently reliable. Using more comprehensive 454 pyrosequencing, the goal of this experiment was to determine how much change occurs over a period of one year. If bacterial communities do change, can the variable phyla be identified? Lastly, how might such change impact the ability to use soil as associative evidence?

Figure 14 depicts the bacterial phyla distributions for Beach Hall 2006, 2007 and 2008 soils, including a table list of the relative sequence abundance and percent composition in the soil library. Absolute sequence totals for each phyla include classified and unclassified sequences. [Note: Figures 15-19 also follow this format].





	BCH 2006		BCH 2007		BCH 2008	
Nitrospira	32	0.30%	91	0.46%	34	0.27%
Cyanobacteria	1	--	2	--	4	0.03%
Verrucomicrobia	670	5.90%	180	0.91%	521	4.14%
Firmicutes	28	0.25%	46	0.23%	18	0.14%
Bacteroidetes	33	0.30%	7	0.04%	15	0.12%
Actinobacteria	1,097	9.61%	2,387	12.1%	1,364	10.8%
Acidobacteria	1,041	9.12%	1,852	9.4%	870	6.91%
Proteobacteria	3,385	29.7%	5,829	29.5%	3,702	29.4%
Bacterial Sequence total:	11,409		19,752		12,582	

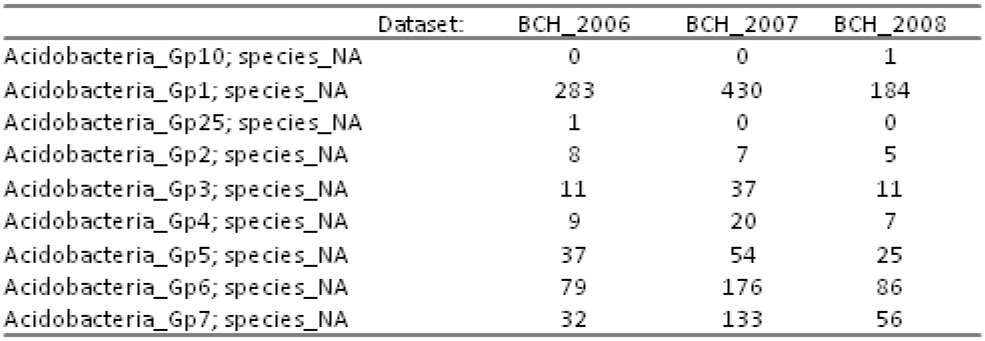
Figure 14 – Distribution of major bacterial phyla in year to year sampling

Pie charts represent the relative distribution of bacterial sequences in Beach Hall soils collected in years 2006, 2007 and 2008. Table lists absolute sequence counts which include both classified and unclassified sequences. Classification was done using the Ribosomal Database project. Percentage calculations were based on the total number of bacterial sequences generated from each library.

When percent composition is used to compare phyla, the data show that bacterial abundance in most phyla remains relatively consistent from year to year. There are two instances of noticeable change, however. In BCH 2007, only 0.91% of the soil library was *Verrucomicrobia*, whereas this phylum comprised 5.90% and 4.14% in the BCH 2006 and BCH 2008 libraries, respectively. Change in abundance can be due to a variety of environmental factors like decreased moisture content [Buckley *et al.*, 2001]. This is a plausible explanation since these soils were sampled during the middle of summer in July. There is also a slight drop in the abundance of *Acidobacteria* in BCH 2008. While fluctuations in soil pH may play a role in altering *Acidobacteria* abundance, without pH data we cannot say for certain [Jones *et al.*, 2009].

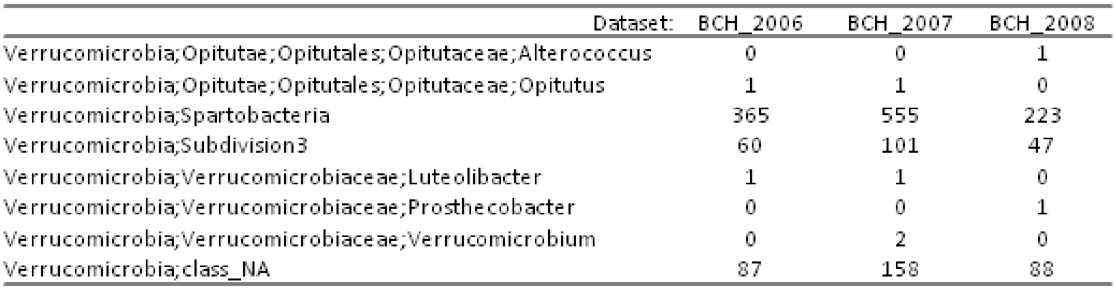
The data show slight changes in the relative abundance of bacterial phyla members in different years. It is important to address whether this change is strictly quantitative or qualitative, specifically in the *Verrucomicrobia* and *Acidobacteria* phyla. In terms of forensic potential, bacterial phyla that only change in quantity are more desirable than groups that change in richness. A modification to richness may result in changes to a DNA profile. Table 11A lists the *Acidobacteria* data in Beach Hall annual soil libraries which show that there is similar representation of *Acidobacteria* classes from year-to-year. The same is true for the *Verrucomicrobia* data in Table 11B. The data show that these two phyla change in quantity annually; species richness remains consistent. This is positive information for possible forensic applications.

Table 11A – Classification of *Acidobacteria* from Beach Hall libraries



Dataset:	BCH_2006	BCH_2007	BCH_2008
Acidobacteria_Gp10; species_NA	0	0	1
Acidobacteria_Gp1; species_NA	283	430	184
Acidobacteria_Gp25; species_NA	1	0	0
Acidobacteria_Gp2; species_NA	8	7	5
Acidobacteria_Gp3; species_NA	11	37	11
Acidobacteria_Gp4; species_NA	9	20	7
Acidobacteria_Gp5; species_NA	37	54	25
Acidobacteria_Gp6; species_NA	79	176	86
Acidobacteria_Gp7; species_NA	32	133	56

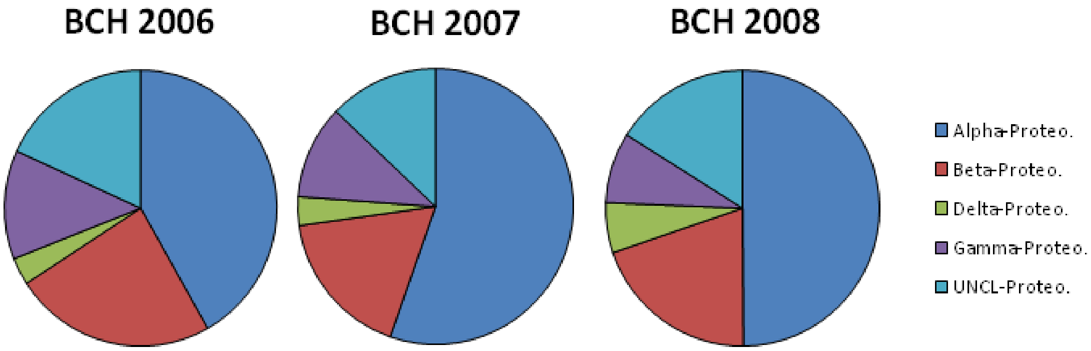
Table 11B – Classification of select *Verrucomicrobia* from Beach Hall libraries



Dataset:	BCH_2006	BCH_2007	BCH_2008
Verrucomicrobia;Opitutae;Opitutales;Opitutaceae;Alterococcus	0	0	1
Verrucomicrobia;Opitutae;Opitutales;Opitutaceae;Opitutus	1	1	0
Verrucomicrobia;Spartobacteria	365	555	223
Verrucomicrobia;Subdivision3	60	101	47
Verrucomicrobia;Verrucomicrobiaceae;Luteolibacter	1	1	0
Verrucomicrobia;Verrucomicrobiaceae;Prosthecobacter	0	0	1
Verrucomicrobia;Verrucomicrobiaceae;Verrucomicrobium	0	2	0
Verrucomicrobia;class_NA	87	158	88

Taxonomy classification for sequences in Beach Hall annual libraries. Classification was done using the Visualization and Analysis of Microbial Population Structures (VAMPS) database (http://vamps.mbl.edu/index.php). Table A lists all *Acidobacteria* classes. Order, family, genus and species classification is not available for *Acidobacteria.* Table B lists several *Verrucomicrobia* classifications. Each sequence category contains taxonomy classification beginning with the *Verrucomicrobia* phylum, followed by class, order, family, genus and species information where available.

Figure 15 shows community data on the *Proteobacteria*. This phylum is very robust, retaining a consistent abundance among classes from year-to-year. All meteorological and seasonal events that took place during the course of one year did not seem to dramatically affect the *Proteobacteria* phylum distribution. These data further supports its potential as a forensically informative group given its prevalence in mineral soils and demonstrated robustness.



	BCH 2006		**BCH 2007**		**BCH 2008**	
Alphaproteobacteria	1,417	41.9%	3,228	55.4%	1,848	49.9%
Betaproteobacteria	810	23.9%	1,016	17.4%	735	19.85%
Deltaproteobacteria	109	3.2%	195	3.35%	218	5.9%
Gammaproteobacteria	434	12.8%	637	10.9%	306	8.3%
Unclassified	615	18.2%	753	12.9%	595	16.1%
***Proteobacteria* Sequence total:**	3,385		5,829		3,702	

Figure 15 – Distribution of *Proteobacteria* in year to year sampling

Pie charts represent the relative distribution of *Proteobacteria* sequences in Beach Hall soils collected in years 2006, 2007 and 2008. Table lists absolute sequence counts for each class. Classification was done using the Ribosomal Database project. Percentage calculations were based on the total number of *Proteobacteria* sequences generated from each library.

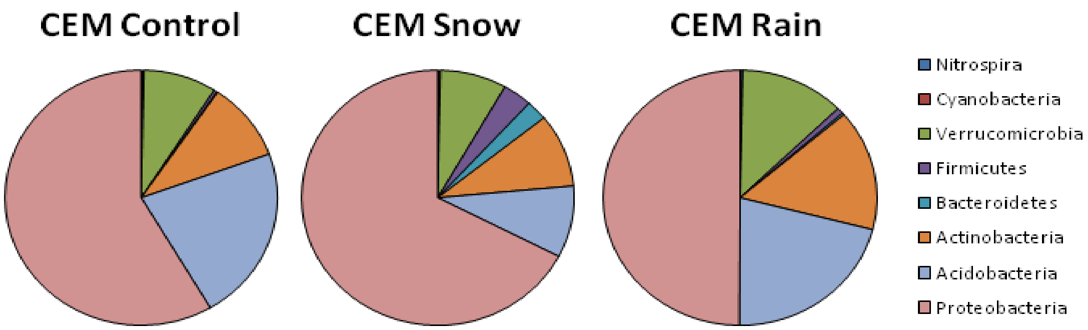
II.g. Bacterial community fluctuation after meteorological events

Bacterial communities are sensitive to environmental variations environment such as temperature change [Lipson *et al.*, 2004; Smit *et al.*, 2001; Walker *et al.*, 2006]. As a result, microbes have adapted survival mechanisms to guarantee viability even in times when growth conditions are not optimal. An example of this is that members of a soil bacterial community that experienced freeze-thaw cycles adapted to withstand damage from ice crystals [Walker *et al.*, 2006]. Such genetic modifications allow bacterial communities to retain a structural balance over time. This adaptation is also critical to the prosperity of the ecosystem, as microorganisms are known to play key roles in the maintenance of the soil ecosystem [Nannipieri *et al.*, 2003].

As previously mentioned, for soil to be useful as forensic evidence the bacterial communities native to a given location must not significantly change over time. A significant change to richness may alter the native DNA profile. If bacterial communities fail to maintain structural equilibrium then associating a soil sample to a location would never be possible. The data in the previous section demonstrated that although some quantitative changes occur from year to year, the overall richness remained consistent. In this section the community in one location is examined after meteorological events to explore how bacterial phyla respond to environmental variables.

Figure 16 shows community data for the top 8 bacterial phyla in the Cemetery meteorological event libraries. It is clear that there are some differences in the snow sample as compared to the other two: *Firmicutes* and *Bacteroidetes* phyla are increased. Member of the *Firmicutes* phylum are gram positive, so their increase in abundance in the snow sample may actually result from ice crystals breaking open more cells. It is unclear why *Bacteroidetes* increased in abundance. Whether the increase of these two phyla are a result of physical disruption exposing more cells, or a direct biological adaptation to cold environment, the data show that snow cover changes community structure. This is significant to forensics because it speaks to the importance of timely sample collection. Furthermore, it demonstrates that group-specific assays done on samples collected from the same location may generate misleading results. This information is useful for interpretation of both universal and group-specific DNA profiles.

These data also show that rain does not drastically change the bacterial community in the Cemetery soil. This is equally important to forensics because it demonstrates that not all environmental variables will negatively impact bacterial community structure. Furthermore, this experiment includes the variable of time. Three weeks separated the collection of control and rain samples. The community structure within Cemetery soil was not affected by time or heavy rain.

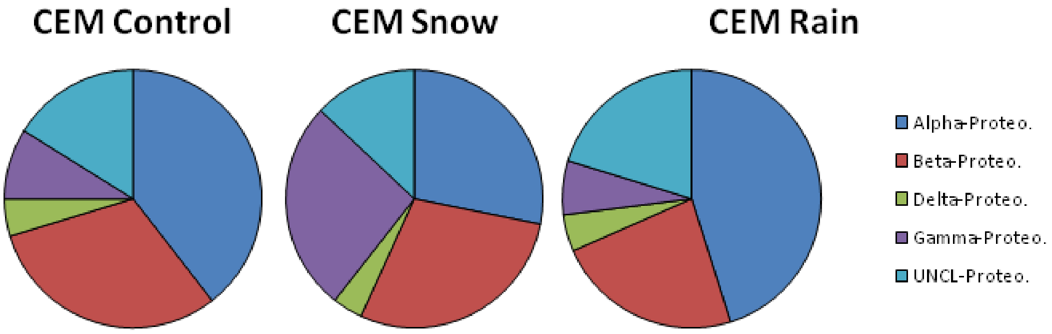


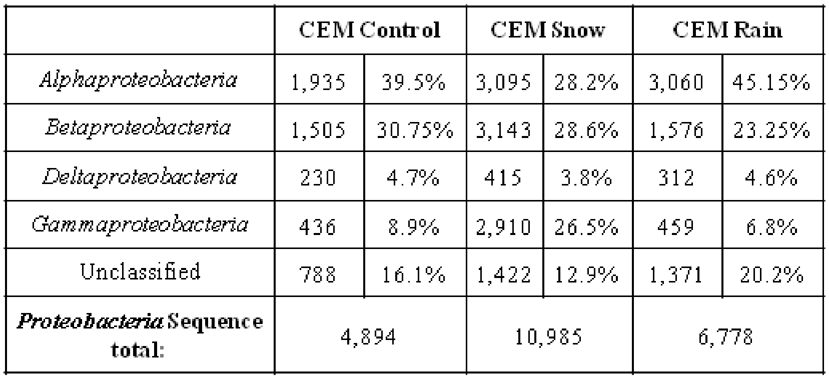
	CEM Control		CEM Snow		CEM Rain	
Nitrospira	12	0.08%	8	0.03%	41	0.16%
Cyanobacteria	13	0.09%	45	0.19%	2	--
Verrucomicrobia	727	5.00%	1,280	5.55%	1,680	6.67%
Firmicutes	31	0.21%	570	2.47%	105	0.42%
Bacteroidetes	9	0.06%	410	1.78%	36	0.14%
Actinobacteria	834	5.74%	1,520	6.60%	2,078	8.25%
Acidobacteria	1,857	12.8%	1,489	6.46%	2,864	11.38%
Proteobacteria	4,894	33.7%	10,985	47.7%	6,778	26.9%
Bacterial Sequence total:	14,530		23,039		25,174	

Figure 16 – Distribution of major bacterial phyla in meteorological event sampling

Pie charts represent the relative distribution of bacterial sequences in Cemetery soils from the control, rain and snow dataset. Table lists absolute sequence counts which include both classified and unclassified sequences. Classification was done using the Ribosomal Database project. Percentage calculations were based on the total number of bacterial sequences generated from each library.

Next, the *Proteobacteria* phylum was examined to determine if snow cover and rain changed the distribution of its members. In the previous section, *Proteobacteria* was determined to be resistant to drastic change over time. As shown in Figure 17, the *Gammaproteobacteria* class has a 3.5-fold average increase in abundance in the snow library as compared to the control and rain libraries.





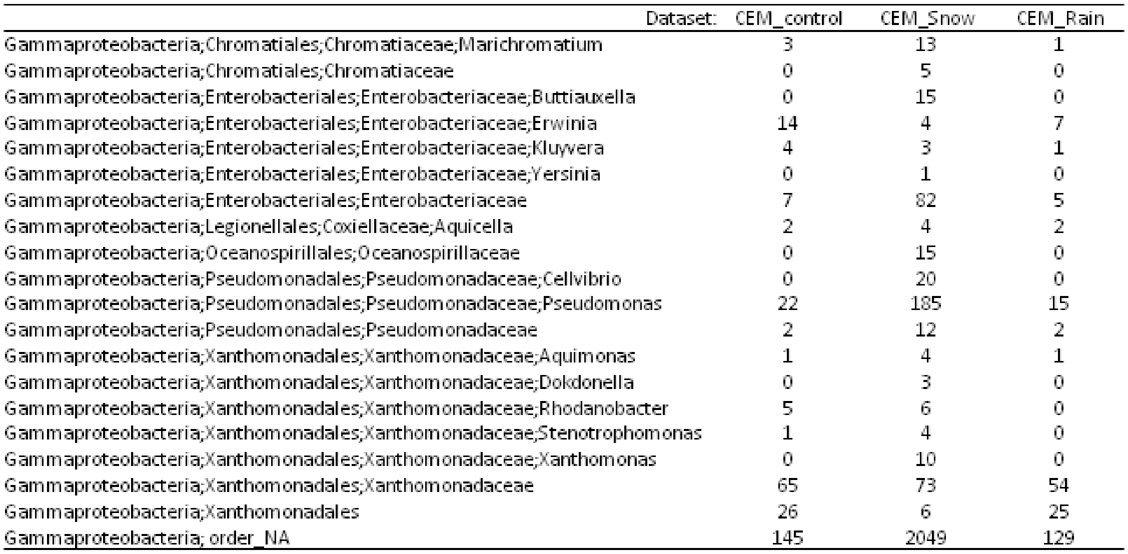
	CEM Control		CEM Snow		CEM Rain	
Alphaproteobacteria	1,935	39.5%	3,095	28.2%	3,060	45.15%
Betaproteobacteria	1,505	30.75%	3,143	28.6%	1,576	23.25%
Deltaproteobacteria	230	4.7%	415	3.8%	312	4.6%
Gammaproteobacteria	436	8.9%	2,910	26.5%	459	6.8%
Unclassified	788	16.1%	1,422	12.9%	1,371	20.2%
***Proteobacteria* Sequence total:**	4,894		10,985		6,778	

Figure 17 – Distribution of *Proteobacteria* in meteorological event sampling

Pie charts represent the relative distribution of *Proteobacteria* sequences in Cemetery control, rain and snow datasets. Table lists absolute sequence counts for each class. Classification was done using the Ribosomal Database project. Percentage calculations were based on the total number of *Proteobacteria* sequences generated from each library.

To ensure that the increase in *Gammaproteobacteria* was not accompanied by new species, taxonomic classification was compared between the three datasets. Table 12 highlights some of the sequence classification. The largest increase in representation in the snow library came from the *Enterobacteriales* and *Pseudomonadales* orders. In general, all of the taxa found in the snow library were also found in the control and rain datasets. This demonstrates that this change is quantitative, a desirable characteristic for group-specific forensic applications.

Table 12 – Species classification of select *Gammaproteobacteria* from Cemetery libraries



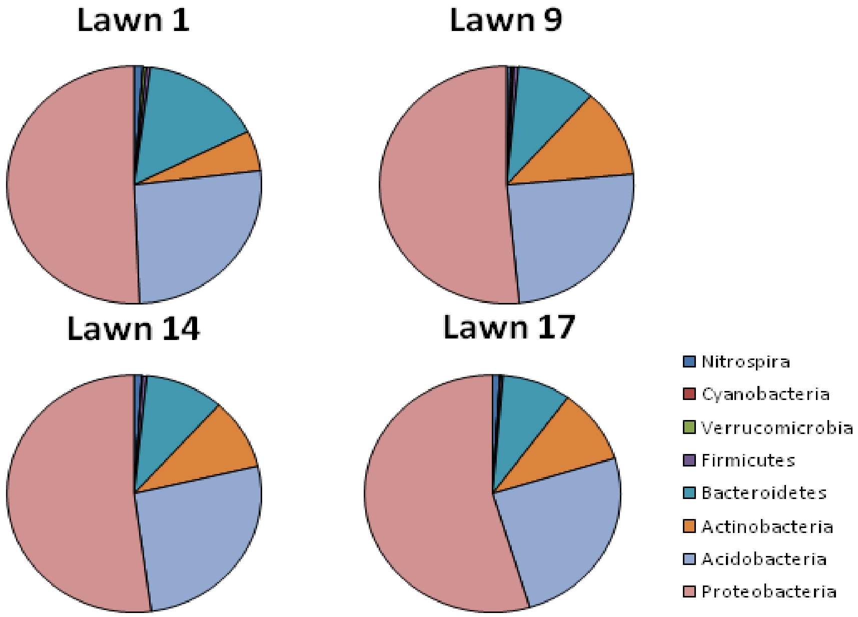
Dataset:	CEM_control	CEM_Snow	CEM_Rain
Gammaproteobacteria;Chromatiales;Chromatiaceae;Marichromatium	3	13	1
Gammaproteobacteria;Chromatiales;Chromatiaceae	0	5	0
Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Buttiauxella	0	15	0
Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Erwinia	14	4	7
Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Kluyvera	4	3	1
Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Yersinia	0	1	0
Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae	7	82	5
Gammaproteobacteria;Legionellales;Coxiellaceae;Aquicella	2	4	2
Gammaproteobacteria;Oceanospirillales;Oceanospirillaceae	0	15	0
Gammaproteobacteria;Pseudomonadales;Pseudomonadaceae;Cellvibrio	0	20	0
Gammaproteobacteria;Pseudomonadales;Pseudomonadaceae;Pseudomonas	22	185	15
Gammaproteobacteria;Pseudomonadales;Pseudomonadaceae	2	12	2
Gammaproteobacteria;Xanthomonadales;Xanthomonadaceae;Aquimonas	1	4	1
Gammaproteobacteria;Xanthomonadales;Xanthomonadaceae;Dokdonella	0	3	0
Gammaproteobacteria;Xanthomonadales;Xanthomonadaceae;Rhodanobacter	5	6	0
Gammaproteobacteria;Xanthomonadales;Xanthomonadaceae;Stenotrophomonas	1	4	0
Gammaproteobacteria;Xanthomonadales;Xanthomonadaceae;Xanthomonas	0	10	0
Gammaproteobacteria;Xanthomonadales;Xanthomonadaceae	65	73	54
Gammaproteobacteria;Xanthomonadales	26	6	25
Gammaproteobacteria; order_NA	145	2049	129

Taxonomy classification for sequences in Cemetery meteorological event libraries. Classification was done using the Visualization and Analysis of Microbial Population Structures (VAMPS) database (http://vamps.mbl.edu/index.php). Each classification begins with the *Gammaproteobacteria* class, followed by order, family, and genus information where available.

II.h. Presumed biological replicate soils

In instances where the origin of a soil sample is unknown, it would be useful to be able to identify potential ecosystems based on the bacterial community structure. Specific soil types influence which bacteria are present in the native community [Hackl *et al.*, 2004; Givran *et al.*, 2003; Louzoupne *et al.*, 2007; Nanniperieri *et al.*, 2003], so it is reasonable to ask if locations that look the same (superficially) contain similar bacterial communities. By surveying maintained lawns from park sites from different locations it is possible to determine whether there are bacterial signatures that characterize this ecosystem.

The first step is to determine whether the lawn soils collected contain similarly structured bacterial communities. As a reminder, these maintained lawns were collected on the same day in November from community parks. Soils were sampled ten feet away from a tree/shrub border. Figure 18 shows data from Lawn samples 1, 9, 14 and 17. The pie charts illustrate a very high similarity among all 4 lawns. The only major difference is seen in Lawn 9, where there are approximately 20-fold more *Cyanobacteria* sequences. Identifying 63 sequences strongly suggests that its presence in the library is real and not artifact (see Table 13 for taxonomy classification). *Cyanobacteria* can be found in damp soils, but are primarily found in aquatic environments. An increase in *Cyanobacteria* was seen in the CT River library. The presence of *Cyanobacteria* in the Lawn 9 library may be evidence of a nearby water source but at the time of collection no water sources were observed in the area.

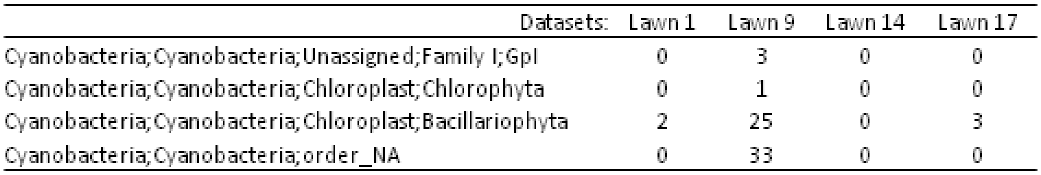


	Lawn 1		**Lawn 9**		**Lawn 14**		**Lawn 17**	
Nitrospira	296	0.94%	116	0.40%	230	0.71%	254	0.72%
Cyanobacteria	3	0.01%	63	0.22%	0	--	3	0.01%
Verrucomicrobia	114	0.36%	23	0.80%	16	0.05%	28	0.08%
Firmicutes	136	0.43%	126	0.44%	133	0.41%	86	0.24%
Bacteroidetes	4,174	13.22%	2,246	7.82%	2,373	7.37%	2,425	6.84%
Actinobacteria	1,482	4.73%	2,712	9.44%	2,312	7.18%	2,846	8.03%
Acidobacteria	7,068	22.5%	5,584	19.4%	6,336	19.7%	7,026	19.8%
Proteobacteria	13,621	43.4%	11,568	40.3%	12,411	38.5%	15,284	43.1%
Bacterial Sequence total:	31,355		28,725		32,198		35,457	

Figure 18 – Distribution of major bacterial phyla in Maintained Lawn samples

Pie charts represent the relative distribution of bacterial sequences in Lawn soils 1, 9, 14 and 17. Table lists absolute sequence counts which include both classified and unclassified sequences. Classification was done using the Ribosomal Database project. Percentage calculations were based on the total number of bacterial sequences generated from each library.

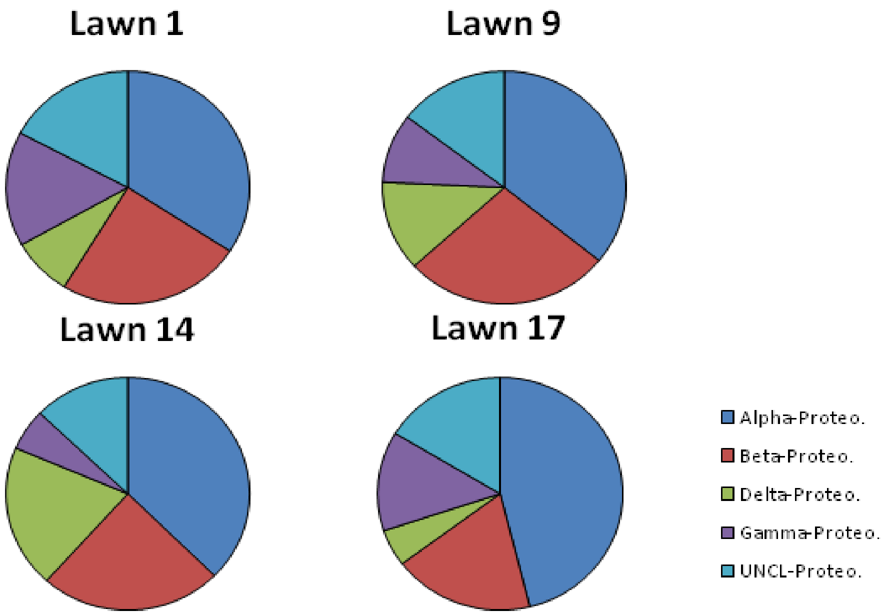
Table 13 – Classification of *Cyanobacteria* from Lawn libraries

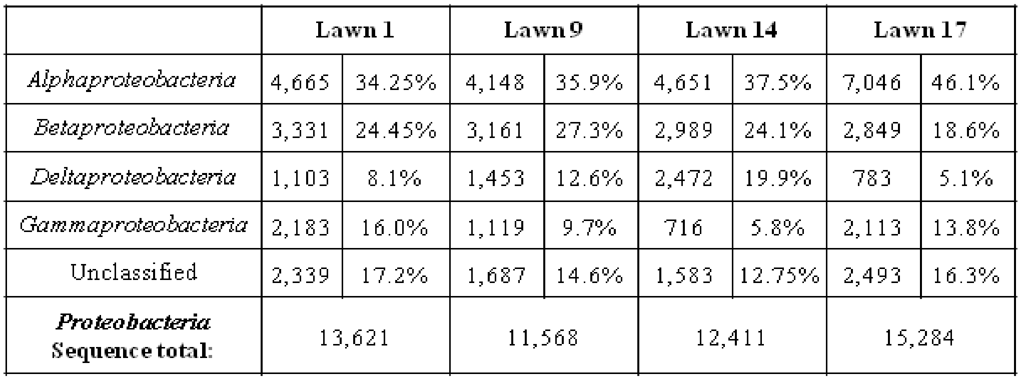


Datasets:	Lawn 1	Lawn 9	Lawn 14	Lawn 17
Cyanobacteria;Cyanobacteria;Unassigned;Family I;GpI	0	3	0	0
Cyanobacteria;Cyanobacteria;Chloroplast;Chlorophyta	0	1	0	0
Cyanobacteria;Cyanobacteria;Chloroplast;Bacillariophyta	2	25	0	3
Cyanobacteria;Cyanobacteria;order_NA	0	33	0	0

Taxonomy classification for sequences in Lawn libraries. Classification was done using the Visualization and Analysis of Microbial Population Structures (VAMPS) database (http://vamps.mbl.edu/index.php) as well as Ribosomal Database Project . Classification begins with the *Cyanobacteria* phylum, followed by class, order, family and genus information when available.

The distributions of *Proteobacteria* (Figure 19) are also very similar among these four libraries, although there are slight differences. Specifically, the *Deltaproteobacteria* class comprises almost 20% of the *Proteobacteria* in Lawn 14 (average representation in the other libraries is 8.6%). In the same library, the *Gammaproteobacteria* representation is lower than the average among the other three (5.8% abundance as compared to the average 13.2%). This data is consistent with the previous *Proteobacteria* data demonstrating their prevalence in mineral soils. The additional point made from the lawn data is that all major phyla have impressively similar bacterial community structures.





	Lawn 1		**Lawn 9**		**Lawn 14**		**Lawn 17**	
Alphaproteobacteria	4,665	34.25%	4,148	35.9%	4,651	37.5%	7,046	46.1%
Betaproteobacteria	3,331	24.45%	3,161	27.3%	2,989	24.1%	2,849	18.6%
Deltaproteobacteria	1,103	8.1%	1,453	12.6%	2,472	19.9%	783	5.1%
Gammaproteobacteria	2,183	16.0%	1,119	9.7%	716	5.8%	2,113	13.8%
Unclassified	2,339	17.2%	1,687	14.6%	1,583	12.75%	2,493	16.3%
Proteobacteria **Sequence total:**	13,621		11,568		12,411		15,284	

Figure 19 – Distribution of *Proteobacteria* in maintained lawn samples

Pie charts represent the relative distribution of *Proteobacteria* sequences in Lawn soils 1, 9, 14 and 17. Table lists absolute sequence counts for each class. Classification was done using the Ribosomal Database project. Percentage calculations were based on the total number of *Proteobacteria* sequences generated from each library.

There are three forensically relevant questions that can be addressed from this data. First, do what we refer to as biological replicates look the same based on community structure? Second, are there sufficient differences within these biological replicates to differentiate them from one another? Third, are the biological replicate lawn libraries measurably different than the other libraries generated in this chapter? The first question can be answered with the data presented in Figure 18. The pie charts demonstrate a remarkable similarity among the unrelated lawn locations. The second question can be answered with data presented in chapter 2. Universal DNA typing was able to differentiate a majority of the lawn soils. So, although community structure is quantitatively similar, there are enough qualitative differences present to distinguish them. To answer the third question, bacterial diversity among all libraries using four different community similarity indices was compared (Figure 20) and portrayed as a heat map.

The similarity measures used are listed above the respective heat map. Each heat map represents the diversity of the communities differently because each index takes into account different community variables. The Morisita-Horn index uses information about the number of unique species found as well as abundance to determine similarity. The Bray-Curtis index measures dissimilarity by comparing the number of unique species to total richness. The Jaccard index calculates the ratio of shared species to the total number of species between two communities and therefore does not account for any species that are absent in one community. The Yue-Clayton index also calculates the ratio of shared to total species but is assigns greater significance for shared species that are most abundant in the community. Each index calculates distance measures across a range of 0 (most similar, blue) to 1 (most different, red). The range of values is coded across a color scale beginning with blue →light blue →white →pink →red.

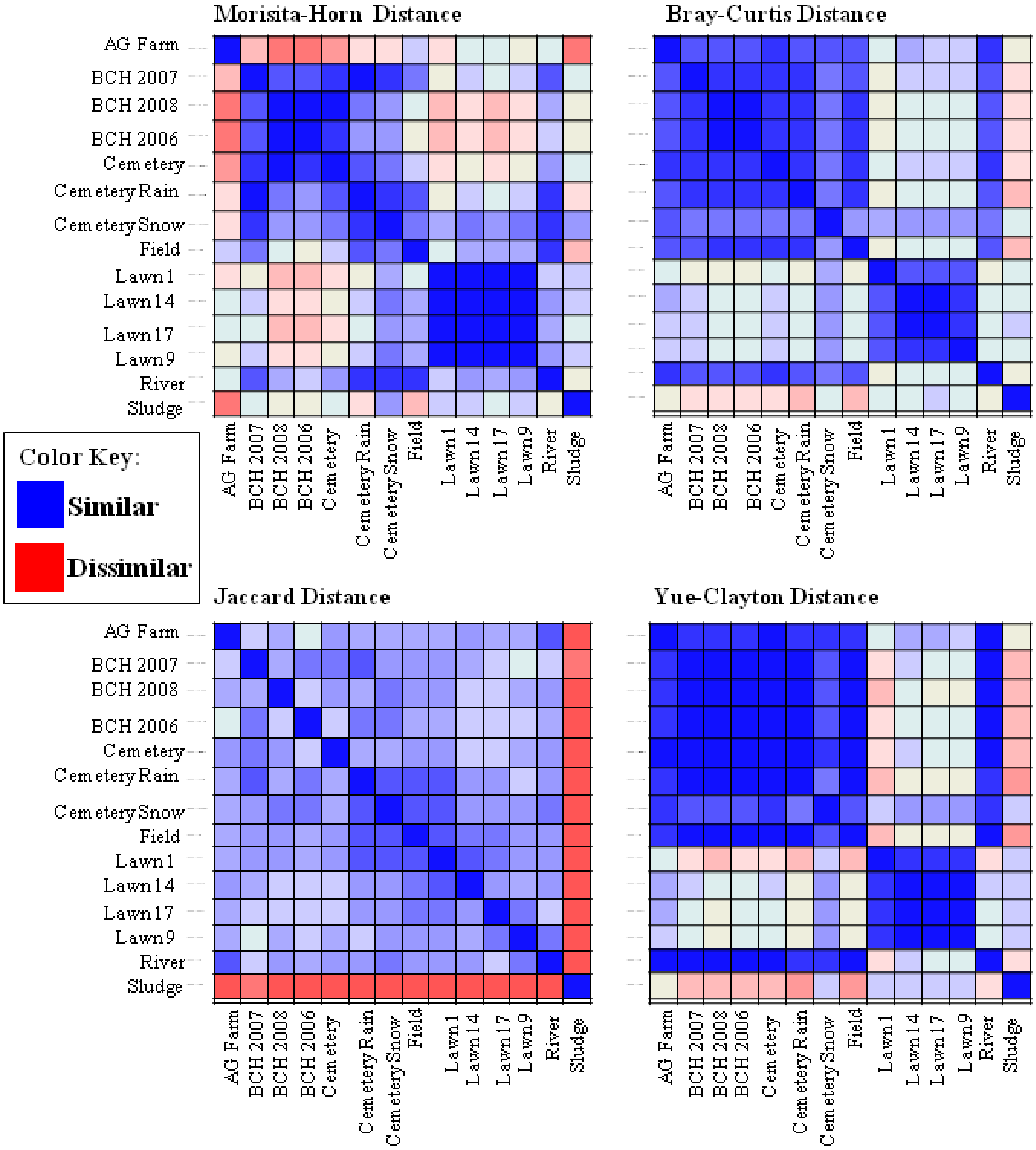


Figure 20 – Community heat maps using various distance measure calculations

Four heat maps are shown comparing all soil libraries. Each heat map was generated by the community similarity index listed above the map. Heat maps were generated using the Visualization and Analysis of Microbial Populations Structures (VAMPS) (http://vamps.mbl.edu/index.php)

The presence, absence, richness, abundance and evenness of bacterial communities are differently represented by each diversity measure. Therefore, it is difficult to say which measure will accurately represent the bacterial community. Because this application has never been explored for forensic potential, all four measures were evaluated. The context that this analysis can be used in forensics would be to identify a location or ecosystem of an unknown sample. A measure that would have the most forensic potential would be one that creates an association between like samples.

The Jaccard Index does not appear to be forensically informative because it does not cluster any of the known locations together. This is likely because the calculation does not account for species that are unique to one location. However, the remaining three indices do produce potentially forensically informative data. The Bray-Curtis and Yue-Clayton diversity measures provide similar information on bacterial diversity. Both cluster the AG Farm, Beach Hall, Cemetery, and Field mineral soils together. Each measure also finds the CT River sediment to be similar in community structure. Based on the community data, the biological replicate lawn samples cluster very tightly and can be differentiated from all other samples. This suggests that bacterial signatures can be used to identify potential ecosystems if reference sample are unavailable.

The data provided by the Morisita-Horn calculation is the most stringent, making it a desirable forensic measure. Using this measure, the AG Farm bacterial community stands apart from all other samples, as does the sewage sludge. Even the Beach Hall and Cemetery samples do not cluster as tightly as they do with the other measures. However, the Lawn samples remain highly similar, further validating the potential for using community diversity measures to suggest an ecosystem of origin.

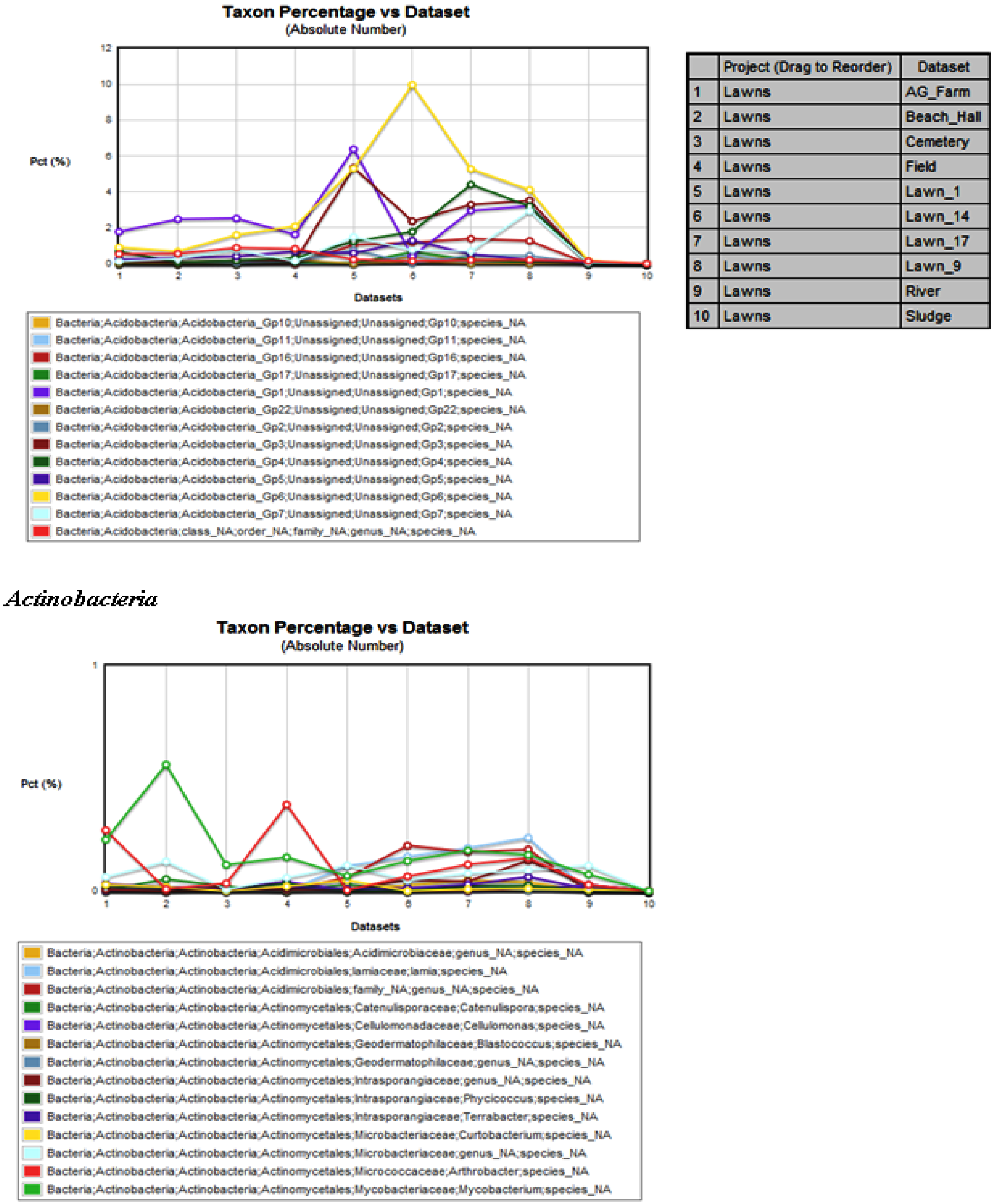
The 454 data presented thus far has provided a detailed survey of several bacterial communities. It is clear that there are quantitative and qualitative differences among bacterial communities. The presence of common phyla in mineral soils is positive to forensics because it gives feasibility to group-specific assays. Heat map data demonstrates that there are qualitative inter-group differences throughout the bacterial community. The key to a successful forensic application will be to find and exploit these differences for the purpose of differentiating soils. The next step in choosing informative group-specific targets for forensic analysis is to demonstrate that single bacterial groups can genetically distinguish samples.

Figure 21 contains several panels that show the intra-group species representation of select taxa for all 454 datasets. The purpose of this data was to show that there was genetic variability within groups. This was important to demonstrate, as the objective to targeting bacterial groups for forensic applications is to differentiate soils. Panels A – E contain species information for one phylum each. Given the abundance of *Proteobacteria*, four subclasses of *Proteobacteria* are shown in Panels F – I. The data show that there is genetic variability within each group shown. All groups do not contain the same levels of diversity. For example, the *Acidobacteria* and *Bacteroidetes* groups show unique richness among datasets, while the *Betaproteobacteria* and *Firmicutes* groups have a more uniform representation within the datasets. However, each panel demonstrated intra-group variation within datasets. These data support the potential for targeted analysis.

Figure 21 – Intra-group species distribution within 454 datasets

Percent composition within 454 datasets of select bacterial species are shown for the following taxonomic groups: Panel A – *Acidobacteria*; Panel B – *Actinobacteria*; Panel C – *Bacteroidetes*; Panel D – *Firmicutes*; Panel E – *Verrucomicrobia*; Panel F – *Alphaproteobacteria*; Panel G – *Betaproteobacteria*; Panel H – *Gammaproteobacteria*; Panel I – *Deltaproteobacteria*. Line graphs were created using the Community Visualization tool available on the Visualization and Analysis of Microbial Population Structures database (http://vamps.mbl.edu).

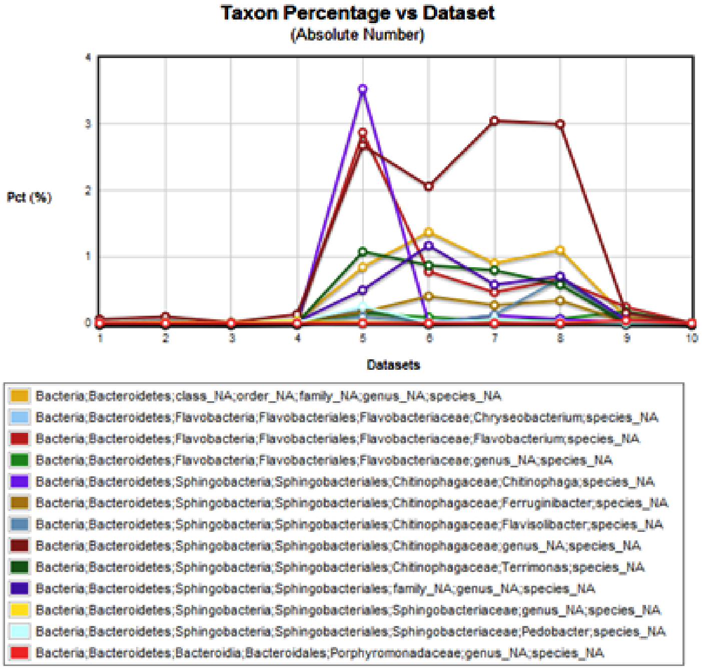
A. Acidobacteria

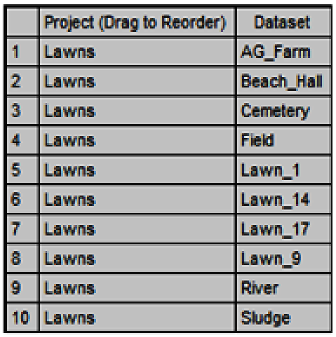


	Project (Drag to Reorder)	Dataset
1	Lawns	AG_Farm
2	Lawns	Beach_Hall
3	Lawns	Cemetery
4	Lawns	Field
5	Lawns	Lawn_1
6	Lawns	Lawn_14
7	Lawns	Lawn_17
8	Lawns	Lawn_9
9	Lawns	River
10	Lawns	Sludge

B. Actinobacteria

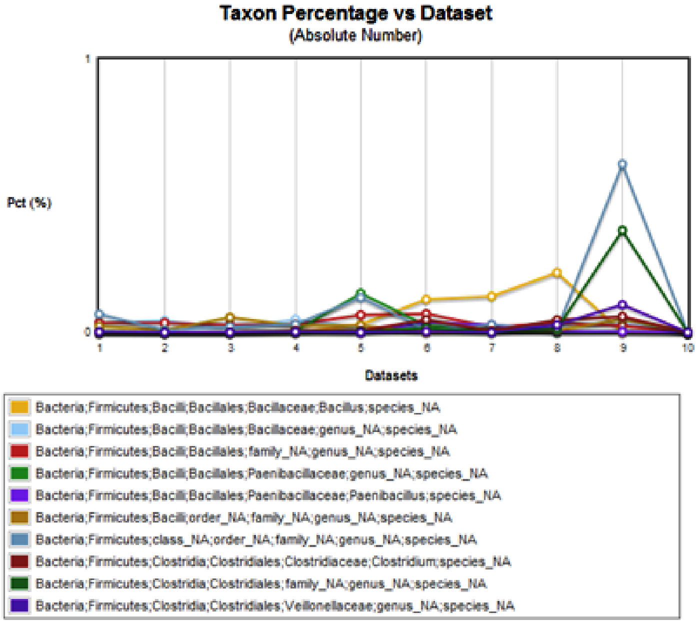
C. Bacteroidetes



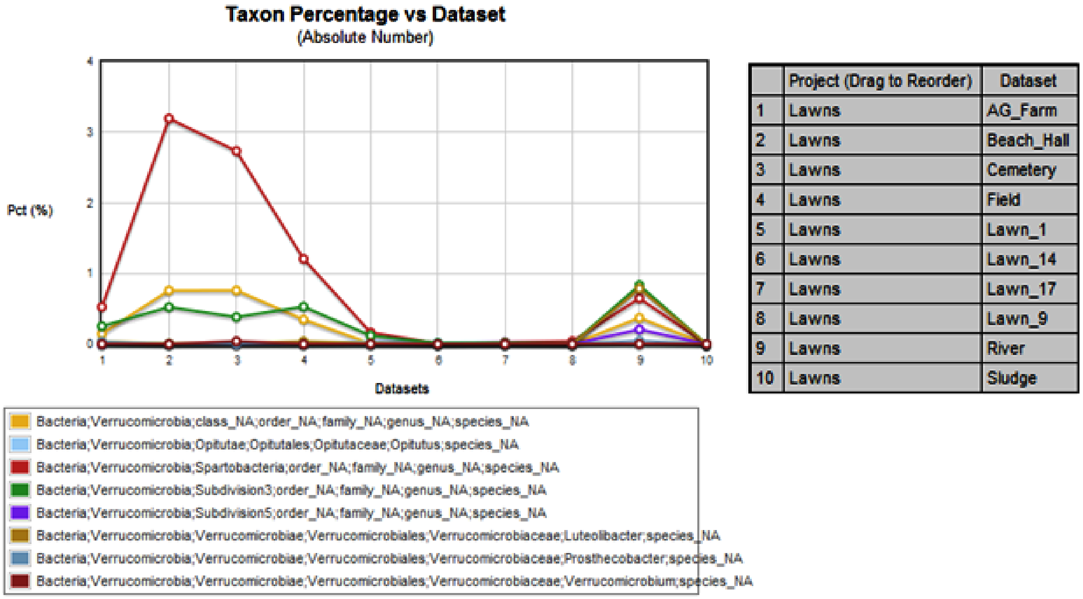


	Project (Drag to Reorder)	Dataset
1	Lawns	AG_Farm
2	Lawns	Beach_Hall
3	Lawns	Cemetery
4	Lawns	Field
5	Lawns	Lawn_1
6	Lawns	Lawn_14
7	Lawns	Lawn_17
8	Lawns	Lawn_9
9	Lawns	River
10	Lawns	Sludge

D. Firmicutes

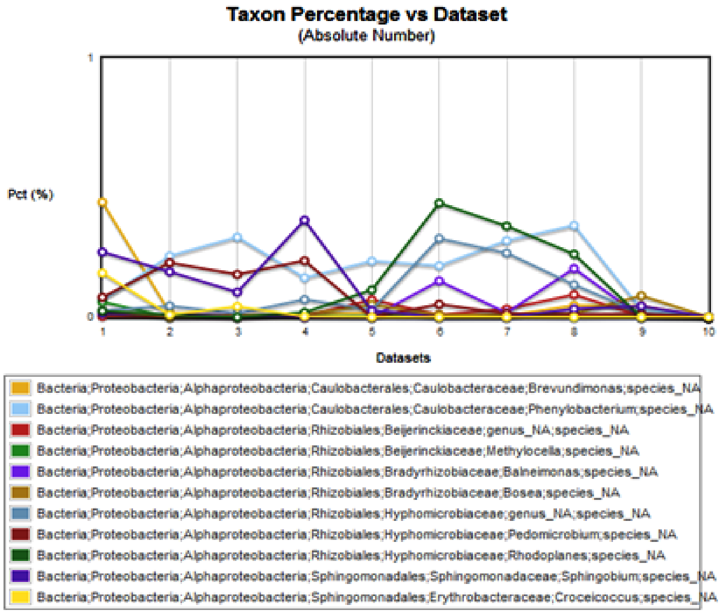


E. Verrucomicrobia

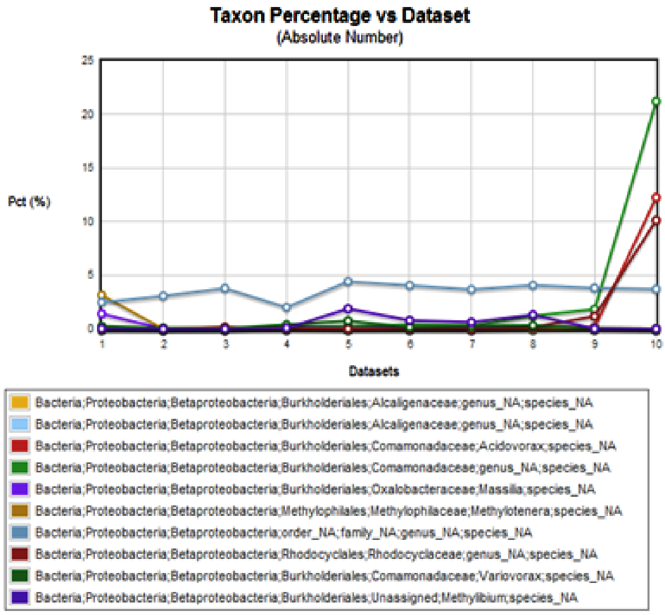


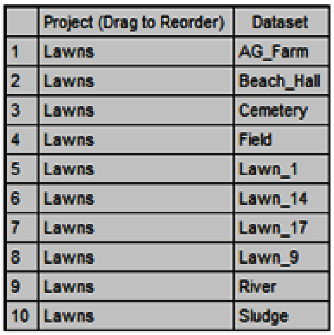
	Project (Drag to Reorder)	Dataset
1	Lawns	AG_Farm
2	Lawns	Beach_Hall
3	Lawns	Cemetery
4	Lawns	Field
5	Lawns	Lawn_1
6	Lawns	Lawn_14
7	Lawns	Lawn_17
8	Lawns	Lawn_9
9	Lawns	River
10	Lawns	Sludge

F. Alphaproteobacteria



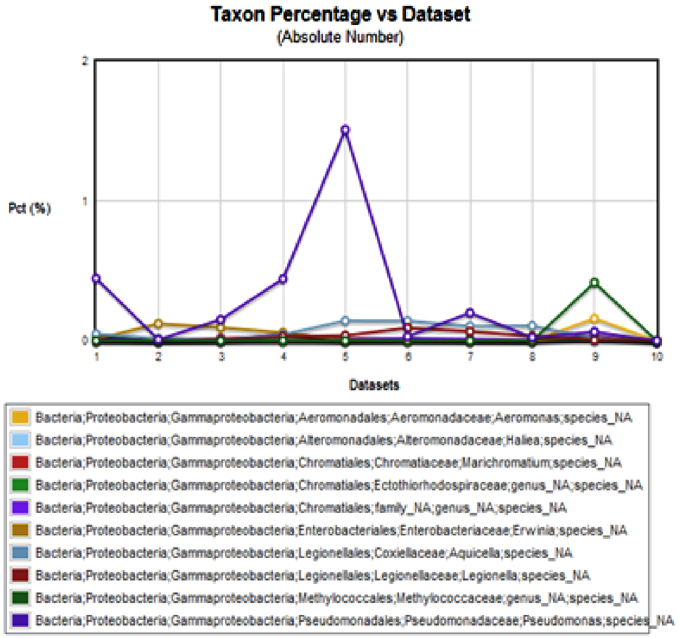
G. *Betaproteobacteria*



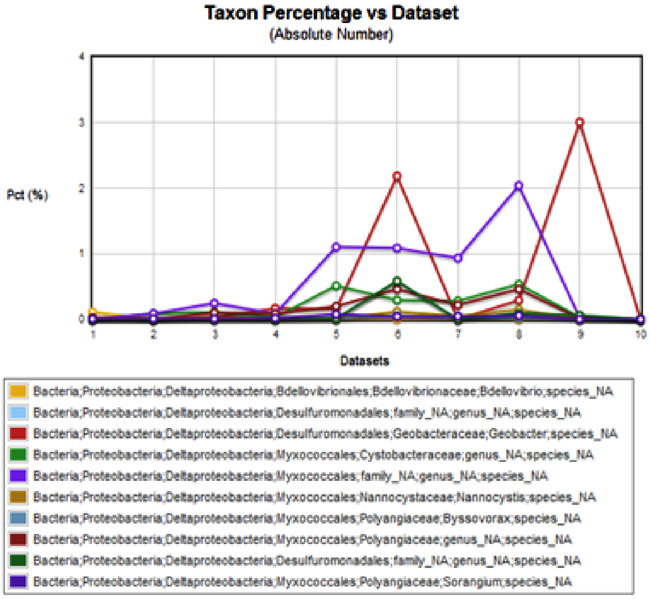


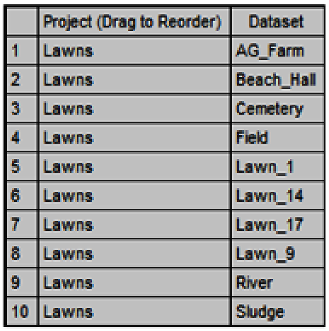
	Project (Drag to Reorder)	Dataset
1	Lawns	AG_Farm
2	Lawns	Beach_Hall
3	Lawns	Cemetery
4	Lawns	Field
5	Lawns	Lawn_1
6	Lawns	Lawn_14
7	Lawns	Lawn_17
8	Lawns	Lawn_9
9	Lawns	River
10	Lawns	Sludge

H. *Gammaproteobacteria*



I. Deltaproteobacteria





	Project (Drag to Reorder)	Dataset
1	Lawns	AG_Farm
2	Lawns	Beach_Hall
3	Lawns	Cemetery
4	Lawns	Field
5	Lawns	Lawn_1
6	Lawns	Lawn_14
7	Lawns	Lawn_17
8	Lawns	Lawn_9
9	Lawns	River
10	Lawns	Sludge

III. Discussion and Conclusions

The goal of this chapter was to generate in-depth sequence libraries from bacterial communities in soil so that forensically relevant questions could be addressed. First, what are the major and minor components to bacterial communities? Second, are there any bacterial groups that are shared among unrelated soils? Third, can soils be classified by ecological origin based on their bacterial community? Fourth, is there sufficient intra-group variation among datasets to support the feasibility of group-specific typing? The purpose of forensic soil analysis is to establish an association between two samples. A positive association between two soil samples suggests a common location. Currently, this type of forensic analysis is based on universal bacterial DNA typing using T-RFLP. However, modern techniques like 454 pyrosequencing paint an accurate picture of community structure, providing data that can improve the way soil is typed.

454 amplicon libraries are a cost-effective way to generate thousands of sequences from a variety of soils. Fourteen libraries containing an average 41,685 high quality sequences were generated. Based on richness and diversity calculations, the soils collected in this study contained more species diversity than was actually sampled. However, rarefaction data demonstrated that more than 50% of OTUs were observed in all soils. Taking the community heat map data into account, this coverage was enough to draw conclusions about community relatedness.

Rarefaction data show that in-depth amplicon libraries identified thousands of bacterial genera. The question for forensics is whether there is a minimum number of genera that must be identified to accurately describe the native community. In most cases, 5,000 sequences can potentially identify approximately 2,000 OTUs. However, the OTUs that are identified by a small number of sequences are likely major components of the soil community. Larger libraries would include genera in minority representation. Such lesser-represented bacteria may provide information about what makes two bacterial communities different. From these data, we recommend that a minimum of 15,000 sequences be collected on any given sample to create an informative survey. The Beach Hall data set produced the poorest rarefaction curve likely because these libraries contained the lowest amount of sequences (just under 15,000). Ideally, new soils and unexplored ecosystems should be more fully sequenced at least once to verify coverage.

As more information about community structure on a multitude of soils is collected, it will be possible to fine tune the number of ideal sequence reads for a screen. This will also help determine whether specialized soils or uncommon ecosystems require more (or less) than the average number of sequences to differentiate their bacterial communities. In this study, agricultural farm soil can be considered specialized, as it is maintained for the sole purpose of growing corn. Sewage sediment may also be considered specialized because of its role in sewage decomposition. However, the number sequences that are required to accurately describe each community is very different. The AG Farm will require more sequences because of the community's natural complexity while sewage sludge will require less. Although there are plenty of bacteria in sludge, the community lacks species richness.

This chapter also explored the community structures of presumed similar and diverse ecosystems. The data show that bacterial communities are not evenly distributed (in terms of species representation). *Proteobacteria* dominated all soils sampled regardless of ecosystem. This finding supports the potential of using *Proteobacteria*-specific targeted analysis because it will consistently generate data. Moving away from universal bacterial DNA typing has appeal to forensics because it increases the amount of information that can be extrapolated from a DNA profile. For example, rather than simply comparing the presence and absence of an unknown peak, investigators can attribute peak similarity to a specific bacterial group. The 454 data presented in this chapter has identified potentially informative phyla for mineral soils. Both the *Acidobacteria* and *Actinobacteria* phyla are abundant in mineral soils. The data also suggest that the *Cyanobacteria* phyla can be used to identify freshwater sediment. There are exceptions to every rule, as *Cyanobacteria* were also detected in Lawn 9. In general, the data presented in this chapter supports the investigation of the following phyla as potential candidates for group-specific soil analysis: *Proteobacteria, Acidobacteria, Actinobacteria, Bacteriodetes, Verrucomicrobia, Firmicutes, Nitrospira,* and *Cyanobacteria.*

The data also emphasize the importance of timing of sample collection. An ideal situation from a forensic perspective would be to obtain evidentiary and reference samples from one location within a short time span. Unfortunately this may not be possible for most cases. The data show that over one year's time, bacterial communities do not change drastically. Even if reference samples are not collected at exactly the same time as when evidentiary samples are deposited, it still may be possible to establish positive association.

The data show that collecting a soil sample from a location that is covered in snow can generate inaccurate information about association.   As discussed in chapter 2, universal DNA typing was able to detect differences in bacterial communities in response to environmental variables. Depending on the group-specific assay, similar results can be achieved. However, the data show that there are some groups (like *Proteobacteria*) that do not change drastically. Assays targeting these phyla would give desirable results. Interestingly, heavy rain did not seem to alter the community much from the control library. A study by Cruz-Martinez *et al*. in 2009 demonstrated that bacterial communities were able to maintain native structure after periods of rain [Cruz-Martinez, 2009]. This information suggests that rain-soaked soil is most similar to the average native bacterial community than the same soil covered in snow.

The data in this chapter supports the potential for using amplicon libraries to suggest probable ecosystems if the origin of a sample is unknown. Based on community diversity measures, the bacterial communities of the biological replicate maintained lawn samples were unique to only the lawns. This was true for all diversity measures except for the Jaccard index. The Jaccard index did not find any measurable difference in any of the libraries except the sludge. In terms of forensic potential, this diversity measure was not informative. The remaining three measures provided informative data that clustered like locations together. The Morisita-Horn index provided the most stringent data and appears to have to most forensic potential. Given the novelty of this application, it is best to include all measures, taking a consensus of each to determine potential ecosystem.

Future continuation of this work should include soils from ecosystems other than maintained lawns. Forests and areas near bodies of water would be forensically relevant locations to study. Establishing microbial community databases like VAMPS would also be useful for forensics so that amplicon library data from various diverse ecosystems could be compared and organized according to state or region. A comprehensive database would strengthen forensic soil analysis. Also, continual site monitoring that documents bacterial community change in response to environmental variables would be very helpful to forensics. These data would fine-tune our expectations for what groups are considered to have forensic potential. The data presented in this chapter only focused on two meteorological events at one location. In order for any concrete conclusions to be made about how bacterial communities respond to environmental variables more data is necessary.

The research presented in this chapter provides a foundation for the exploration of group-specific DNA typing. In this chapter, common soil bacterial groups have been identified. These common soil groups are good preliminary candidates for group-specific analysis because they are found in mineral soils and contain intra-group variation.   In the next chapter, five targets will be evaluated for their potential in forensically differentiating soils.

Chapter 4 – Exploring the potential for group-specific bacterial analysis in the forensic differentiation of soils

I. Introduction

Soil evidence can be potentially very valuable to criminal investigations by linking a suspect (or victim) to a crime scene or object. In its infancy, soil evidence was examined by physical classification using color comparisons and chemical and organic composition percentages [Sugita and Marumo, 1996]. A commentary by Morgan and Bull (2007) recognized the first use of soil evidence by George Popp in 1904 to solve a murder [Morgan and Bull, 2007]. Although physical classification was successful at that time, modern day forensic scientists understand the limitations associated with subjective analysis. Physical classification is not only difficult to perform but requires a skilled expert to interpret the data [Heath and Saunders, 2006; Horswell *et al.*, 2002]. Considering the world-wide variability of soils, a molecular approach targeting bacteria was the next logical step in advancing soil evidence.

The research by Horswell *et al.* provided a foundation for the improvement of forensic soil analysis. Forensically relevant questions could be addressed, including how time, seasons and meteorological events changed the bacterial communities in soil [Meyers and Foran, 2008]. The forensic community also gained valuable information on these questions from basic research on the same topics [Griffiths *et al.*, 2003; Lipson and Schmidt, 2004; Smit *et al.*, 2001; Walker *et al.*, 2006]. In the span of less than a decade since the study by Horswell *et al.* was published, the forensic community continued to investigate and improve soil as evidence. Modern molecular techniques like 454 pyrosequencing could be used to comprehensively describe bacterial community diversity, providing insight into how it can best be used in forensic applications. Specifically for soil analysis, the previous chapter demonstrated how high throughput sequencing of bacteria can be used to identify potential targets for group-specific analysis. The forensic potential for such targeted analysis of soils was previously suggested by Meyers and Foran [Meyers and Foran, 2008].

The research presented in this chapter investigates the potential of group-specific DNA typing for forensic soil analysis. Current universal detection methods (like T-RFLP) represent all bacteria in soil. This method does not readily allow extrapolation of the types of bacteria present in the soil sample. Knowing the precise identity of bacteria provides an additional layer of potentially informative data. In the previous chapter, 454 high-throughput sequencing was used to survey 10 soil samples to catalog the major and minor components present in each community. From these extensive DNA libraries, 5 bacterial groups have been selected as candidates for group-specific bacterial typing, employing group-specific primers with the previously described comprehensive restriction fragment length polymorphism (C-RFLP) analysis method. The resulting HPLC chromatograms are easy to read, contain peaks that can be used to objectively compare soil samples together, and are relatively inexpensive to generate. The main question being addressed in this study is whether group-specific data can be used as an alternative to universal DNA typing.

There are many target choices for group-specific evaluation. Our approach to choosing potential targets was based on representation, i.e. group-specific targets that are common to many soils. Choosing phyla that are known to be found in most soils will generate the most data for the limited number of samples we have. Given that the group-specific approach to forensic soil analysis is fairly new, it was important to design our initial experiments based on a proof-of-principle framework. Thus far, 454 amplicon pyrosequencing has shown that there is very high taxonomic diversity in soil and that within taxonomic groups species richness and abundance will vary from sample to sample. The next step was to determine the potential of select groups to forensically differentiate soils. This study will focus on members of the *Proteobacteria* and *Acidobacteria* phyla. We have also included *Firmicutes* in our study so that we can compare the performance of major and minor groups.

II. Results

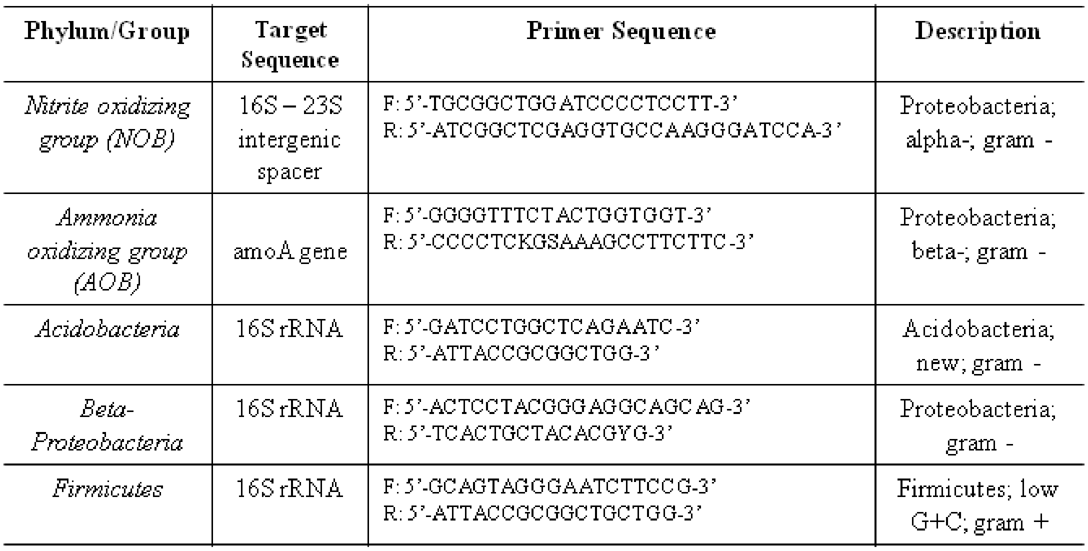
II.a. Development of group-specific assays

The main goal of this chapter is to investigate the forensic potential of targeted bacterial DNA typing to differentiate soil samples, and determine if two samples could have originated from the same location. Universal bacterial typing generates DNA profiles with unknown genetic content so taxonomic information cannot be readily extrapolated. In contrast, group-specific amplification products generate informative data that can be traced to specific bacteria. This data can be used to strengthen the conclusions made about the relationship of two soil samples. Additionally, group-specific typing alleviates the issue of amplification bias toward major bacterial phyla.

Many soil microbiome studies focus on specific bacterial groups [Blackwood *et al.*, 2005; Fierer *et al.*, 2005; Jones *et al.*, 2009; Poly *et al.*, 2008]. Although these studies do not address forensically relevant questions, they provide models for choosing targets for analysis. First and foremost, the DNA from a bacterial group must be able to be reliably extracted and amplified from a variety of soil types using standard methods. Second, there must be enough genetic variation within the genomes of the selected bacterial group to allow for sample differentiation.

Table 14 describes the targets chosen for analysis in this study. Members of the nitrite oxidizing (*NOB*) and ammonia oxidizing (*AOB*) bacterial groups were chosen because of their known roles in the nitrification of soils [Teske *et al.*, 1994]; they represent *alpha*- and *betaproteobacteria*, respectively. The high sequence similarity in the *NOB* and *AOB* groups in their 16S genes makes differentiating them difficult based on 16S data [Chu *et al.*, 2007; Grundmann *et al.*, 2000]. Gundmann *et al.* reported that the intergenic spacer region between the 16S and 23S genes contained sufficient genetic variability to differentiate members of the *NOB* group [Grundmann *et al.*, 2000]. For the *AOB* group, Horz *et al.* successfully identified ammonia-oxidizing bacteria targeting the ammonia monooxygenase gene (*amoA*) [Horz *et al.*, 2000].

Table 14 – Description of group specific targets used for HPLC analysis



Phylum/Group	Target Sequence	Primer Sequence	Description
Nitrite oxidizing group (NOB)	16S – 23S intergenic spacer	F: 5'-TGCGGCTGGATCCCCTCCTT-3' R: 5'-ATCGGCTCGAGGTGCCAAGGGATCCA-3'	Proteobacteria; alpha-; gram -
Ammonia oxidizing group (AOB)	amoA gene	F: 5'-GGGGTTTCTACTGGTGGT-3' R: 5'-CCCCTCKGSAAAGCCTTCTTC-3'	Proteobacteria; beta-; gram -
Acidobacteria	16S rRNA	F: 5'-GATCCTGGCTCAGAATC-3' R: 5'-ATTACCGCGGCTGG-3'	Acidobacteria; new; gram -
Beta-Proteobacteria	16S rRNA	F: 5'-ACTCCTACGGGAGGCAGCAG-3' R: 5'-TCACTGCTACACGYG-3'	Proteobacteria; gram -
Firmicutes	16S rRNA	F: 5'-GCAGTAGGGAATCTTCCG-3' R: 5'-ATTACCGCGGCTGCTGG-3'	Firmicutes; low G+C; gram +

NOB group primers – Grundmann *et al.*, 2000; *AOB* group primers – Horz *et al.*, 2000; *Acidobacteria*, *Betaproteobacteria*, and *Firmicutes* primers – Fierer *et al.*, 2005

Group-specific amplification of the 16S rRNA gene for the three remaining groups in this study is accomplished by utilizing conserved primer sequences flanking the variable regions. The *Acidobacteria* phylum is newly recognized, with its members commonly found in many types of soil environments [Barns *et al.*, 1999; Jones *et al.*, 1999; Kielak *et al.*, 2009]. Members of the gram positive *Firmicutes* group are not as abundant in the soil community as other phyla [Fierer *et al.*, 2005]. This was supported by the previously described 454 data. Successful detection of *Firmicutes* in soil will demonstrate that gram positive cells were being lysed during the extraction process. The *Proteobacteria* group was chosen because of its predominance in the amplicon libraries generated in this study (Table 9). In order to focus analysis on a smaller set of *Proteobacteria*, members of the *Betaproteobactera* subclass were targeted. Although the *Betaproteobactera* panel did not show as much intra-group variation as the other *Proteobacteria* subclasses, it was interesting to determine how a seeming low-diversity group would perform at differentiating soils.

Soil samples were individually amplified using these group-specific primer sets; to minimize PCR bias and to ensure sufficient PCR product, reactions were performed in triplicate and combined prior to digestion. Figure 22 depicts sample HPLC chromatograms for each group-specific assay.

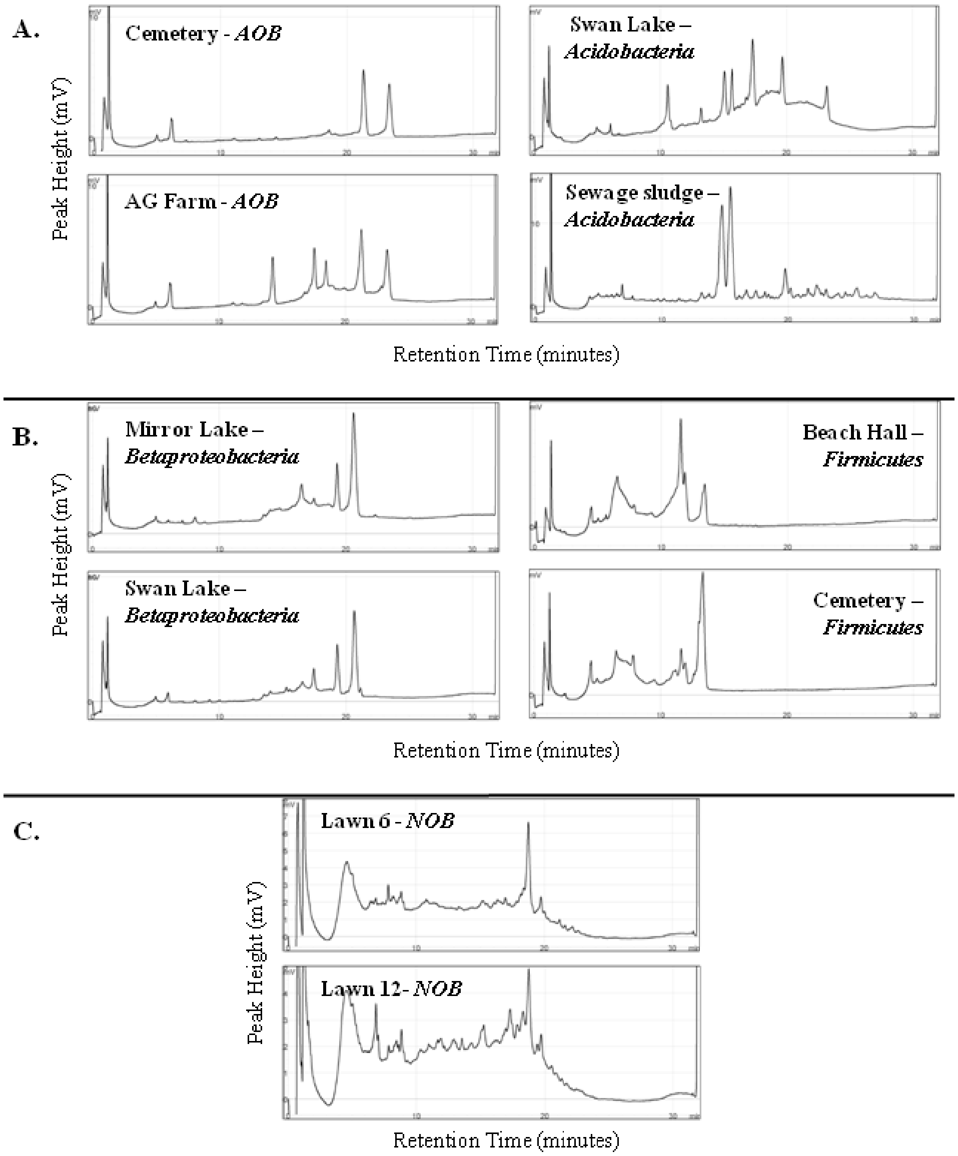


Figure 22 – Group-specific C-RFLP profiles

Examples of HPLC group-specific chromatograms. Figure shows profiles for soils from diverse ecosystems (panel A: Cemetery/AG Farm; Swan Lake/Sludge), similar ecosystems sharing local geography (panel B: Mirror Lake/Swan Lake; Beach Hall/Cemetery), and presumed biological replicate ecosystems (panel C: Lawn 6/Lawn 12) .

In chapter 2, HPLC fragment separation and detection were shown to be highly reproducible and that resulting chromatograms could be used to compare two soil samples. HPLC-based fragment analysis is a valuable method for exploring the potential of group-specific targets because DNA profiles are easy to generate and are cost-effective (fluorophores are not needed).

As evident in Figure 22, each group-specific assay produces a distinct DNA profile with varying numbers of peaks. An example of each group-specific assay is shown for either diverse ecosystems (panel A), similar ecosystems sharing local geography (panel B) and presumed biological replicate ecosystems (panel C). The data show that there are some group-specific profiles that are easier to visually interpret than others. For example, there is no question that the *Acidobacteria* profiles between Swan Lake and Sewage Sludge are different. In contrast, the profiles between Mirror Lake and Swan Lake using the *Betaproteobacteria* target share a greater number of similarities, making it more difficult to individualize the profiles by eye. In some cases, visual comparison of HPLC chromatograms is enough to individualize soils, especially when two soil samples originate from diverse ecosystems. From a forensics perspective, a profile type most amenable to this type of analysis contains clean, sharp peaks (like those seen in the *AOB, Acidobacteria,* and *Betaproteobacteria* profiles. However, the profiles generated by the *NOB* and *Firmicutes* groups are more complex and harder to interpret by eye alone. It is important to recognize that a clean profile does not necessarily mean that the assay is better suited for forensic differentiation, or vice versa. Therefore, profile similarity was based on a statistical measure so that subjective interpretation can be avoided.

II.b. Establishment of group-specific match criteria

The premise of using bacterial DNA profiles to establish the relatedness of two soil samples was first addressed by Horswell *et al.* [Horswell *et al.*, 2002]. Both Horswell *et al.* and Meyers and Foran included the Sorensen similarity index (SI) in their studies [Horswell *et al.*, 2002; Meyers and Foran, 2008]. For the SI to be a suitable metric for forensic applications, match criteria for determining whether the profiles of two samples are the same must be established.

Forensic match criteria was modeled as previously described in chapter 2, using soil from the grid experiment. Extracted DNA from each grid was analyzed using all 5 primer sets and similarity indices were calculated, using each successfully profiled grid as a reference for all others to ensure outlier references were not chosen. Figure 23 depicts the distribution of SI values for each group. The data for the *Acidobacteria* group is closest to having a normal distribution, while all the others appear to be classified by either a slightly bimodal distribution or platykurtic distribution. The fact that each group-specific test produces a wide range of SI values in the grid tells us that the targeted bacterial groups are heterogeneously dispersed in the soil. This is consistent with what is known about bacterial microenvironments [Ettema and Wardle, 2002; Grundmann and Normand, AEM, 2000]. The data also show that the range of SI values for each group-specific assay is different. The most desirable characteristic for forensic analysis is an assay that generates a high SI range within a single location (much like the distributions seen for both *Acidobacteria* and *Firmicutes*). A high SI range within a known single location subsequently provides a larger range of values below it that can describe unrelated locations.

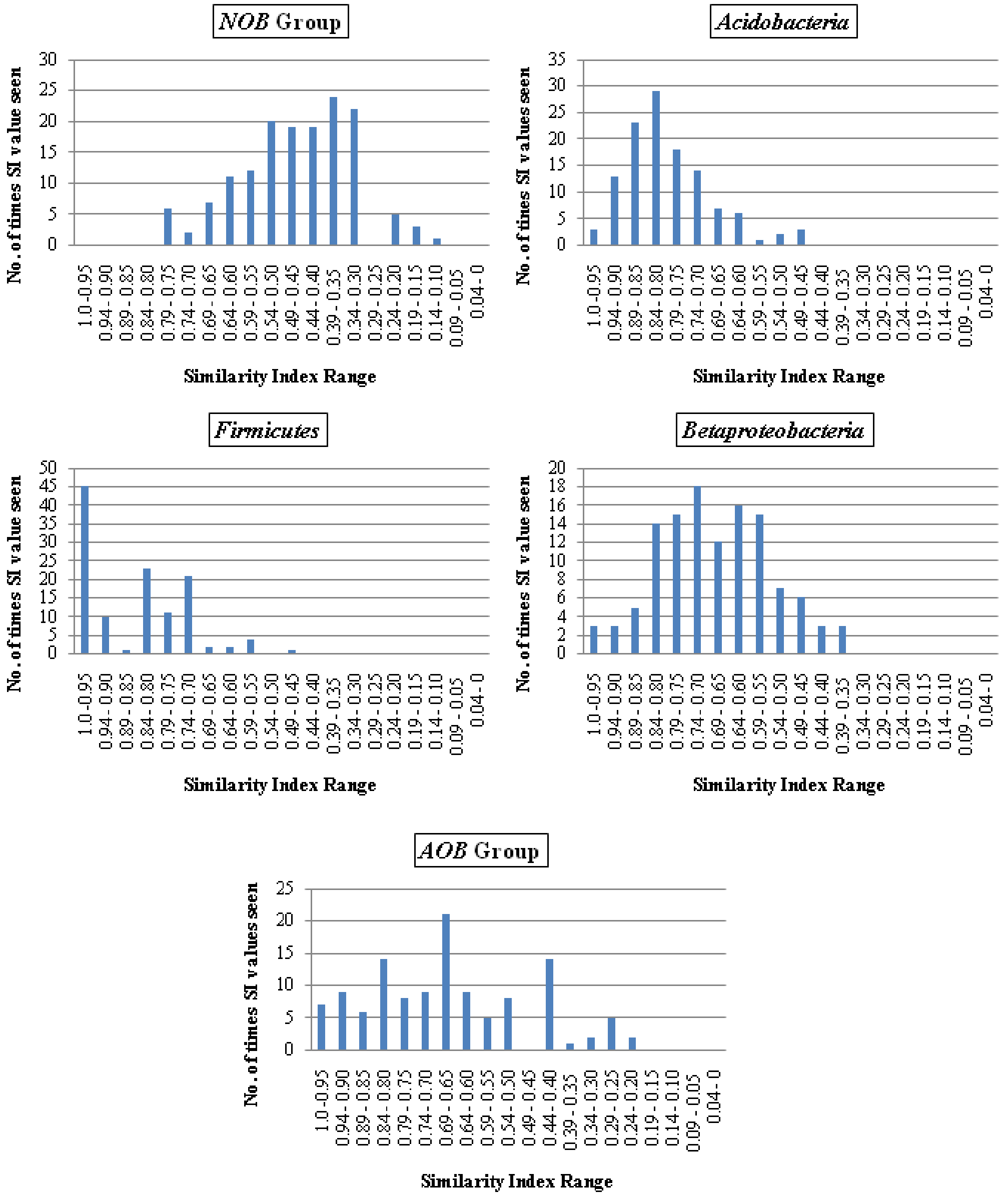
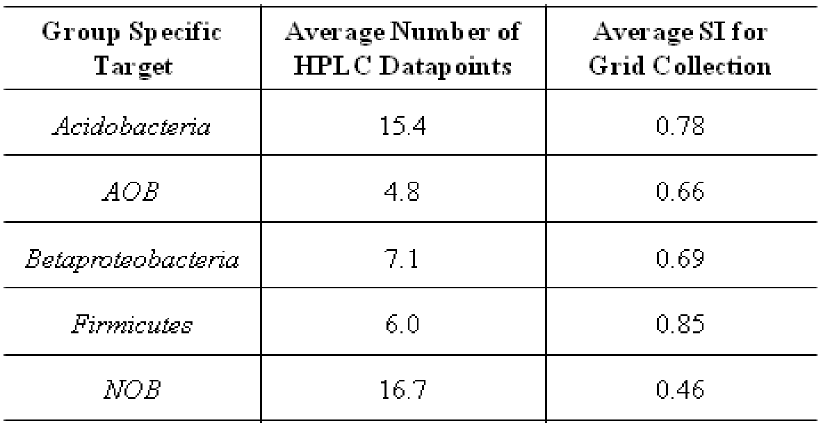


Figure 23 – Range of similarity index values observed for group-specific targets in grid analysis

Bar graphs depict the number of times an SI value is observed in grid analysis (y-axis). Data from rows 1, 2 and 3 are included in the bar graphs.

Our approach to determining whether establishing match criteria is possible is to use the average SI for each group-specific grid assay. The values used for each group are shown in Table 15. If two samples have an SI equal to or greater than the average SI for that group, then that group-specific test suggests that the two soil samples *likely came* from the same location. To parallel human STR typing, each group-specific test is considered a locus where all loci are equally informative. First, these data will indicate whether a rigid SI criterion is appropriate for each group-specific test. Second, concordance of information provided by each group-specific assay can be determined. That is, do all group-specific assays reach the same conclusion about the relatedness of known samples?

Table 15 – Grid collection data for group-specific targets



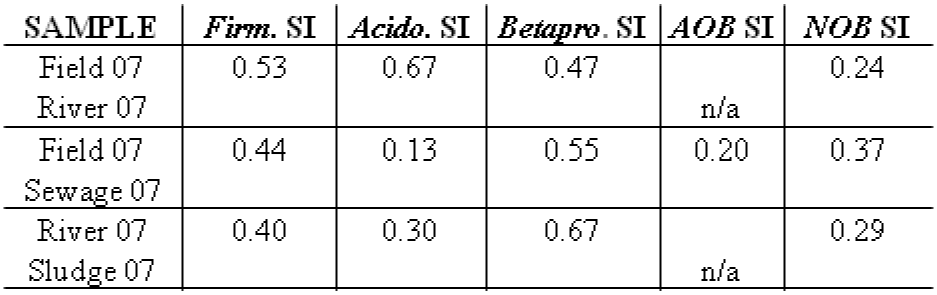
Group Specific Target	Average Number of HPLC Datapoints	Average SI for Grid Collection
Acidobacteria	15.4	0.78
AOB	4.8	0.66
Betaproteobacteria	7.1	0.69
Firmicutes	6.0	0.85
NOB	16.7	0.46

Grid sampling data was combined for each bacterial group target to generate an average similarity index (SI) for each group. This value would be used as a criterion to establish the likelihood of two samples originating from the same location. The SI data generated from each group per soil sample would be weighed equally and individually

II.c. Comparison of samples using group-specific loci

Soil samples collected from radically different ecosystems will be characterized by very different bacterial communities. Any method potentially suitable for forensic applications should reflect this. Soil/sediment collected from a field, river and sewage sludge fit the radically different classification. Table 16 shows data from these sites using the 5 group-specific tests. All 5 group-specific tests produced SI values lower than the established match criteria.

Table 16 – Comparison of presumed radically different ecosystems



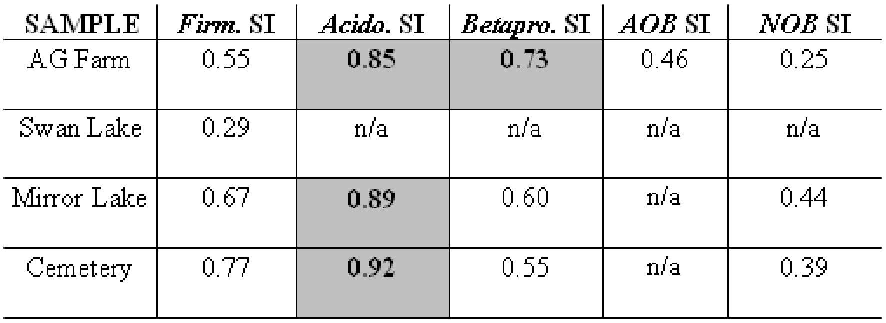
SAMPLE	*Firm.* SI	*Acido.* SI	*Betapro.* SI	*AOB* SI	*NOB* SI
Field 07 River 07	0.53	0.67	0.47	n/a	0.24
Field 07 Sewage 07	0.44	0.13	0.55	0.20	0.37
River 07 Sludge 07	0.40	0.30	0.67	n/a	0.29

Similarity indices (SI) for all samples and groups do not meet the match criteria established in the grid experiment. Grid similarity index averages for groups: *Firmicutes*, 0.85; *Acidobacteria*, 0.78; *Betaproteobacteria.*, 0.69; *AOB*, 0.66; *NOB*, 0.46. “n/a” indicates no comparison was possible for that group because no PCR product was generated for one sample.

Furthermore, when these values are plotted against the grid SI distributions shown in Figure 23, they are primarily found toward the lower ends of the ranges. Also worth noting is that absence of PCR product at the AOB locus in the CT River sample can be as informative as a positive PCR test, suggesting that the target is absent or in insufficient amounts to be amplified. The use of control DNA (plasmid containing known target sequence from a single species belonging to each group used) monitored positive PCR amplification. The lack of amplification in the AOB group is consistent with the biology of the AOB group, known to be slow growing and therefore in low biomass in bacterial communities [Chu *et al.*, 2007; Horz *et al.*, 2000]. Based on these results, we can conclude that the 5 loci chosen for analysis are sufficient for forensically differentiating soils originating from very different ecosystems; each assay reaches the same conclusion about relatedness.

Differentiation of such radically different soil types is not particularly challenging. How will group-specific assays perform to differentiate more closely related soils? Soils collected from the University of Connecticut, Storrs, share a common geography as well as more similar soil composition (at least compared to river sediment and sludge). The closer in proximity two bacterial communities are, the more similar they tend to be [Horner-Devine *et al.*, 2004]. Table 17 lists SI results for all 5 group-specific assays, using Beach Hall soil as the reference for other samples. Based on the match criteria established for each assay, there are some similarities among these samples. In the Beach Hall and Agricultural Farm comparison, both the *Acidobacteria* and *Betaproteobacteria* loci exceed the grid SI values, suggesting that there are more members of these groups that are shared between the two samples than the other groups. Members of the *Acidobacteria* group in the Beach Hall sample are also very similar to those within the Mirror Lake and Cemetery communities based on SI values above the match criteria. These data reveal similarities among the *Acidobacteria* populations within this sample set. The *Acidobacteria* assay does differentiate soils collected from the same geography by the established criteria.

Table 17 – Comparison of samples to Beach Hall soil using all group specific tests



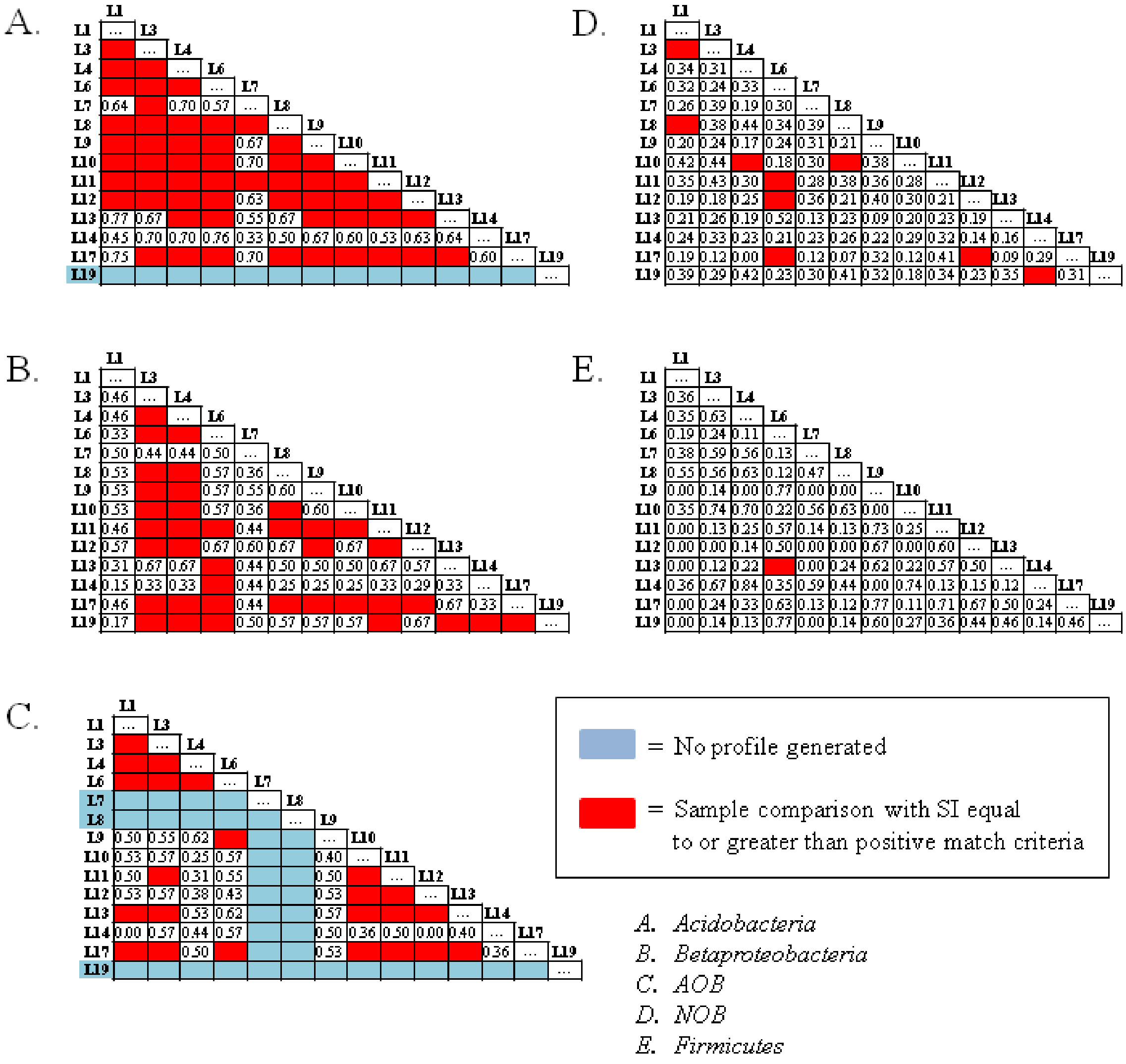
SAMPLE	*Firm.* SI	*Acido.* SI	*Betapro.* SI	*AOB* SI	*NOB* SI
AG Farm	0.55	**0.85**	**0.73**	0.46	0.25
Swan Lake	0.29	n/a	n/a	n/a	n/a
Mirror Lake	0.67	**0.89**	0.60	n/a	0.44
Cemetery	0.77	**0.92**	0.55	n/a	0.39

Similarity indices (SI) highlighted in grey indicate a positive match greater than the similarity threshold. Grid similarity index averages for groups: *Firmicutes*, 0.85; *Acidobacteria*, 0.78; *Betaproteobacteria.*, 0.69; *AOB*, 0.66; *NOB*, 0.46. “n/a” indicates no comparison was possible for that group because no PCR product was generated for one sample.

The *Firmicutes* and *NOB* assays differentiated the Storrs soils collected from different locations. Combined with the results from studies of the radically diverse ecosystems, there is forensic potential for these two assays to differentiate soils. Again, there was no amplification of the *AOB* group in the Swan Lake, Mirror Lake and Cemetery samples. This inconsistency in amplification may exclude this group from forensic testing since group-specific targets that generate interpretable profile data are preferable.

Are bacterial communities collected from biological replicate soils (such as maintained lawns) from more widespread geographical locations the same (bacterial communities are ecosystem driven) or different (communities are location/geography driven)? From a forensics perspective, it is important to determine whether soils from locations that superficially look the same produce profiles that are distinguishable. It is equally important to determine if soil samples can be analyzed to suggest habitats of origin. For example, if soil evidence was obtained in an investigation but its origin was unknown, could the soil’s bacterial community suggest a possible habitat like a forest or maintained lawn? Such information could be of great value by providing investigative leads. Since specific soil types influence which bacteria are present in the native community [Hackl *et al.*, 2004; Louzoupne *et al.*, 2007; Givran *et al.*, 2003; Nanniperieri *et al.*, 2003]. It can be hypothesized that lawns will be more similar to each other than to other soils. This hypothesis is further supported by the 454 community data presented in Figure 20 that tightly clusters the maintained lawn samples..

Figure 24 shows color-coded SI data for all Lawn samples: negative PCR amplification (blue) and SI values equal to or greater than established match criteria (red). Each group-specific assay is represented by a triangle (A-F) that depicts results of all pairwise comparisons.



A.

	L1	L3	L4	L6	L7	L8	L9	L10	L11	L12	L13	L14	L17	L19
L1	...													
L3	🟥	...												
L4	🟥	🟥	...											
L6	🟥	🟥	🟥	...										
L7	0.64	🟥	0.70	0.57	...									
L8	🟥	🟥	🟥	🟥	🟥	...								
L9	🟥	🟥	🟥	🟥	0.67	🟥	...							
L10	🟥	🟥	🟥	🟥	0.70	🟥	🟥	...						
L11	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	...					
L12	🟥	🟥	🟥	🟥	0.63	🟥	🟥	🟥	🟥	...				
L13	0.77	0.67	🟥	🟥	0.55	0.67	🟥	🟥	🟥	🟥	...			
L14	0.45	0.70	0.70	0.76	0.33	0.50	0.67	0.60	0.53	0.63	0.64	...		
L17	0.75	🟥	🟥	🟥	0.70	🟥	🟥	🟥	🟥	🟥	🟥	0.60	...	
L19	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	...

B.

	L1	L3	L4	L6	L7	L8	L9	L10	L11	L12	L13	L14	L17	L19
L1	...													
L3	0.46	...												
L4	0.46	🟥	...											
L6	0.33	🟥	🟥	...										
L7	0.50	0.44	0.44	0.50	...									
L8	0.53	🟥	🟥	0.57	0.36	...								
L9	0.53	🟥	🟥	0.57	0.55	0.60	...							
L10	0.53	🟥	🟥	0.57	0.36	🟥	0.60	...						
L11	0.46	🟥	🟥	🟥	0.44	🟥	🟥	🟥	...					
L12	0.57	🟥	🟥	0.67	0.60	0.67	🟥	0.67	🟥	...				
L13	0.31	0.67	0.67	🟥	0.44	0.50	0.50	0.50	0.67	0.57	...			
L14	0.15	0.33	0.33	🟥	0.44	0.25	0.25	0.25	0.33	0.29	0.33	...		
L17	0.46	🟥	🟥	🟥	0.44	🟥	🟥	🟥	🟥	🟥	0.67	0.33	...	
L19	0.17	🟥	🟥	🟥	0.50	0.57	0.57	0.57	🟥	0.67	🟥	🟥	🟥	...

C.

	L1	L3	L4	L6	L7	L8	L9	L10	L11	L12	L13	L14	L17	L19
L1	...													
L3	🟥	...												
L4	🟥	🟥	...											
L6	🟥	🟥	🟥	...										
L7	🟦	🟦	🟦	🟦	...									
L8	🟦	🟦	🟦	🟦	🟦	...								
L9	0.50	0.55	0.62	🟥	🟦	🟦	...							
L10	0.53	0.57	0.25	0.57	🟦	🟦	0.40	...						
L11	0.50	🟥	0.31	0.55	🟦	🟦	0.50	🟥	...					
L12	0.53	0.57	0.38	0.43	🟦	🟦	0.53	🟥	🟥	...				
L13	🟥	🟥	0.53	0.62	🟦	🟦	0.57	🟥	🟥	🟥	...			
L14	0.00	0.57	0.44	0.57	🟦	🟦	0.50	0.36	0.50	0.00	0.40	...		
L17	🟥	🟥	0.50	🟥	🟦	🟦	0.53	🟥	🟥	🟥	🟥	0.36	...	
L19	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	...

D.

	L1	L3	L4	L6	L7	L8	L9	L10	L11	L12	L13	L14	L17	L19
L1	...													
L3	🟥	...												
L4	0.34	0.31	...											
L6	0.32	0.24	0.33	...										
L7	0.26	0.39	0.19	0.30	...									
L8	🟥	0.38	0.44	0.34	0.39	...								
L9	0.20	0.24	0.17	0.24	0.31	0.21	...							
L10	0.42	0.44	🟥	0.18	0.30	🟥	0.38	...						
L11	0.35	0.43	0.30	🟥	0.28	0.38	0.36	0.28	...					
L12	0.19	0.18	0.25	🟥	0.36	0.21	0.40	0.30	0.21	...				
L13	0.21	0.26	0.19	0.52	0.13	0.23	0.09	0.20	0.23	0.19	...			
L14	0.24	0.33	0.23	0.21	0.23	0.26	0.22	0.29	0.32	0.14	0.16	...		
L17	0.19	0.12	0.00	🟥	0.12	0.07	0.32	0.12	0.41	🟥	0.09	0.29	...	
L19	0.39	0.29	0.42	0.23	0.30	0.41	0.32	0.18	0.34	0.23	0.35	🟥	0.31	...

E.

	L1	L3	L4	L6	L7	L8	L9	L10	L11	L12	L13	L14	L17	L19
L1	...													
L3	0.36	...												
L4	0.35	0.63	...											
L6	0.19	0.24	0.11	...										
L7	0.38	0.59	0.56	0.13	...									
L8	0.55	0.56	0.63	0.12	0.47	...								
L9	0.00	0.14	0.00	0.77	0.00	0.00	...							
L10	0.35	0.74	0.70	0.22	0.56	0.63	0.00	...						
L11	0.00	0.13	0.25	0.57	0.14	0.13	0.73	0.25	...					
L12	0.00	0.00	0.14	0.50	0.00	0.00	0.67	0.00	0.60	...				
L13	0.00	0.12	0.22	🟥	0.00	0.24	0.62	0.22	0.57	0.50	...			
L14	0.36	0.67	0.84	0.35	0.59	0.44	0.00	0.74	0.13	0.15	0.12	...		
L17	0.00	0.24	0.33	0.63	0.13	0.12	0.77	0.11	0.71	0.67	0.50	0.24	...	
L19	0.00	0.14	0.13	0.77	0.00	0.14	0.60	0.27	0.36	0.44	0.46	0.14	0.46	...

🟦 = No profile generated

🟥 = Sample comparison with SI equal to or greater than positive match criteria

A. Acidobacteria
B. Betaproteobacteria
C. AOB
D. NOB
E. Firmicutes

Figure 24 – Group-specific SI data for lawn replicate soils.

Triangles represent all comparisons between lawn soils. A. *Acidobacteria;* B. *Betaproteobacteria;* C. *AOB;* D. *NOB;* E. *Firmicutes.* Query samples are listed down the left side of the triangles; reference samples are diagonally across the top. Some SI values have been color coded for ease of interpretation (see legend).

Members of the *Acidobacteria* group (panel A) are widely shared between maintained lawns. The large number of SI values higher than the grid match criteria indicates that habitat acts as a driving force rather than local geography itself. These data show potential for the *Acidobacteria* group assay to identify maintained lawns, but it is not sufficiently discriminatory alone to differentiate soils from similar habitats.

Panel B shows data for the *Betaproteobacteria* group assay which produced the second highest incidence of SI values meeting grid match criteria. Given that a the majority of SI values fell below the match criteria, the results suggest that *Betaproteobacteria* may not be as reliable in identifying maintained lawns as the *Acidobacteria* group assay. The *Betaproteobacteria* group assay is better suited for soil differentiation than the *Acidobacteria* group assay. However, the true potential of this assay is inconclusive given that the SI values falling above and below the match criteria are just about equal.

Data shown in Panel C for the *AOB* group are consistent with previously described results in that there is a high incidence of failure to amplify. The forensic potential of the *AOB* group is limited since this assay produces inconsistent profile data. A useful experiment would be to explore the *AOB* group using the 16S rRNA gene (or another genomic region). If other molecular targets confirm the low abundance of members of the *AOB* group, this assay is undesirable for forensic testing.

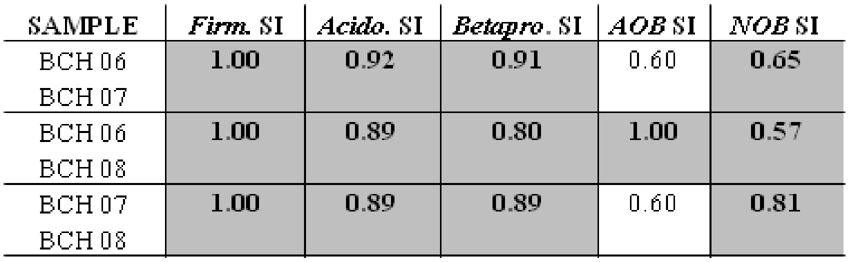
The *NOB* and *Firmicutes* groups (panels D and E, respectively) provided data that best differentiated maintained lawn soils from widespread locations. The *Firmicutes* group performed the strongest, successfully differentiating 99% of lawns and generating the highest number of 0% SI values. Both assays reveal that similar *Firmicutes* and *NOB* patterns are not consistently found in maintained laws and that similarity between two soils reflects a very local, highly restricted geography. The profile data generated from these two groups would not be useful to identify a maintained lawn.

II.d. Year-to-year sampling

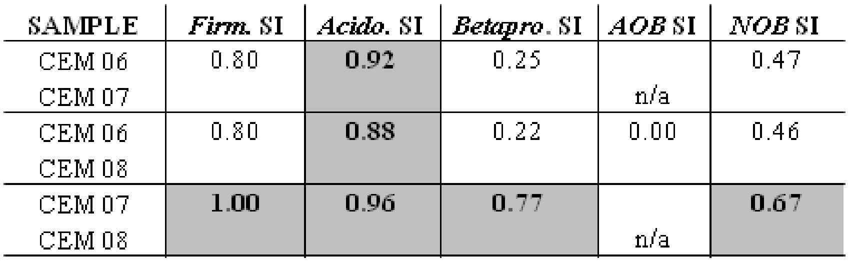
The premise of determination of soil sample origin is predicated on the notion that bacterial communities do not change very much over time within the same location. In fact, bacterial communities do fluctuate over time, as shown in both the forensic study by Meyers and Foran [Meyers and Foran, 2008] and in other studies [Lipson *et al.*, 2004; Smit *et al.*, 2001; Walker *et al.*, 2006]. Environmental influences over bacterial community structure cannot be controlled, but its effects on DNA profiles can be gauged to guide and control for time of collection. This portion of the research investigated whether samples collected from the same area at different time points might be falsely interpreted as originating from unrelated locations based on group-specific assay data.

Previously extracted soils from the year-to-year experiment sampling were re-analyzed using the group-specific assays. The purpose of this experiment was to determine how much each group's profile changed over time. The similarity index results for this section are shown in Table 18.

Table 18 – Year-to-year sampling of soils collected at the University of Connecticut



SAMPLE	*Firm.* SI	*Acido.* SI	*Betapro.* SI	*AOB* SI	*NOB* SI
BCH 06 BCH 07	**1.00**	**0.92**	**0.91**	0.60	**0.65**
BCH 06 BCH 08	**1.00**	**0.89**	**0.80**	**1.00**	**0.57**
BCH 07 BCH 08	**1.00**	**0.89**	**0.89**	0.60	**0.81**



SAMPLE	*Firm.* SI	*Acido.* SI	*Betapro.* SI	*AOB* SI	*NOB* SI
CEM 06 CEM 07	0.80	**0.92**	0.25	n/a	0.47
CEM 06 CEM 08	0.80	**0.88**	0.22	0.00	0.46
CEM 07 CEM 08	**1.00**	**0.96**	**0.77**	n/a	**0.67**

Similarity indices (SI) highlighted in grey indicate a positive match greater than the similarity threshold. Grid similarity index averages for groups: *Firmicutes*, 0.85; *Acidobacteria*, 0.78; *Betaproteobacteria.*, 0.69; *AOB*, 0.66; *NOB*, 0.46. “n/a” indicates no comparison was possible for that group because no PCR product was generated for one sample.

Table 18 shows year-to-year data for Beach Hall and Cemetery. Beach Hall soil achieved the most forensically desirable result, with soil from 2007 and 2008 matching at all 5 loci, demonstrating that bacterial groups native to this location did not change much from year to year. However, this does not imply that the community did not change at all during the course of one year. Month-to-month sampling would shed light on whether there are slight fluctuations over time in response to seasons, for example. Additionally, other members of the bacterial community not examined in group-specific assays could have fluctuated.

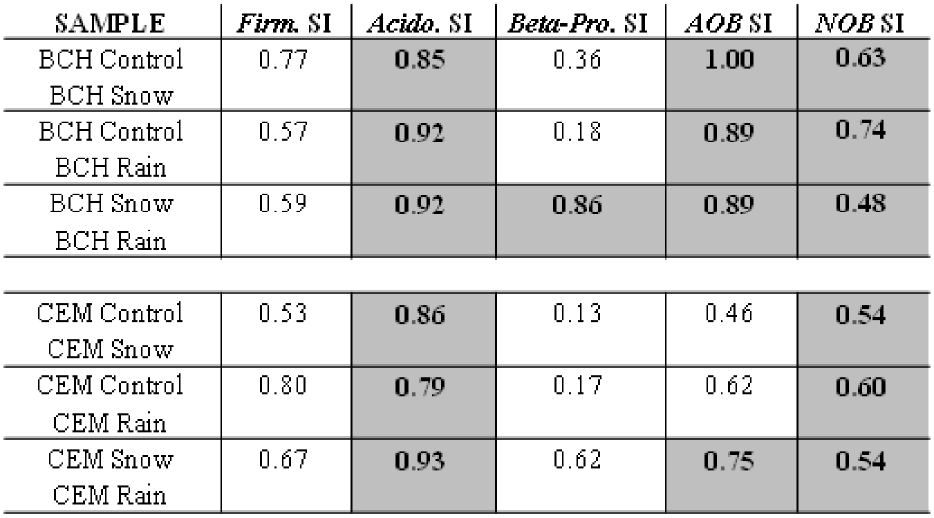
On the contrary, the bacterial groups assayed in Cemetery soil did not remain consistent over time. Only one comparison (CEM 2007 and CEM 2008) generated data similar to the Beach Hall data. In this very limited study time influenced bacterial communities differently, and the extent of these changes will be dependent on the local conditions. As previously stated, both locations were accessible to the public. The data generated for Cemetery soil could have been influenced by another variable in addition to time.

These results demonstrate that time can influence the structure of certain bacterial communities. The degree to which this happens is not uniform, as shown with the Beach Hall and Cemetery sets. In both sets, members of the *Acidobacteria* group retain a consistent structure from year-to-year. This is desirable for forensics given the potential for there to be a difference in the collection times of evidentiary and reference samples. As a whole, each group-specific assay generated different information about relatedness. What this data means for forensic soil analysis is that if reference and evidentiary samples are not collected within a certain time frame of each other, then there is a possibility that the resulting DNA profiles may not accurately reflect their actual relatedness (if in fact the two samples did originate from the same location).

II.e. Meteorological event sampling

Addressing the impact of meteorological events on bacterial communities in soil is another forensically relevant topic. Based on the data in Table 19, it is clear that some bacterial groups change over one year’s time. To further complicate matters, this fluctuation is not consistent. In the next experiment, sampling locations around the University of Connecticut were evaluated for measurable changes after heavy rainfall and when the ground is covered in snow. The data in Table 19 lists SI values for group-specific assays at all locations.

Table 19 – Impact on group-specific analysis after heavy rainfall and snow cover



SAMPLE	*Firm.* SI	*Acido.* SI	*Beta-Pro.* SI	*AOB* SI	*NOB* SI
BCH Control BCH Snow	0.77	**0.85**	0.36	**1.00**	**0.63**
BCH Control BCH Rain	0.57	**0.92**	0.18	**0.89**	**0.74**
BCH Snow BCH Rain	0.59	**0.92**	**0.86**	**0.89**	**0.48**
CEM Control CEM Snow	0.53	**0.86**	0.13	0.46	**0.54**
CEM Control CEM Rain	0.80	**0.79**	0.17	0.62	**0.60**
CEM Snow CEM Rain	0.67	**0.93**	0.62	**0.75**	**0.54**

Similarity indices (SI) highlighted in grey indicate a positive match greater than the similarity threshold. Grid similarity index averages for groups: *Firmicutes*, 0.85; *Acidobacteria*, 0.78; *Betaproteobacteria.*, 0.69; *AOB*, 0.66; *NOB*, 0.46. "n/a" indicates no comparison was possible for that group because no PCR product was generated for one sample.

Data from the Beach Hall set demonstrate that the groups tested respond differently to environmental variables. Each *Firmicutes* test generates an SI value lower than the grid match threshold. Although the *Firmicutes* group fluctuated annually, the SI values from that experiment remained above the match criteria. The *Acidobacteria* group retained its native structure in response to meteorological events. Also retaining consistency were members of the *AOB* and *NOB* groups. Lastly, the *Betaproteobacteria* group appeared to change in response to snow and rain.

The Cemetery soil set shared similarity to the Beach Hall set in that both the *Acidobacteria* and *NOB* groups generated SI values above the match criteria for each comparison. This is of importance to forensic applications, since these two groups are robust enough to withstand the impact of meteorological events. Again, we also see that the Cemetery set responds differently to snow and rain than the Beach Hall soil. Forensic interpretation of the relatedness of soil samples after meteorological events must be approached with caution, as environmental variables can have a broad impact on soils depending on the location, as well as the group being assayed.

II.f. Peak identification using HPLC fragment collection

One of the advantages to using the HPLC WAVE® system is the potential for fragment identification by DNA sequencing. In instances where more information is needed in order to solidify DNA typing results, the WAVE® system can collect peaks for further analysis. For high-profile forensic cases, this feature is desirable as it can provide an added layer of information that strengthens the evidence.

The purpose of this experiment was to determine the efficacy of fragment collection and identification using mixed templates. Although group-specific assays targeted a small subset of the bacterial community, the amplicons generated were still heterogeneous. For standard Sanger sequencing to work, the DNA collected must be single copy. Lawn 1 and Lawn 9 were used in this experiment, comparing the results from the *AOB* group and the *Acidobacteria* group. The *AOB* group assay targets a functional gene that is conserved in sequence, while the *Acidobacteria* group targets the 16S gene which is highly variable.

Figure 25 shows C-RFLP profiles for Lawn 1 and Lawn 9 (top panel). Undigested amplicons from both assays were separated by HPLC and their respective chromatograms are shown in the bottom panel.

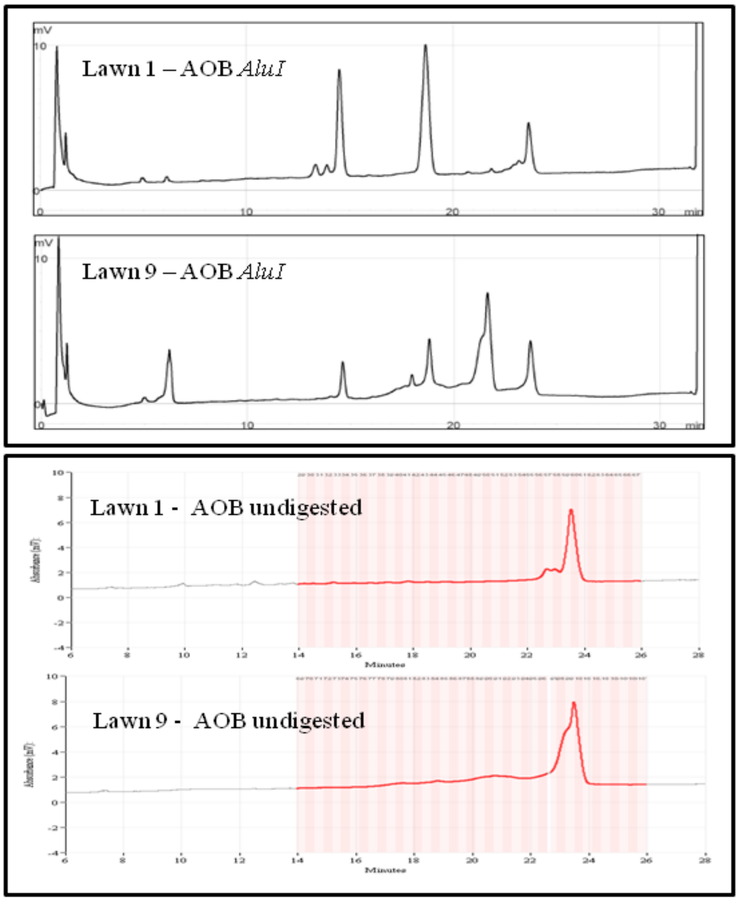
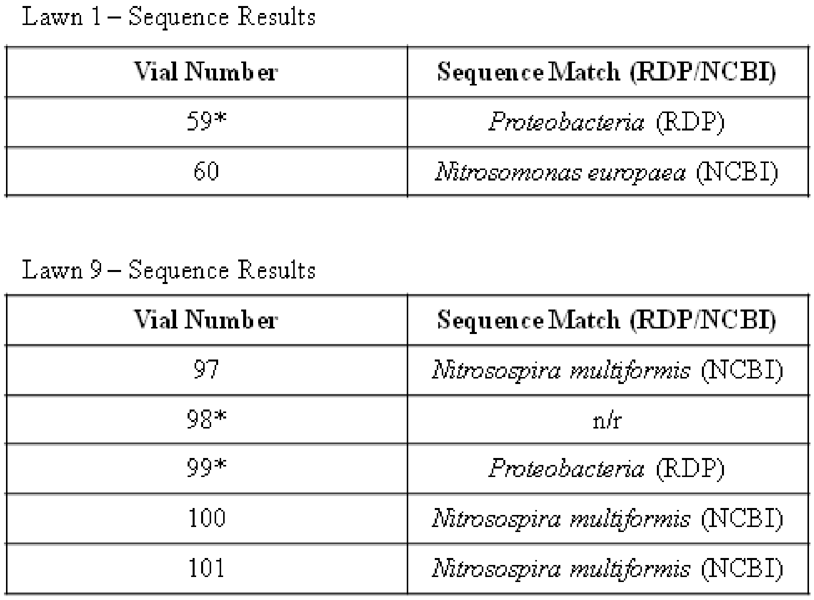


Figure 25 – HPLC separation of PCR products targeting the *AOB* group [*amoA* gene]: Lawn 1and Lawn 9

Top panel shows C-RFLP profiles for Lawn 1 and Lawn 9. Bottom panel shows fragment collection data for undigested *amoA* amplicon for Lawn 1 and Lawn 9.

It is clear that the C-RFLP profiles from Lawn 1 and 9 are different. If it were necessary to provide additional data supporting this conclusion, sequence confirmation can be done. The WAVE® system has the ability to dispense peaks into multiple vials. Initially, whole peaks were collected and sequenced. The results from combining vials together were problematic, as the DNA sequences appeared to contain more than one sequence. To minimize the undesirable effect of multiple templates, single vials were sequenced. The results from the *AOB* group for Lawns 1 and 9 are shown in Table 20.

Table 20 – Sequence match results for Lawn 1 and Lawn 9 - *AOB* group



Lawn 1 – Sequence Results

Vial Number	Sequence Match (RDP/NCBI)
59*	*Proteobacteria* (RDP)
60	*Nitrosomonas europaea* (NCBI)

Lawn 9 – Sequence Results

Vial Number	Sequence Match (RDP/NCBI)
97	*Nitrosospira multiformis* (NCBI)
98*	n/r
99*	*Proteobacteria* (RDP)
100	*Nitrosospira multiformis* (NCBI)
101	*Nitrosospira multiformis* (NCBI)

Table lists sequence match results for Lawn 1 (top) and Lawn 9 (bottom) *AOB* group analysis for all vials collected. Both RDP and NCBI were used to classify the sequence. Result shown only lists the most complete result. (*) Vial #59 had poor sequence quality; could not confidently call nucleotides. Vial #98 was not able to generate sequence. Vial #99 had poor sequence quality; could not confidently call nucleotides.

The undigested amplicon from Lawn 1 was dispensed into 2 vials; Lawn 9 amplicon was dispensed into 5 vials. Keeping each vial separate was sufficient enough to eliminate multiple template contamination in most of the vials. There was enough sequence data generated from each sample to confidently determine the bacterial species that was being targeted in each assay. [Sequence alignments for Lawn 1 and Lawn 9 are shown in supplementary figures S28 and S29, respectively]. The C-RFLP profile generated from Lawn 1 is a result of the presence of the species, *Nitrosomonas europaea,* while the Lawn 9 profile is a result of the presence of the species, *Nitrosospira multiformis.* These results further validate the specificity of the *AOB* assay.

The results for the *Acidobacteria* group were not as successful as the *AOB* group. Figure 26 shows *Acidobacteria* data for Lawns 1 and 9.

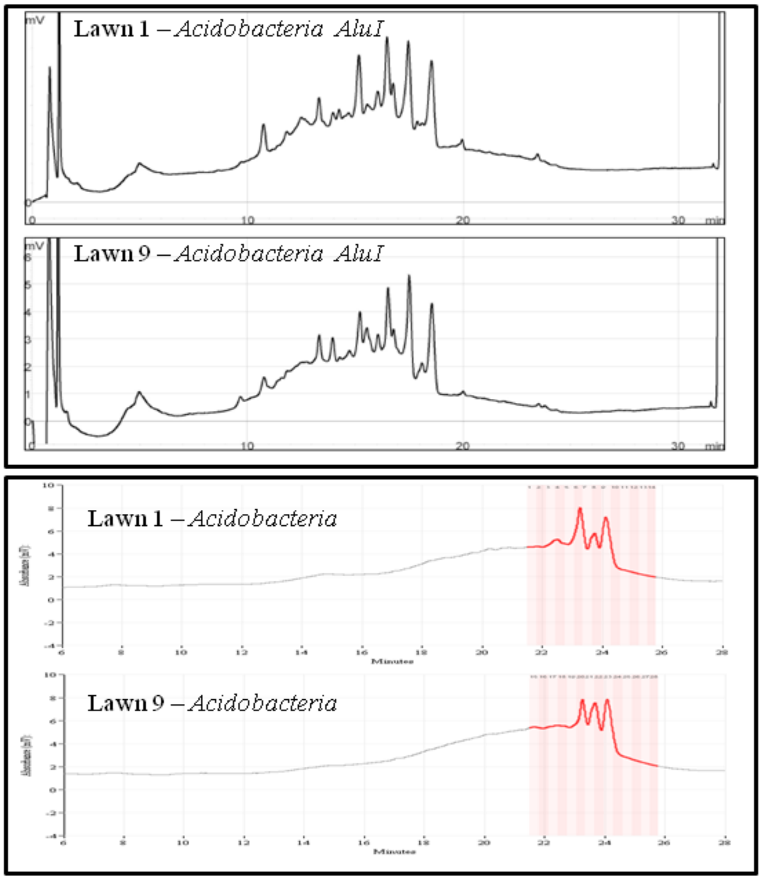
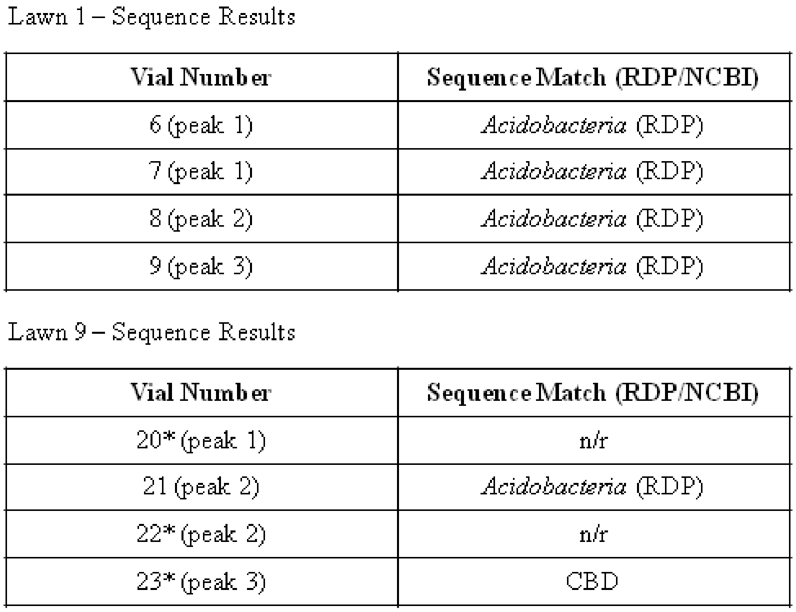


Figure 26 – HPLC separation of PCR products targeting *Acidobacteria* [16S gene]: Lawn 1 and Lawn 9

Top panel shows C-RFLP profiles for Lawn 1 and Lawn 9. Bottom panel shows fragment collection data for undigested *Acidobacteria* amplicon for Lawn 1 and Lawn 9.

As compared to Figure 25, the *Acidobacteria* assay generates more peaks because of the hypervariability of the region being targeted. HPLC separation of the undigested amplicons reveals three peaks, suggesting that there are three amplicons of different size and possibly sequence. Sequence results from each single vial are reported in Table 21.

Table 21 – Sequence match results for Lawn 1 and Lawn 9 - *Acidobacteria*



Lawn 1 – Sequence Results

Vial Number	Sequence Match (RDP/NCBI)
6 (peak 1)	*Acidobacteria* (RDP)
7 (peak 1)	*Acidobacteria* (RDP)
8 (peak 2)	*Acidobacteria* (RDP)
9 (peak 3)	*Acidobacteria* (RDP)

Lawn 9 – Sequence Results

Vial Number	Sequence Match (RDP/NCBI)
20* (peak 1)	n/r
21 (peak 2)	*Acidobacteria* (RDP)
22* (peak 2)	n/r
23* (peak 3)	CBD

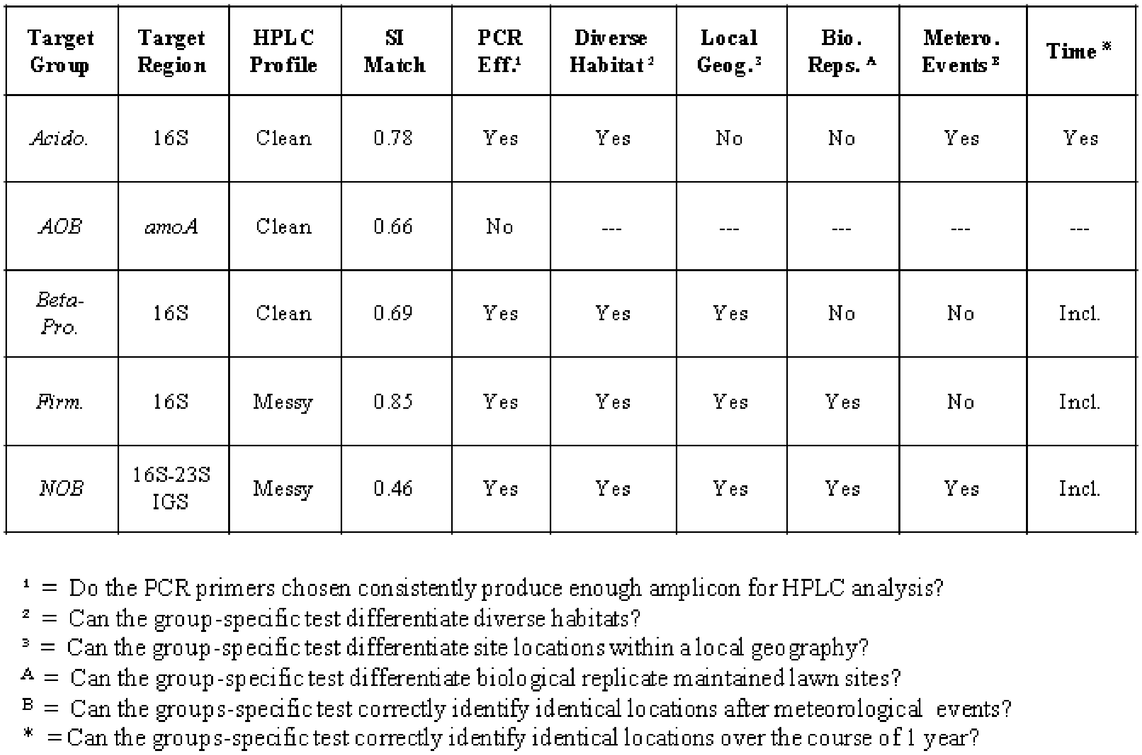
Table lists sequence match results for Lawn 1 (top) and Lawn 9 (bottom) *Acidobacteria* analysis for all vials collected. Both RDP and NCBI were used to classify the sequence. Result shown only lists the most complete result. (*) Vial #20 and #22 was not able to generate sequence. Vial #23 produced truncated sequence with poor sequence quality; could not confidently call nucleotides.

Sequences generated from Lawn 1 were of better quality than Lawn 9. However, none of the vials in either samples generated clean sequences (i.e. without evidence of mixed template). Data analysis of the sequences was done conservatively, only manually calling bases that were clearly missed by the software. Classification of the sequences revealed *Acidobacteria* taxonomy. Further classification past the phylum level could not be done because of the high amount of ambiguous bases. Separation of the amplicons would have to be done using traditional cloning methods. Although the results were not as specific as the *AOB* group, the data show that HPLC fragment collection can be used to identify peaks based on sequence. This type of analysis is better suited for amplicon pools that are as close to single source as possible.

III. Discussion and Conclusions

The purpose of this study was to evaluate the forensic potential of bacterial group-specific assays to establish the relatedness of soils. Universal bacterial typing generates DNA profiles with unknown genetic content so taxonomic information cannot be readily extrapolated. In contrast, group-specific amplification products generate more informative data that can identify specific bacteria. These data can strengthen the conclusions made about the relationship of two soil samples. Additionally, group-specific typing alleviates the issue of amplification bias toward major bacterial phyla. Using modern sequencing techniques, we have chosen target phyla for DNA typing, primarily based on abundance. This study focused on two major phyla (*Proteobacteria* and *Acidobacteria*) and one minor phylum (*Firmicutes*). Table 22 provides a general summary of the performance of each group-specific assay with respect to several criteria (listed below table).

Table 22 - Summary of findings for groups-specific assays



Target Group	Target Region	HPLC Profile	SI Match	PCR Eff.[1]	Diverse Habitat[2]	Local Geog.[3]	Bio. Reps.[A]	Metero. Events[B]	Time[*]
Acido.	16S	Clean	0.78	Yes	Yes	No	No	Yes	Yes
AOB	*amoA*	Clean	0.66	No	---	---	---	---	---
Beta-Pro.	16S	Clean	0.69	Yes	Yes	Yes	No	No	Incl.
Firm.	16S	Messy	0.85	Yes	Yes	Yes	Yes	No	Incl.
NOB	16S-23S IGS	Messy	0.46	Yes	Yes	Yes	Yes	Yes	Incl.

[1] = Do the PCR primers chosen consistently produce enough amplicon for HPLC analysis?
[2] = Can the group-specific test differentiate diverse habitats?
[3] = Can the group-specific test differentiate site locations within a local geography?
[A] = Can the group-specific test differentiate biological replicate maintained lawn sites?
[B] = Can the groups-specific test correctly identify identical locations after meteorological events?
[*] = Can the groups-specific test correctly identify identical locations over the course of 1 year?

Table provides a summary of the performance of each group-specific assay evaluated in this study: *Acidobacteria, AOB, Betaproteobacteria, Firmicutes,* and *NOB.* Table lists each assay's genomic target region, subjective classification of the resulting HPLC chromatograms (i.e. clean or messy), similarity index match criteria.

Five group-specific targets were evaluated on a variety of soil samples. HPLC analysis effectively evaluated the potential of multiple targets without the financial burden of fluorophore-labeled primers. Grid experiments for each group-specific assay demonstrated that members were heterogeneously dispersed in soil. In terms of forensic analysis, this result is problematic. Group-specific profiles can be different even when two soil samples are collected from the same location, complicating the establishment of match criteria. Each group-specific assay revealed a wide range of differently distributed SI values. Using an absolute average SI value as a threshold for determining the relatedness of soil samples may not be appropriate for all assays and future research should explore profile analysis using less stringent thresholds.

The data presented in this chapter show that group-specific analysis can differentiate diverse soils. As soils become more similar (i.e. similar composition or local geography), bacterial groups can be shared. Throughout all sample sets, the *Acidobacteria* group had the highest similarity among mineral soils from both related and unrelated locations. Although this assay did not perform well in differentiating soils, it has the potential to serve as a biological marker for mineral soils from maintained lawns. Both the *NOB* and *Firmicutes* groups generated the most consistent data that accurately differentiated soils from different sites, showing the most potential in a forensic context.

Given the heterogeneity of soil, using multiple group-specific assays to determine the relatedness of two soil samples will provide the most detailed information. However, accurate differentiation of nearly all lawn samples was accomplished using only the *Firmicutes* assay. The key to choosing forensically informative assays will be the assay itself, and not the number of tests included. Most importantly, the choice of assay will be dependent on whether sample relatedness or ecosystem suggestion is the analysis goal. Furthermore, it is possible that different group-specific assays will be better/worse for different soil types. This study has identified forensically informative assays for maintained lawns. These assays may (or may not) perform as well on organic or sandy soils.

Meteorological events and time can alter the structure of certain bacterial groups. The *Acidobacteria* group did not appear to fluctuate in response to these variables, further validating its potential as a forensically informative group. Although able to consistently differentiate soils, it was shown that the *Firmicutes* group was influenced by these variables. Altogether, these results speak to the importance of timely evidence and reference sample collection. The most desirable situation would be for reference sample collection to occur within hours of a crime. In a majority of cases this is an unrealistic scenario. However, being aware of the impact these environmental variables have is critical to proper analysis.

Each assay was shown to represent bacterial groups uniquely, in turn providing different conclusions about the relatedness of soils samples. Some groups are better at distinguishing soils, while others are more useful for suggesting locations of origin. Data generated by the *AOB* group did not provide useful information about the relatedness of soils. This was likely due to the low amount of *AOB* members in the soils collected.

There are interpretational limitations to determining the relatedness of soil samples using group-specific targets. If group-specific tests are to be used, the forensic community must determine how many tests must "match" in order for the samples to be interpreted as originating from the same location. Conversely, there must be guidelines for interpretation of soils that fall below match criteria. For example, there can be two explanations for soil samples having a lower than expected SI value. Either soil samples A and B are not from the same location, or soil samples A and B are *likely* from the same location, but the time between the collection of these samples has changed the native bacterial community in one sample. Perhaps the most difficult variable to measure in forensic soil analysis will be the environment. Not all locations are exposed to the same environmental variables so the impact to bacterial communities will vary. Translating this variation into a probability will be useful for interpretation.

Chapter 5 – Thesis synopsis and future directions

I. Synopsis

The data presented herein aims to close the gap between basic research and related forensic applications. The ultimate goal of all translational research is to generate data that can be used in the design of a specific application or diagnostic test. In this case, the data herein adds to an already strong foundation of basic soil research with the ultimate goal of a multiplex typing kit for the forensic community. We recognize that 454 pyrosequencing and HPLC analysis of nucleic acids may never be part of routine analysis in crime labs; these two methods can be a part of a national, regional or commercial service labs. However, these modern techniques have provided valuable data from which forensically relevant topics have been explored.

Bacterial communities in soil are demonstratively complex. Fortunately, we have instrumentation at our disposal that makes analyzing these communities easier. In-depth pyrosequencing generates tens of thousands of bacterial sequences per sample in a few days, something that traditional cloning methods could never do. Regardless of the complexity of the instrument or quantity of data, the basic questions are still the same. The first basic question addressed the forensic potential of two DNA typing methods.

Universal bacterial DNA profiles from soil can be used to establish the relatedness of two samples. However, neither T-RFLP nor C-RFLP can be used to determine the exact location of origin of a soil sample. This does not speak to the efficiency of either method, rather it is a limitation due to the natural, heterogenic dispersal of bacteria in soil. The sensitivity of the T-RFLP method highlights this characteristic more so than C-RFLP. With the capillary sequencer's resolution capability of one base pair, in addition to a very sensitive fluorophore detection system, both major and minor terminal fragments are readily visualized. If the goal of T-RFLP analysis is to estimate bacterial diversity in soil, then this feature is desirable. From a forensics perspective, this sensitivity complicates analysis. The key to choosing an appropriate soil analysis method is that it must be sensitive enough to detect differences in samples, but not go so far as to highlight all differences such that a positive association will never be achieved no matter where the soils originate. C-RFLP analysis meets this criterion and therefore has more forensic potential for future applications. Even if HPLC separation and detection is not the forensic method of choice, the important point here is that the sensitivity of the method is crucial.

Regardless of the analysis method eventually chosen for forensic soil analysis, a match criterion must be established for determining whether two samples are the same. We attempted to address this issue with the grid experiment. The data from that experiment (for both the universal and group-specific assays) showed that one absolute number may not be appropriate for establishing relatedness. Again, this has to do with the natural heterogeneity of bacterial communities. Multiple samplings from one location can generate a range of similarity indices. Taking the average of these numbers and using that as rigid match criteria can give misleading results. Future work might evaluate match criteria with an appropriate standard deviation. Furthermore, the data from the grid experiments emphasize the need to collect multiple reference samples from a single location. This will ensure that outlier samples are not inadvertently used as a single reference point, possibly leading to incorrect conclusions. Collection of multiple samples will also ensure back-up samples in case one or more extractions or amplifications fail. Throughout the course of this research, there were several samples that could not be profiled either because of low extraction yield or high humic substance contamination.

This research also identified other considerations, such as meteorological events and time, which should be factored in to soil analysis. These data demonstrated that bacterial communities can change over the course of one year. Changes can also take place over a shorter period of time, as in response to heavy rainfall or snow. These environmental factors would not be as big of a problem for forensic analysis if their impacts on bacterial communities were consistent. Soil communities can have unique responses to these variables. The data show that changes from “native” structure may be either slight or severe. This point emphasizes the importance of timeliness of sample collection. For forensic applications, it will be necessary to collect reference and evidentiary samples over a narrow time frame so that the effects of these potential variables can be minimized.

The application of bacterial community structure to identify possible ecosystems of origin was also explored. When the origin of a soil sample is unknown, it would be of great forensic value to be able to suggest potential environments. For example, if there is soil on a body found in the middle of a parking lot, there is considerable forensic interest to be able to infer that the soil likely came from a forest in the area. The use of 454 pyrosequencing has potential for this type of analysis. As shown with heat map similarity measures, bacterial communities can be clustered to determine which samples are most alike. Biological replicate samples (maintained lawns bordered by trees) were readily identified as a cluster and could be differentiated from other mineral soils. Conversely, the data might also be able to reveal to investigators where a sample *did not* originate from.

II. The future of forensic soil analysis

The purpose of this research was to address how soil might be best used for forensic applications. The quest for the answer to this question began with universal bacterial T-RFLP analysis as described by Horswell *et al.* [Horswell *et al.*, 2002]. The data also addressed the limitations to this type of analysis. Most importantly, it prompted the forensic community to start asking questions about the analysis of soil evidence. Soon after, the promise of universal bacterial DNA typing gave way to the potential for group-specific typing, as suggested by Meyers and Foran (2008). The data presented herein provides support for this type of analysis, adding to the body of scientific evidence that is absolutely essential if soil samples are to be critically examined and ultimately accepted for forensic applications. The kinds of information resulting from our experiments are necessary and valuable regardless of the method ultimately chosen.

There are still many fundamental questions that need to be addressed before soil analysis can be accepted by the forensic community. Spatial analyses of single locations, like our grid experiments, are critical to understanding the heterogeneity of soil. It would be useful to sample across micro- and macro-scales to determine how much DNA profiles can vary across space. These data will provide valuable information for how locations are to be properly sampled. Sampling at various depths would also be useful in this capacity. Our sample set focused on soils collected from the top two inches of the soil surface. Soil taken from deeper below may contain different diversity.

Studies exploring forensically relevant locations are also needed. This research was focused around a convenience soil sample set primarily composed of mineral soils from maintained lawns. More isolated, heterogeneous environments like forests, wetlands and prairies would be of interest to study as crimes can also occur in these remote areas. Heterogeneous environments may pose their own set of difficulties in terms of analysis. For example, bacterial community variation across a spatial scale may be more exaggerated in these environments than across a maintained lawn.

The consensus among the forensic community is that PCR-based typing methods are best suited for all DNA typing methods, including soil analysis. Although there are biases associated with PCR-based methods, an understanding of them can minimize their impact on analysis. The choice of an appropriate statistic for the chosen method also needs careful consideration. The Sorensen similarity index may not be the best selection for statistical comparison given its relatively simple approach to measuring percent relatedness. Multivariate methods, like the Bray-Curtis and Morisita-Horn measures, that cluster samples based on similarity/dissimilarity can be of use to forensics. These measures take into account more than just presence and absence of data points.

Validating potential extraction and molecular typing methods is also critical. There are many variables to consider when formulating any DNA typing protocol. Beginning with soil extraction, there are several commercially available kits as well as chemical extraction protocols that can extract nucleic acids from soils. Five kits were evaluated in addition to the Yeates *et al.* protocol for this work. The Yeates *et al.*

protocol consistently performed the best, generating high molecular weight DNA with the lowest amount of PCR inhibitors. The forensic community must work to validate all possible methods for DNA quality, cost-effectiveness, and reliability. There also must be agreement on proper storage, time between storage and extraction, and what the minimum amount of soil needed for analysis is. In this study, soil samples were stored at four degrees in plastic zip top bags. Other storage methods, like freezing or air-drying samples prior to storage were not evaluated. It is difficult to conclude which method is best for forensic applications. Extractions were carried out within one week of collection from one gram of soil. Smaller starting amounts should be evaluated for ability to generate DNA profiles.

There are also many PCR variables to consider such as starting template, target region, and the number of replicates to perform on a single sample. If T-RFLP and C-RFLP remain viable options for analysis, the restriction enzyme choice will also be critical. Performing double digests on samples, as well as several single digests on one sample may show to be more informative than one digestion. On the other hand, methods that contain a restriction enzyme step introduce the artifact of incomplete digestion. Generating a DNA profile does not necessarily have to include digestion. For example, a length-based assay could be created similar to the current human STR typing method. Specific primer sets could be tagged with different fluorophores creating uniquely sized fragments for each taxon probed. These length variants could be analyzed by capillary electrophoresis. This type of multiplex assay would require a lot more work, as informative taxa have not yet been identified.

The ultimate goal of this translational research is to provide useful information so that a multiplex soil typing assay can be created. The method employed can be based on a modification of T-RFLP, 454 analysis or microarray chips. For example, studies using multiplex (M) T-RFLP have been presented as a way to identify bacteria in environmental samples [Singh *et al.*, 2006; Singh and Thomas, 2006]. This has potential forensic applications as T-RFLP is run on instrumentation crime labs already have. 454 Life Sciences has recently introduced (June 2010) a Junior FLX sequencer that generates less data at a greatly reduced cost. Additionally, one could use multiplex ID tags with the original GS FLX system to allow for more samples to be analyzed at one time, also driving down cost. The use of MID tags for bacterial surveys has been successful in previous studies [Dowd *et al.*, 2008; Huse *et al.*, 2008]. The group-specific data presented can also be used in the development of a bacterial microarray chip containing forensically informative markers. Chips can be created for common phyla and rare ecosystem-specific groups. A similar microarray system was used to survey the human oral microbiome [Huyghe *et al.,* 2008].

Whatever the method chosen, incorporation of a streamlined protocol to be used across the nation is desirable for forensics because it ensures consistency. A soil multiplex typing assay is likely to be centered on group-specific loci, just as human STR typing is built around a core set of markers. The data presented herein demonstrate potential for the use of group-specific markers. There are some bacterial groups that are best suited for sample differentiation (*Firmicutes* and *NOB* groups), while others may be best suited for ecosystem association (*Acidobacteria).* The successful design of any multiplex assay will rely on the researcher’s ability to recognize that not all bacterial groups are created equal. There are many classes and subclasses of phyla, each harboring potentially valuable forensic information. When that is added to the many choices of genetic targets, the combination of assays is limitless. A superior assay does not necessarily have to contain 16 markers like human typing kits do. Rather, the combination of markers must be able to successfully type most soils, and generate enough information about the sample in question so that the forensically relevant question being asked can be accurately answered. Finally, a successful application of bacterial DNA analysis to forensics must be user-friendly and cost-effective, implementing a typing method that delivers the most discriminate information in the shortest amount of time.

Chapter 6 - Materials and Methods

I. Nucleic acid extraction and quantitation from soil

Samples were subject to nucleic acid extraction within one week of collection. Solid materials from the Connecticut River and the sewage treatment plant were isolated from their liquid portions prior to extraction. Two milliliters of each sample were centrifuged at 13,000 x g for 5 minutes. A total of 1.0 gram of wet soil/sediment was used for extraction. The nucleic acid extraction protocol used in this study exactly followed the method published by Yeates *et al*. (1997), except for modifications to reagent volume to accommodate a smaller amount of starting material [Yeates *et al.*, 1997].

The quantity of nucleic acids was estimated using gel electrophoresis. A portion of the extract was run on a 1.0% w/v agarose gel containing ethidium bromide. Query DNA bands were compared to known DNA standards ranging from 12.5 ng to 400 ng.

II. Preparation of amplicons for T-RFLP and C-RFLP universal bacterial typing

All stock solutions were diluted to working concentrations of 2.0 ng/µl of DNA. PCR reactions were set up in triplicate (technical replicates) for each sample, using bacterial-specific universal primers for the 16S ribosomal RNA gene: 27F 5’ – AGAGTTTGATCCTGGCTCAG – 3’ and 926R 5’ – CCGTCAATTCATTTRAGTTT – 3’ (Primer positions based on *Escherichia coli* numbering). PCR for C-RFLP analysis used unlabeled primers, whereas T-RFLP analysis included the use of the following fluorophore-labeled forward primer: 27F 5’ – /56-FAM/AGAGTTTGATCCTGGCTCAG – 3’. Reverse primer 926R was not modified.

The PCR mixture (30µl total volume) contained the following for both T-RFLP and C-RFLP analysis: 10 ng of template, 1X PCR buffer (Bio-Rad Laboratories, Hercules, CA, USA), 5 µl of GeneReleaser® (BioVentures, Inc., Murfreesboro, TN, USA), 1X Cresol Red, 0.22 µM each of primers 27F and 926R, deoxynucleoside triphosphates at a final concentration of 0.13 mM, 1.67 mM MgCl2, and 1.0U of *Taq* polymerase (Bio-Rad Laboratories, Hercules, CA, USA). PCR cycling conditions were as follows: 5 minute initial denaturation at 95º C, followed by 25 cycles of 95 º C for 45 seconds, 52 º C for 1 minute, and 72 º C for 1 minute. Cycle was completed with final extension at 72 º C for 15 minutes. Genomic *Escherichia coli* DNA was used as a positive control.

PCR products were separated on a 1.0% w/v agarose gel containing ethidium bromide. Bands were visualized under ultraviolet light, and amplicons were compared to a 1 Kb+ ladder (Invitrogen, Carlsbad, CA, USA) for size verification.

III. Restriction enzyme digestion

Amplicons from technical replicates were combined and centrifuged briefly to pellet the GeneReleaser®. An aliquot of 60µl of PCR product was transferred to another tube for restriction enzyme digestion. Amplicons generated for C-RFLP and T-RFLP analysis were digested with the restriction enzyme, *AluI* (New England BioLabs, Ipswich, MA, USA). Prior to selecting this restriction enzyme for analysis, several different restriction enzymes were evaluated (data not shown): *AluI, HaeIII, HhaI, HinfI, EcoR1, MspI, and TaqIα.* After testing these enzymes on several soil samples, *AluI* was selected because it created an optimal set of numerous fragments from the 16S rRNA amplicons. Digestion reactions (100µl total volume) contained the following: (approximately) 350 ng of 16S rRNA amplicons (60µl), 1X *AluI* Buffer, and 20U of *AluI*. Reactions were incubated at 37 ºC for 8 hours, and the enzyme then inactivated at 65 ºC for 20 minutes. Digested DNA was purified using the QiaQuick® PCR Purification Kit (Qiagen, Valencia, CA, USA), according to the manufacturer's protocol. Fragments were eluted in 60µl of sterile MilliQ water.

IV. Separation and detection of DNA fragments using C-RFLP

The total eluate from purification of unlabeled digested PCR products was loaded onto the HPLC/WAVE® Nucleic Acid Fragment Analysis System (Transgenomic, Inc., Omaha, NE, USA). Fifty microliters of each sample was injected into the system containing a DNASep® cartridge, and analyzed using the Universal Linear application (non-denaturing conditions). The oven temperature was held constant at 50.0º C. The separation specifications were constant for all samples: "Fast" clean type; 28 minute gradient time; 0.90 slope distribution for Solution A [0.1M triethylammonium acetate (TEAA), obtained from Transgenomic, Inc.] at 55%, and Solution B [0.1M TEAA in 25% acetonitrile, obtained from Transgenomic, Inc.] at 45%. Reagent flow rate through the column was 0.65mL/min. Fragments were continually detected by ultraviolet light at

260nm. Chromatogram data was analyzed using Navigator™ Software (Transgenomic, Inc.). All peaks that met the 0.05 mV (millivolt) threshold were included in analysis.

V. Separation and detection of DNA fragments using T-RFLP

A portion of the purified 56-FAM labeled, digested amplicons was used for T-RFLP analysis. Approximately 100 – 150 ng of DNA (10-15 µl) was combined with formamide (9.5-14.5 µl) and 0.5 µl of GeneScan ™ - 500 LIZ ™ Size Standard (Applied Biosystems, Foster City, CA, USA). Prior to injection, prepared samples were added to a 96-well plate, denatured for 3 minutes at 95° C, and snap cooled on ice for 3 minutes. Terminal fragments were separated by capillary electrophoresis on Applied Biosystem's 3130 Genetic Analyzer using GeneScan ™ software for fragment sizing. Fragments were visualized using GeneMapper ™ ID software version 3.1. Only fragments within the range of the genomic size standards, and above 100 rfu (relative fluorescence units) were considered for preliminary analysis. Peaks were further eliminated from analysis through normalization [Meyers and Foran, 2008].

VI. 454 GS FLX amplicon pyrosequencing – Standard Chemistry Preparation

To prepare samples for standard chemistry Genome Sequencer FLX amplicon pyrosequencing (454 Life Sciences/Roche, Branford, CT, USA), amplicon libraries were created from the following samples: [Set A] - AG Farm, CT River, Field, and Sewage Sludge; [Set B] - Beach Hall (BCH) 2006, BCH 2007, BCH 2008, Cemetery 2008, Cemetery Snow and Cemetery Rain. PCR reactions were set up in triplicate (technical replicates) for each sample, using universal bacterial primers containing 454 adaptor sequences A (forward primer) and B (reverse primer) targeting the V6 region of the 16S gene [Sogin, *et al.* 2006]: A967F 5’ – GCCTCCCTCGCGCCATCAGCAACGCGAAGAACCTTACC – 3’; B1046R 5’ – GCCTTGCCAGCCCGCTCAGCGACAGCCATGCANCACCT – 3’. [Note: adaptor sequences used with standard chemistry for amplicon sequencing]. The PCR mixture (25µl total volume) contained the following: 8 ng of template, 1X PCR buffer (Bio-Rad Laboratories, Hercules, CA, USA), 5 µl of GeneReleaser® (BioVentures, Inc., Murfreesboro, TN, USA), 1X Cresol Red, 0.4 µM each of primer, deoxynucleoside triphosphates at a final concentration of 0.16 mM, 2.0 mM MgCl2, and 1.0U of *Taq* polymerase (Bio-Rad Laboratories, Hercules, CA, USA). PCR cycling conditions were as follows: 5 minute initial denaturation at 94º C, followed by 23 cycles of 94 º C for 30 seconds, 57 º C for 45 seconds, and 72 º C for 1 minute. Cycle was completed with final extension at 72 º C for 15 minutes.

PCR products were separated on a 1.0% w/v agarose gel containing ethidium bromide. Bands were visualized under ultraviolet light, and amplicons were compared to a 1 Kb+ ladder (Invitrogen, Carlsbad, CA, USA) for size verification. Amplicons from technical replicates were combined and centrifuged briefly to pellet the GeneReleaser®. DNA was purified using the QiaQuick® PCR Purification Kit (Qiagen, Valencia, CA, USA), according to the manufacturer’s protocol. Fragments were eluted in 50µl of sterile MilliQ water.

For samples in Set A, amplicon library concentrations were measured using an Agilent Bioanalyzer system (Agilent Technologies, Foster City, CA, USA), following manufacturers protocol. For samples in Set B, an Experion Automated Gel Electrophoresis System (Bio-Rad Laboratories, Hercules, CA, USA) was used following manufacturer's protocol.

For the samples in Set A, a 2.0 E+05 double stranded DNA molecules/µl dilution of the amplicon library was created. A target copy number of 1.0 molecule per bead was used in the emulsion PCR. For samples in Set B, a 4.0 E+06 double stranded DNA molecules/µl dilution of the amplicon library was created. A target copy number of 2.0 molecules per bead was used in the emulsion PCR. All DNA libraries were prepared for unidirectional sequencing from the A end. Emulsion PCR amplification was carried out using manufacturer's protocol. After emulsion PCR was complete, emulsions were broken and positive beads were enriched for following the manufacturer's protocol. A 4-region 454 sequencing run was done using a 70x75 GS PicoTiterPlate (PTP) on the GS FLX System for Set A. An 8-region sequencing run was done using a 70x75 GS PicoTiterPlate (PTP) for Set B. All sequencing procedures followed the manufacturer's instructions.

VII. 454 GS FLX amplicon pyrosequencing – Titanium Chemistry Preparation

To prepare samples for standard chemistry Genome Sequencer FLX amplicon pyrosequencing (454 Life Sciences/Roche, Branford, CT, USA), amplicon libraries were created from the following samples: Lawn 1, Lawn 9, Lawn 14 and Lawn 17. PCR reactions were set up in triplicate (technical replicates) for each sample, using universal bacterial primers containing 454 adaptor sequences A (forward primer) and B (reverse primer) targeting the V1-V2 region of the 16S gene [Sundquist, 2007]: A_8F 5' – CGTATCGCCTCCCTCGCGCCATCAGAGAGTTTGATCMTGGCTCAG – 3'; B_361R 5' – CTATGCGCCTTGCCAGCCCGCTCAGCYIACTGCTGCCTCCCGTAG – 3'. The PCR mixture (30µl total volume) contained the following: 10 ng of template, 1X PCR buffer (Bio-Rad Laboratories, Hercules, CA, USA), 5 µl of GeneReleaser® (BioVentures, Inc., Murfreesboro, TN, USA), 1X Cresol Red, 0.33 µM each of primer, deoxynucleoside triphosphates at a final concentration of 0.13 mM, 1.67 mM MgCl2, and 1.0U of *Taq* polymerase (Bio-Rad Laboratories, Hercules, CA, USA). PCR cycling conditions were as follows: 5 minute initial denaturation at 95º C, followed by 23 cycles of 95 º C for 45 seconds, 64 º C for 1 minute, and 72 º C for 1 minute. Cycle was completed with final extension at 72 º C for 15 minutes.

PCR products were separated on a 1.0% w/v agarose gel containing ethidium bromide. Bands were visualized under ultraviolet light, and amplicons were compared to a 1 Kb+ ladder (Invitrogen, Carlsbad, CA, USA) for size verification. Amplicons from technical replicates were combined and centrifuged briefly to pellet the GeneReleaser®. DNA was purified using the QiaQuick® PCR Purification Kit (Qiagen, Valencia, CA, USA), according to the manufacturer's protocol. Fragments were eluted in 50µl of sterile MilliQ water.

For all Lawn samples, an Experion Automated Gel Electrophoresis System (Bio-Rad Laboratories, Hercules, CA, USA) was used following manufacturer's protocol.

A 4.0 E+06 double stranded DNA molecules/µl dilution of the amplicon library was created. A target copy number of 4.0 molecules per bead was used in the emulsion PCR. The DNA libraries were prepared for unidirectional sequencing from the A end. Emulsion PCR amplification was carried out using manufacturer's protocol. After emulsion PCR was complete, emulsions were broken and positive beads were enriched for following the manufacturer's protocol. An 8-region 454 sequencing run was done using a 70x75 GS PicoTiterPlate (PTP) on the GS FLX System. All sequencing procedures followed the manufacturer's instructions.

VIII. Preparation of amplicons for group-specific C-RFLP analysis

All stock solutions were diluted to working concentrations of 2.0 ng/µl of DNA. PCR reactions were set up in triplicate (technical replicates) for each sample, using the group-specific primers in Table 15. The PCR mixture (30µl total volume) contained the following: 10 ng of template, 1X PCR buffer (Bio-Rad Laboratories, Hercules, CA, USA), 5 µl of GeneReleaser® (BioVentures, Inc., Murfreesboro, TN, USA), 1X Cresol Red, 0.33 µM each of primer, deoxynucleoside triphosphates at a final concentration of 0.13 mM, 1.67 mM MgCl2, and 1.0U of *Taq* polymerase (Bio-Rad Laboratories, Hercules, CA, USA). PCR cycling conditions were as follows: 5 minute initial denaturation at 95º C, followed by 35 cycles of 95 º C for 45 seconds, 57 º C for 1 minute, and 72 º C for 1 minute. Cycle was completed with final extension at 72 º C for 15 minutes.

PCR products were separated on a 1.0% w/v agarose gel containing ethidium bromide. Bands were visualized under ultraviolet light, and amplicons were compared to a 1 Kb+ ladder (Invitrogen, Carlsbad, CA, USA) for size verification. Amplicons from technical replicates were combined and centrifuged briefly to pellet the GeneReleaser®. DNA was purified using the QiaQuick® PCR Purification Kit (Qiagen, Valencia, CA, USA), according to the manufacturer's protocol. Fragments were eluted in 50µl of sterile MilliQ water. DNA fragments were run on the HPLC system as described in section 6.5.

IX. Sequence identification of group-specific amplicons as separated by HPLC

Undigested group-specific amplicons were prepared as described in section 6.10. Triplicate PCR reactions were combined and DNA was purified using the QiaQuick® PCR Purification Kit (Qiagen, Valencia, CA, USA), according to the manufacturer's protocol. Fragments were eluted in 50µl of sterile MilliQ water.

Amplicons were initially separated by HPLC in order to generate a chromatogram that pinpointed the location of all peaks. The location of these peaks (retention time) dictated how the fragment collection protocol was designed. Once a fragment collection protocol was established for each group-specific amplicon, a new sample was run on the HPLC. Fragments were collected using a 96-well plate, with each vial containing 200µl of eluate.

IX.a. Single peak sequencing

Using Navigator™ software, all vials containing the desired DNA molecules were identified by vial number. Single peaks that were distributed among multiple vials were combined into a clean 1.5mL tube. Collected peaks were purified using a Microcon YM-30 microcentrifuge unit. The entire volume of each peak was transferred to each microcon unit. Samples were centrifuged for 25 minutes at 4.6 rpm. If the peak volume exceeded 600µl, two rounds of centrifugation were done with the remaining volume. Once the sample passed through, the filter was washed twice with 250µl of sterile water by centrifugation at 4.6 rpm for 25 minutes. Samples were concentrated to a final volume of 30µl using MilliQ water.

To generate enough template for DNA sequencing, the purified DNA was used as template for group-specific PCR. Amplification followed the protocol as described in section 6.8.

Sanger sequencing was then performed on the HPLC peak PCR product using the BigDye Terminator Kit v3.1 (Applied Biosystems, Foster City, CA, USA). DNA sequencing reactions (6µl total volume) contained the following: 2.5µl DNA template, 1X BigDye sequencing buffer, 1µl BigDye sequencing enzyme mix, and 0.83 µM of either the appropriate forward or reverse group-specific primer. PCR cycling conditions were as follows: 5 minute initial denaturation at 96º C, followed by 25 cycles of 96 º C for 10 seconds, 50 º C for 5 seconds, and 60 º C for 4 minutes.

Sequencing products were purified from sequencing reactions by first bringing up the reaction volume to 20µl with sterile MilliQ water. Entire volume was transferred into a clean 1.5mL tube. Two microliters of 3M sodium acetate was added to the reaction as well as 50µl of 95% ethanol. Samples were briefly vortexed and allowed to incubate at room temperature for 10 minutes in the dark. Samples were centrifuged for 10 minutes at 13.2 x g. Supernatant was removed from the tube, carefully avoiding the DNA pellet. Two hundred fifty microliters of 70% ethanol was added to the pellet and briefly vortexed. Tubes were centrifuged for 5 minutes at 13.2 x g. Supernatant was removed from the tube (avoiding the DNA pellet) and the pellet was allowed to fully dry. DNA pellets were resuspended in 20µl of HiDi Formamide by thoroughly vortexing the sample. Samples were loaded onto a 96-well plate and placed into an ABI 3130 DNA Capillary Electrophoresis Instrument (Applied Biosystems, Foster City, CA, USA).

IX.b. Single vial sequencing

Using Navigator™ software, all vials containing the desired DNA molecules were identified by vial number. Single vials were not combined during this protocol. Single vials were concentrated, PCR amplified, DNA sequenced and sequencing products were purified exactly as described in section 6.9.1.

Chapter 7 – References cited

Acosta-Martinez, V., Dowd, S., Sun, Y. and Allen, V., Tag-encoded pyrosequencing analysis of bacterial diversity in a single soil type as affected by management and land use. Soil Biology and Biochemistry 40 (2008) 2762 – 2770.

Baker, G. C., Smith, J. J. and Cowan, D. A., Review and re-analysis of domain-specific 16S primers. Journal of Microbial Methods 55 (2003) 541 – 555.

Barlaan, E. A., Sugimori, M., Furukawa, S. and Takeuchi, K., Profiling and monitoring of microbial populations by denaturing high-performance liquid chromatography. Journal of Microbial Methods 61 (2005) 399 – 412.

Barns, S. M., Takala, S. L. and Kuske, C. R., Wide distribution and diversity of members of the bacterial kingdom *Acidobacterium* in the environment. Applied and Environmental Microbiology 65 (4) (1999) 1731 – 1737.

Blackwood, C. B., Oaks, A. and Buyer, J. S., Phylum- and class-specific PCR primers for general microbial community analysis. Applied and Environmental Microbiology 71 (10) (2005) 6193 – 6198.

Buckley, D. H. and Schmidt, T. M., Environmental factors influencing the distribution of rRNA from Verrucomicrobia in soil. FEMS Microbiology Ecology 35 (2001) 105 – 112.

Chu, H., Fujii, T., Morimoto, S., Lin, X., Yagi, K., Hu, J. and Zhang, J., Community structure of ammonia-oxidizing bacteria under long-term application of mineral fertilizer and organic manure in a sandy loam soil. Applied and Environmental Microbiology 73 (2) (2007 485 – 491.

Claesson, M. J., O'Sullivan, O., Wang, Q., Nikkila, J., Marchesi, J. R., Smidt, H., de Vos, W. M., Ross, R. P. and O'Toole, P. W., Comparative analysis of pyrosequencing and a phylogenetic microarray for exploring microbial community structures in the human distal intestine. PLoS One 4 (8) (2009).

Cruz-Martinez, K., Suttle, K. B., Broide, E. L., Power, M. E., Andersen, G. L. and Banfield, J. F., Despite strong seasonal responses, soil microbial consortia are more resilient to long-term changes in rainfall than overlying grassland. International Society for Microbial Ecology, 3 (2009) 738 – 744.

Domann, E., Hong, G., Imirzalioglu C., Turschnre, S., Kuhle, J., Watzel, C., Hain, T., Hossain, H. and Chakraborty, T., Culture-independent identification of pathogenic bacteria and polymicrobial infections in the genitourinary tract of renal transplant recipients. Journal of Clinical Microbiology 41 (12) (2003) 5500 – 5510.

Dowd, S. E., Callaway, T. R., Wolcott, R. D., Sun, Y., McKeehen, T., Hagevoort, R. G., and Edrington, T. S., Evaluation of the bacterial diversity in the feces of cattle using 16S rDNA bacterial tag-encoded FLX amplicon sequencing (bTEFAP). BMC Microbiology 8 (125) (2008).

Egert, M. and Friedrich, M. W., Formation of pseudo-terminal restriction fragments, a PCR-related bias affecting terminal restriction fragment length polymorphism analysis of microbial community structure. Applied and Environmental Microbiology 69 (5) (2003) 2555 – 2562.

Ettema, C. H. and Wardle, D. A., Spatial soil ecology. TRENDS in Ecology and Evolution 17 (4) (2002) 177 – 183.

Farrelly, V., Rainey, F. A. and Stackenbrandt, E., Effect of genome size and *rrn* copy number on PCR amplification of 16S rRNA genes from a mixture of bacterial species. Applied and Environmental Microbiology 61(7) 2798 – 2801.

Feinstein, L. M., Sul, W. J., Blackwood, C. B., Assessment of bias assicoated with incomplete extraction of microbial DNA from soil. Applied and Environmental Microbiology 75 (16) (2009) 5428 – 5433.

Fierer, N. and Jackson, R. B., The diversity and biogeography of soil bacterial communities. PNAS 103 (3) (2006) 626 – 631.

Fierer, N., Jackson, J. A., Vilgalys, R. and Jackson, R. B., Assessment of soil microbial community structure by use of taxon-specific quantitative PCR assays. Applied and Environmental Microbiology 71 (7) (2005) 4117 – 4120.

Fierer, N., Lauber, C. L., Zhou, N., McDonald, D., Costello, E. K. and Knight, R., Forensic identification using skin bacterial communities. Proc. Natl. Acad. Sci. 107 (2010) 6477 – 6481.

Finley, J. A., Geologic material as physical evidence. FBI Law Enforcement Bulletin 7 (3) (2004) 2 – 7.

Girvan, M. S., Bullimore, J., Pretty, J. N., Osborn, A. M. and Ball, A. S., Soil type is the primary determinant of the composition of the total and active bacterial communities in arable soils. Applied and Environmental Microbiology 69 (3) (2003) 1800 – 1809.

Goldenberg, O., Herrmann, S., Marjoram, G., Noyer-Weidner, M., Hong, G., Bereswill, S. and Gobel, U. B., Molecular monitoring of the intestinal flora by denaturing high performance liquid chromatography. Journal of Microbial Methods 68 (2007) 94 – 105.

Griffiths, R. I., Whiteley, A. S., O'Donnell, A. G. and Bailey, M. J., Influence of depth and sampling time on bacterial community structure in an upland grassland soil. FEMS Microbiology Ecology 43 (2003) 35 – 43.

Grundmann, G. L., Neyra, M. and Normand, P., High-resolution phylogenetic analysis of $\mathrm{NO2}^{-}$-oxidizing *Nitrobacter* species using the *rrs-rrl* IGS sequence and *rrl* genes. International Journal of Systematic and Evolutionary Microbiology 50 (2000) 1893 – 1898.

Grundmann, G. L. and Normand, P., Microscale diversity of the genus *Nitrobacter* in soil on the basis of analysis of genes encoding rRNA. Applied and Environmental Microbiology 66 (10) (2000) 4543 – 4546.

Hackl, E., Zechmeister-Boltenstern, S., Bodrossy, L. and Sessitsch, A., Comparison of diversities and compositions of bacterial populations inhabiting natural forest soils. Applied and Environmental Microbiology 70 (9) (2004) 5057 – 5065.

Halverson, J. L. and Basten, C., Forensic DNA identification of animal-derived trace evidence: tools for linking victims and suspects. Croatian Medical Journal 46 (4) (2005) 598 – 605.

Heath, L. E. and Saunders, V. A., Assessing the potential of bacterial DNA profiling for forensic comparisons. Journal of Forensic Sciences 51 (5) (2006) 1062-1068.

Hill, J., Strausbaugh, L. and Graf, J., "Soil DNA Typing in Forensic Science", Chapter 10. Nonhuman DNA Typing: Theory and Casework Applications. CRC Press, 2008.

Horner-Devine, M. C., Lage, M., Hughes, J. B. and Bohannan J. M., A taxa-area relationship for bacteria. Nature 432 (2004) 750 – 753.

Horswell, J., Corinder, S. J., Maas, E. W., Martin, T. M., K., Sutherland, K. B. W., Speir, T. W., Nogales, B. and Osborn, A. M., Forensic comparison of soils by bacterial community DNA profiling. Journal of Forensic Sciences 47 (2) (2002) 350-353.

Horz, H.-P., Rotthauwe, J.-H., Lukow, T. and Liesack, W., Identification of major subgroups of ammonia-oxidizing bacteria in environmental samples by T-RFLP analysis of *amoA* PCR products. Journal of Microbial Methods 39 (2000) 197 – 204.

Humbolt, C. and Guyot, J-P., Pyrosequencing of tagged 16S rRNA gene amplicons for rapid deciphering of the microbiomes of fermented foods such as pearl millet slurries. Applied and Environmental Microbiology 75 (13) (2009) 4354 – 4361.

Hurtle, W., Shoemaker, D., Henchal, E. and Norwood, D., Denaturing HPLC for identifying bacteria. BioTechniques 33 (2002) 386 – 391.

Huse, S. M., Dethlefsen, L., Huber, J. A., Welch, D. A., Relman, D. A. and Sogin, M. L., Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. PLoS Genetics 4 (11) (2008).

Huyghe, A., Francois, P., Charbonnier, Y., Tangomo-Bento, M., Bonetti, E.-J., Paster, B. J., Bolivar, I., Baratti-Mayer, D., Pittet, D., Schrenzel, J., and the Geneva Study Group on Noma, Novel microarray design strategy to study complex bacterial communities. Applied and Environmental Microbiology 74 (6) (2008) 1876 – 1885.

Janssen, P. H., Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes. Applied and Environmental Microbiology 72 (3) (2006) 1719 – 1728.

Jones, R. T., Robeson, M. S., Lauber, C. L., Hamady, M. and Knight, R., A comprehensive survey of soil acidobacterial diversity using pyrosequencing and clone library analysis. International Society for Microbial Ecology 3 (2009) 442 – 453.

Kielak, A., Pijl, A. S., van Veem, J. A. and Kowalchuck, G. A., Phylogenetic diversity of *Acidobacteria* in a former agricultural soil. International Society for Microbial Ecology 3 (2009) 378 – 382.

Kirk, J. L., Beaudette, L. A., Hart, M., Moutoglis, P., Klironomos, J. N., Lee, H. and Trevors, J. T., Methods for studying soil microbial diversity. Journal of Microbial Methods 58 (2004) 169 – 188.

Klappenbach, J. A., Dunbar, J. M. and Schmidt, T. M., rRNA operon copy number reflects ecological strategies of bacteria. Applied and Environmental Microbiology 66 (4) (2000) 1328 – 1333.

Lerner, A., Shor, Y., Vinokurov, A., Okon, Y. and Jurkevitch, E., Can denaturing gradient gel electrophoresis (DGGE) analysis of amplified 16S rDNA of soil bacterial populations be used in forensic investigations? Soil Biology and Biochemistry 38 (2006) 1188 – 1192.

Liles, M. R., Manske, B. F., Bintrim, S. B., Handelsman, J. and Goodman, R. M., A census of rRNA genes and linked genomic sequences within a soil metagenomic library. Applied and Environmental Microbiology 69 (5) (2003) 2684 – 2691.

Lipson, D. A. and Schmidt, S. K., Seasonal changes in an alpine soil bacterial community in the Colorado Rocky Mountains. Applied and Environmental Microbiology 70 (5) (2004) 2867-2879.

Liu, W,-T., Marsh, T. L., Cheng, H., and Forney, L. J., Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. Applied and Environmental Microbiology 63 (11) (1997) 4516 – 4522.

Liu, Z., Lozupone, C., Hamady, M., Bushman, F. D. and Knight, R., Short pyrosequencing reads suffice for accurate microbial community analysis. Nucleic Acids Research 35 (18) (2007).

Lozupone, C. A. and Knight, R., Global patterns in bacterial diversity. Proc. Natl. Acad. Sci. 104 (27) (2007) 11436 – 11440.

Martin-Laurent, F., Phillippot, L., Hallet, S., Chaussod, R., Germon, J. C., Soulas, G. and Catroux, G., DNA extraction from soils: old bias for new microbial diversity analysis methods. Applied and Environmental Microbiology 67 (5) (2001) 2354 – 2359.

Menking, D. E., Emanuel, P. A., Valdes, J. J. and Kracke, S. K., Rapid cleanup of bacterial DNA from field samples. Resources, Conservation and Recycling 27 (1999) 179 – 186.

Menotti-Raymond, M. A., David, V. A. and O'Brien, S. J., Pet cat hair implicates murder suspect. Nature 386 (1997) 774.

Meyers, M. S. and Foran, D. R., Spatial and temporal influence on bacterial profiling for forensic soil samples. Journal of Forensic Sciences 53 (3) (2008) 652-660.

Miller, S. R., Strong, A. L., Jones, K. L. and Ungerer, M. C., Bar-coded pyrosequencing reveals shared bacterial community properties along the temperature gradient of two alkaline hot springs in Yellowstone National Park. Applied and Environmental Microbiology 75 (13) (2009) 4565 – 4572.

Miller Coyle, Heather (Editor). *Nonhuman DNA Typing: Theory and Casework Applications.* Chapter 10, "Soil DNA Typing in Forensic Science." Boca Raton: Taylor & Francis Group, 2008.

Miller Coyle, H., Ladd, C., Palmbach, T. and Lee, H. C., The green revolution: botanical contributions to forensics and drug enforcement. Croatian Medical Journal 42 (3) (2001) 340 – 345.

Moore, Solomon. "Science Found Wanting at Nations Crime Labs." *The New York Times*. Published 4 Feb. 2009. Accessed 8 July 2010. Article available online at: http://www.nytimes.com.

Morales, S. E., Cosart, T. F., Johnson, J. V. and Holben, W. E., Extensive phylogenetic analysis of a soil bacterial community illustrates extreme taxon evenness and the effects of amplicon length, degree of coverage, and DNA fractionation on classification and ecological parameters. Applied and Environmental Microbiology 75 (3) (2009) 668 – 675.

Morgan, R. M. and Bull, P. A., The philosophy, nature and practice of forensic sediment analysis. Progress in Physical Geography 31 (1) (2007) 43 – 58.

Muyzer, G., De Waal, E. C., Uitterlinden, A. G., Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. Applied and Environmental Microbiology 59 (3) (1993) 695 – 700.

Nannipieri, P., Ascher, J., Ceccherini, M. T., Landi, L., Pietramellara, G. and Renella, G., Microbial diversity and soil functions. European Journal of Soil Science 54 (2003) 655 – 670.

Poly, F., Wertz, S., Brothier, E., and Degrange, V., First exploration of *Nitrobacter* diversity in soils by a PCR cloning-sequencing approach targeting functional gene *nxrA*. FEMS Microbial Ecology 63 (2008) 132 – 140.

Roose-Amsaleg, C. L., Garnier-Silliam, E. and Harry, M., Extraction and purification of microbial DNA from soil and sediment samples. Applied Soil Ecology 18 (2001) 47 – 60.

Singh, B. K., Nazaries, L., Munro, S., Anderson, I. C. and Campbell C. D., Use of multiplex terminal restriction fragment length polymorphism for rapid and simultaneous analysis of different components of the soil microbial community. Applied and Environmental Microbiology 72 (11) (2006) 7278 – 7285.

Singh, B. K. and Thomas, N., Multiplex-terminal restriction fragment length polymorphism. Nature Protocols 1 (5) (2006) 2428 – 2433.

Smit, E., Leeflang, P., Gommans, S., Van Den Broek, J., Van Mil, S. and Wernars, K., Diversity and seasonal fluctuations of the dormant members of the bacterial soil community in a wheat field as determined by cultivation and molecular methods. Applied and Environmental Microbiology 67 (5) (2001) 2284-2291.

Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M. and Neal, P. R., Microbial diversity in the deep sea and the underexplored “rare biosphere”. Proc. Natl. Acad. Sci. 103 (32) (2006) 12115 – 12120.

Sugita, R. and Marumo, Y., Validity of color examination for forensic soil identification. Forensic Science International 83 (1996) 201 – 210.

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

Tate, Robert L. *Soil Microbiology*, Second Edition. New York: John Wiley & Sons, 2000.

Teske, A., Alm, E., Regan, J. M., Toze, S., Rittmann, B. E. and Stahl, D. A., Evolutionary relationships among ammonia- and nitrite-oxidizing bacteria. Journal of Bacteriology 176 (21) (1994) 6623 – 6630.

Torsvik, V., Goksoyr, J. and Daae, F. L., High diversity in DNA of soil bacteria. Applied and Environmental Microbiology 56 (3) (1990) 782 – 787.

Torsvik, V. and Ovreas, L., Microbial diversity and function in soil: from genes to ecosystems. Current Opinion in Microbiology 5 (2002) 240 – 245.

Walker, V. K., Palmer, G. R. and Voordouw, G., Freeze-thaw tolerance and clues to the winter survival of a soil community. Applied and Environmental Microbiology 72 (3) (2006) 1784 – 1792.

Wang, G., Garrity, G. M., Tiedje, J. M., Cole, J. R., Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and Environmental Microbiology 73 (16) (2007) 5261 – 5267.

Wintzingerode, F. v., Gobel, U. B. and Stackebrandt, E., Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. FEMS Microbiology Reviews 21 (1997) 213 – 229.

Yeates, C., Gillings, M. R., Davidson, A. D., Altavilla, N. and Veal, D. A., PCR amplification of crude microbial DNA extracted from soil. Letters in Applied Microbiology 25 (1997) 303-307.

Yoon, K. C., Botanical witness for the prosecution. Science 260 (5110) (1993) 894 – 895.

Zipper, H., Buta, c., Lammle, K., Brunner, H., Bernhagen, J. and Vitzthum, F., Mechanisms underlying the impact of humic acids on DNA quantification by SYBR Green I and consequences for the analysis of soils and aquatic sediments. Nucleic Acids Research 31 (7) (2003).

Chapter 8 – Appendix

Supplementary Figures S27A – S27H – *Validation data for C-RFLP bacterial profiling using HPLC to separate and detect DNA fragments.*

Bacterial community profiles generated by 16S rRNA amplicon fragment separation and detection on the HPLC/WAVE® System. Included for all sampling locations is a chromatogram of consolidated, independently generated profiles. The 5 tallest peaks were chosen for validation testing. Table A lists fragment retention times for the independently run samples in minutes (as determined by Navigator™ Software). Each peak retention time is listed with a rounded value (to the nearest tenth of a minute). Table B lists the elapsed time between the detection of select fragments among independently run trials. Elapsed time reported in minutes. The difference between the average elapsed time and the query peaks is reported in seconds (parentheses).

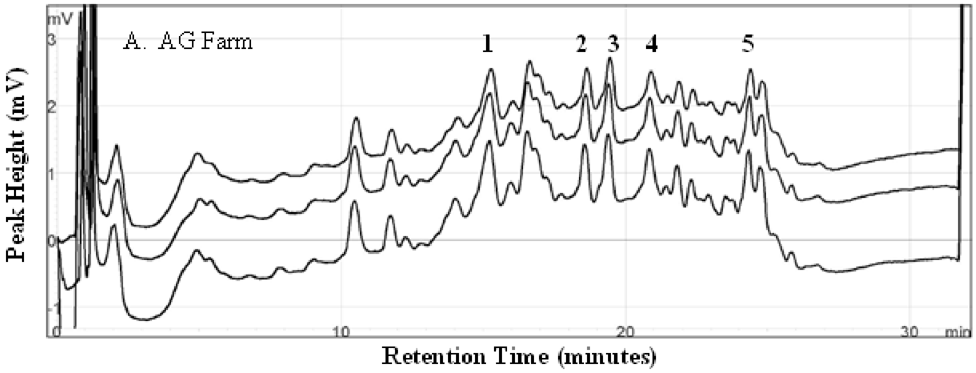


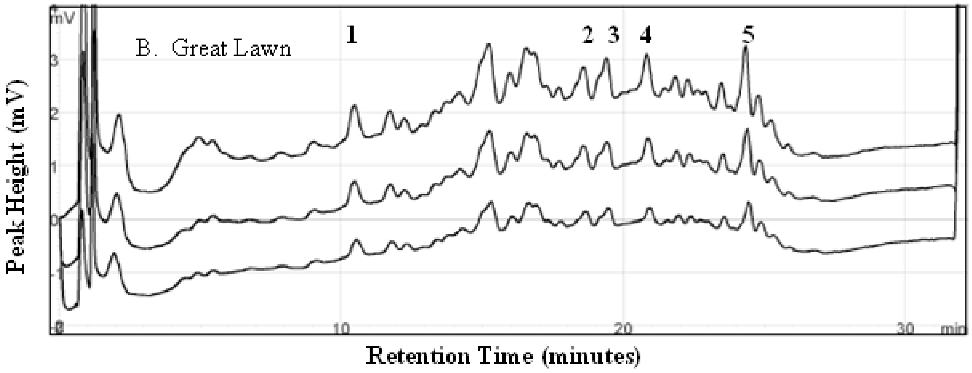
Table A: Individual Peak Retention Time

Soil Sample Trial	No. 1 (rounded)	No. 2 (rounded)	_No. 3 (rounded)	No. 4 (rounded)	No. 5 (rounded)
Trial 1	15.207 (15.2)	18.573 (18.6)	19.393 (19.4)	20.840 (20.8)	24.374 (24.4)
Trial 2	15.233 (15.2)	18.593 (18.6)	19.407 (19.4)	20.853 (20.9)	24.360 (24.4)
Trial 3	15.287 (15.3)	18.640 (18.6)	19.453 (19.5)	20.893 (20.9)	24.413 (24.4)

Table B: Pattern Analysis - Elapsed Time Between Fragment Detection

Soil Sample Trial	Peak 1 and 2 (seconds)	Peak 2 and 3 (seconds)	Peak 3 and 4 (seconds)	Peak 4 and 5 (seconds)
Trial 1	3.366 (0.36*s*)	0.820 (0.24*s*)	1.447 (0.18*s*)	3.534 (0.84*s*)
Trial 2	3.360 (0.00*s*)	0.814 (0.12*s*)	1.446 (0.12*s*)	3.507 (0.78*s*)
Trial 3	3.353 (0.42*s*)	0.813 (0.18*s*)	1.440 (0.24*s*)	3.520 (0.00*s*)

Figure S27A – Validation using Agricultural Farm soil



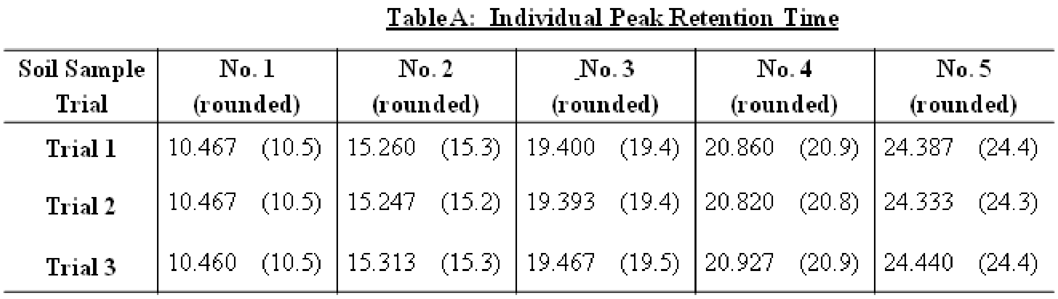


Table A: Individual Peak Retention Time

Soil Sample Trial	No. 1 (rounded)		No. 2 (rounded)		_No. 3 (rounded)		No. 4 (rounded)		No. 5 (rounded)	
Trial 1	10.467	(10.5)	15.260	(15.3)	19.400	(19.4)	20.860	(20.9)	24.387	(24.4)
Trial 2	10.467	(10.5)	15.247	(15.2)	19.393	(19.4)	20.820	(20.8)	24.333	(24.3)
Trial 3	10.460	(10.5)	15.313	(15.3)	19.467	(19.5)	20.927	(20.9)	24.440	(24.4)

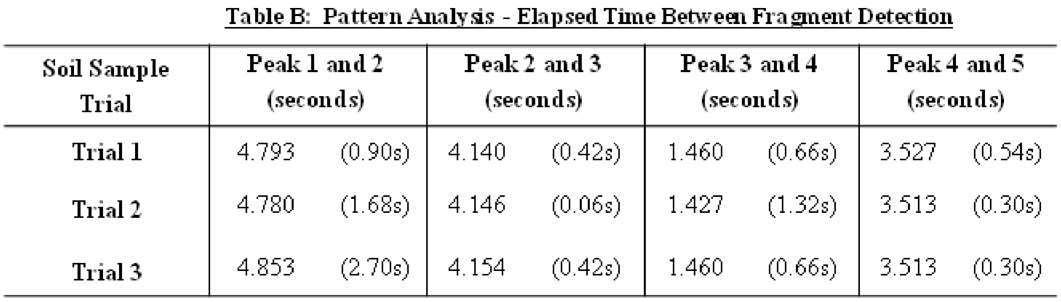


Table B: Pattern Analysis - Elapsed Time Between Fragment Detection

Soil Sample Trial	Peak 1 and 2 (seconds)		Peak 2 and 3 (seconds)		Peak 3 and 4 (seconds)		Peak 4 and 5 (seconds)	
Trial 1	4.793	(0.90s)	4.140	(0.42s)	1.460	(0.66s)	3.527	(0.54s)
Trial 2	4.780	(1.68s)	4.146	(0.06s)	1.427	(1.32s)	3.513	(0.30s)
Trial 3	4.853	(2.70s)	4.154	(0.42s)	1.460	(0.66s)	3.513	(0.30s)

Figure S27B – Validation using Beach Hall Great Lawn soil

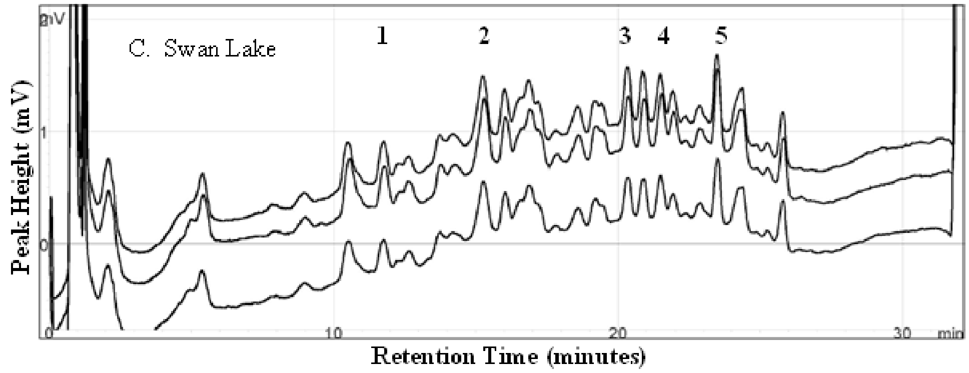
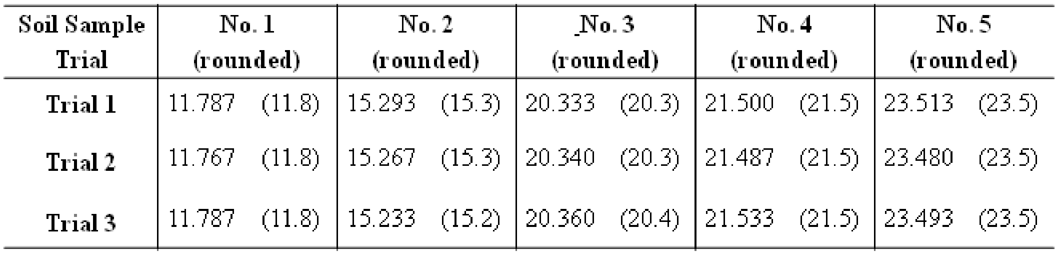
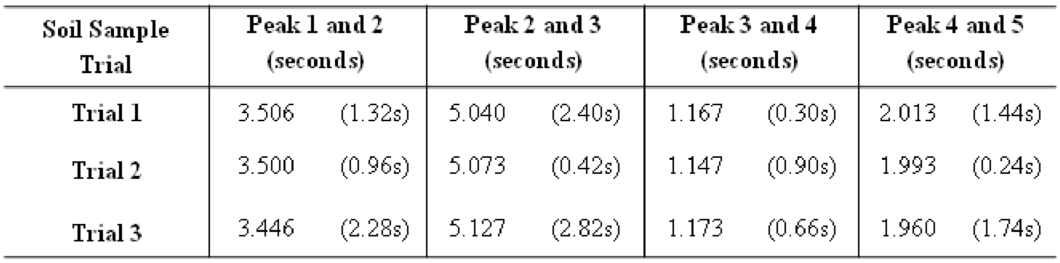


Table A: Individual Peak Retention Time



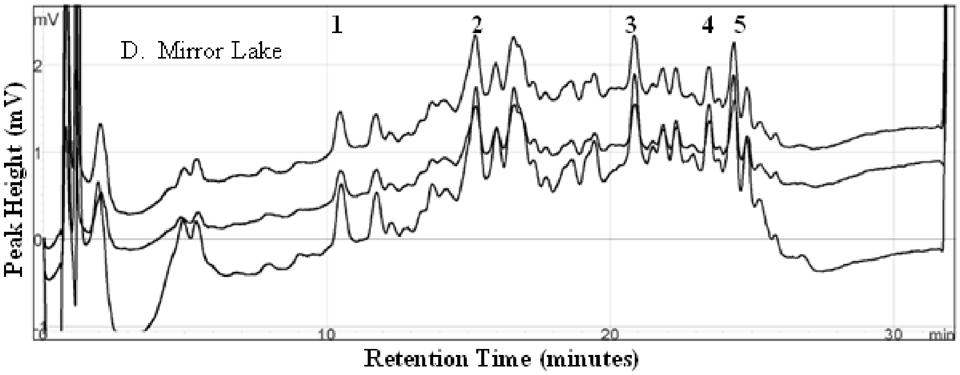
Soil Sample Trial	No. 1 (rounded)		No. 2 (rounded)		No. 3 (rounded)		No. 4 (rounded)		No. 5 (rounded)	
Trial 1	11.787	(11.8)	15.293	(15.3)	20.333	(20.3)	21.500	(21.5)	23.513	(23.5)
Trial 2	11.767	(11.8)	15.267	(15.3)	20.340	(20.3)	21.487	(21.5)	23.480	(23.5)
Trial 3	11.787	(11.8)	15.233	(15.2)	20.360	(20.4)	21.533	(21.5)	23.493	(23.5)

Table B: Pattern Analysis - Elapsed Time Between Fragment Detection



Soil Sample Trial	Peak 1 and 2 (seconds)		Peak 2 and 3 (seconds)		Peak 3 and 4 (seconds)		Peak 4 and 5 (seconds)	
Trial 1	3.506	(1.32s)	5.040	(2.40s)	1.167	(0.30s)	2.013	(1.44s)
Trial 2	3.500	(0.96s)	5.073	(0.42s)	1.147	(0.90s)	1.993	(0.24s)
Trial 3	3.446	(2.28s)	5.127	(2.82s)	1.173	(0.66s)	1.960	(1.74s)

Figure S27C – Validation using Swan Lake soil



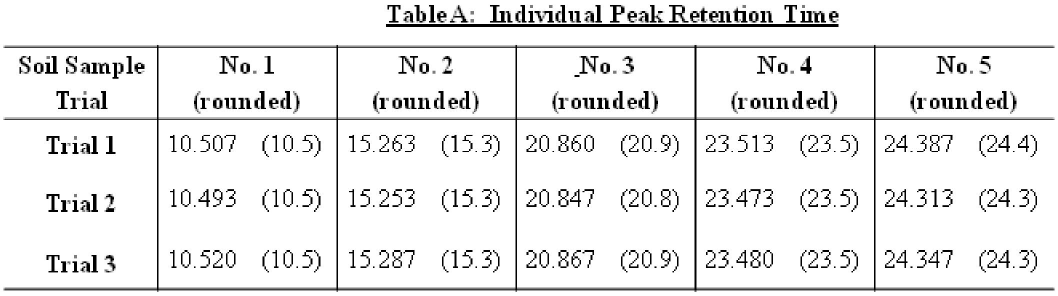


Table A: Individual Peak Retention Time

Soil Sample Trial	No. 1 (rounded)		No. 2 (rounded)		No. 3 (rounded)		No. 4 (rounded)		No. 5 (rounded)	
Trial 1	10.507	(10.5)	15.263	(15.3)	20.860	(20.9)	23.513	(23.5)	24.387	(24.4)
Trial 2	10.493	(10.5)	15.253	(15.3)	20.847	(20.8)	23.473	(23.5)	24.313	(24.3)
Trial 3	10.520	(10.5)	15.287	(15.3)	20.867	(20.9)	23.480	(23.5)	24.347	(24.3)

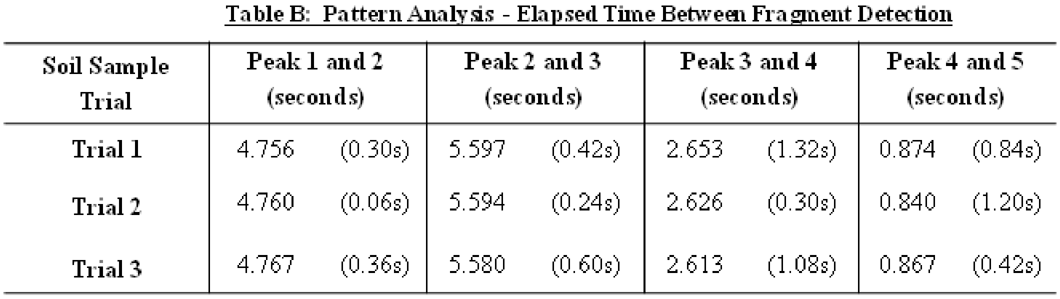
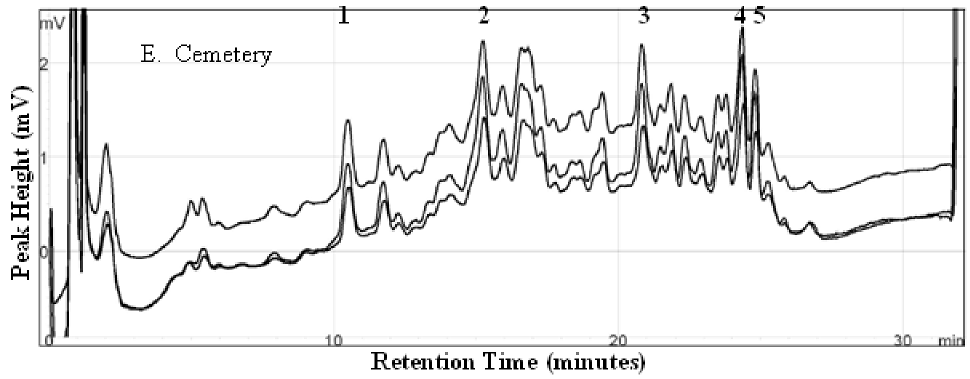


Table B: Pattern Analysis - Elapsed Time Between Fragment Detection

Soil Sample Trial	Peak 1 and 2 (seconds)		Peak 2 and 3 (seconds)		Peak 3 and 4 (seconds)		Peak 4 and 5 (seconds)	
Trial 1	4.756	(0.30s)	5.597	(0.42s)	2.653	(1.32s)	0.874	(0.84s)
Trial 2	4.760	(0.06s)	5.594	(0.24s)	2.626	(0.30s)	0.840	(1.20s)
Trial 3	4.767	(0.36s)	5.580	(0.60s)	2.613	(1.08s)	0.867	(0.42s)

Figure S27D – Validation using Mirror Lake soil



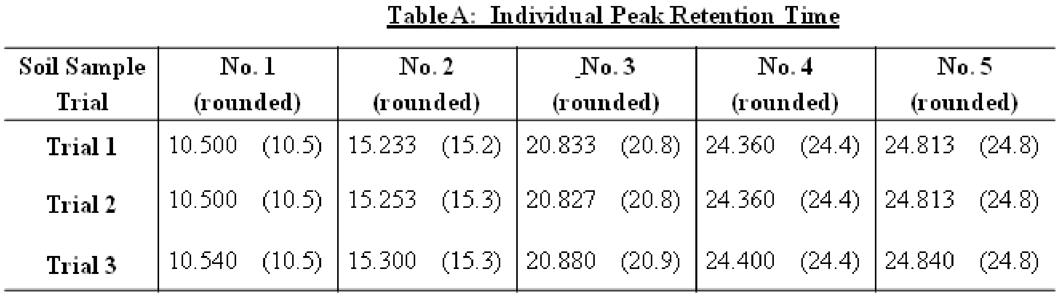


Table A: Individual Peak Retention Time

Soil Sample Trial	No. 1 (rounded)		No. 2 (rounded)		No. 3 (rounded)		No. 4 (rounded)		No. 5 (rounded)	
Trial 1	10.500	(10.5)	15.233	(15.2)	20.833	(20.8)	24.360	(24.4)	24.813	(24.8)
Trial 2	10.500	(10.5)	15.253	(15.3)	20.827	(20.8)	24.360	(24.4)	24.813	(24.8)
Trial 3	10.540	(10.5)	15.300	(15.3)	20.880	(20.9)	24.400	(24.4)	24.840	(24.8)

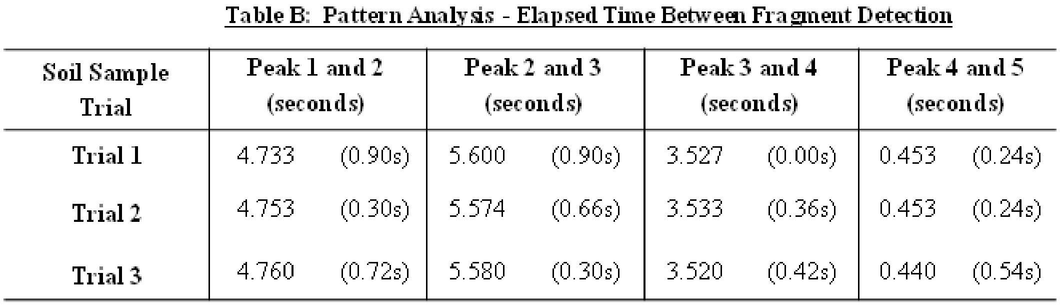
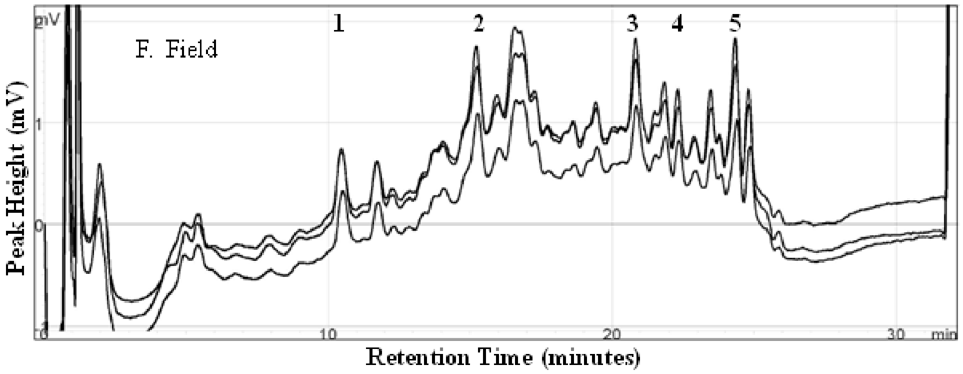


Table B: Pattern Analysis - Elapsed Time Between Fragment Detection

Soil Sample Trial	Peak 1 and 2 (seconds)		Peak 2 and 3 (seconds)		Peak 3 and 4 (seconds)		Peak 4 and 5 (seconds)	
Trial 1	4.733	(0.90s)	5.600	(0.90s)	3.527	(0.00s)	0.453	(0.24s)
Trial 2	4.753	(0.30s)	5.574	(0.66s)	3.533	(0.36s)	0.453	(0.24s)
Trial 3	4.760	(0.72s)	5.580	(0.30s)	3.520	(0.42s)	0.440	(0.54s)

Figure S27E – Validation using Cemetery soil



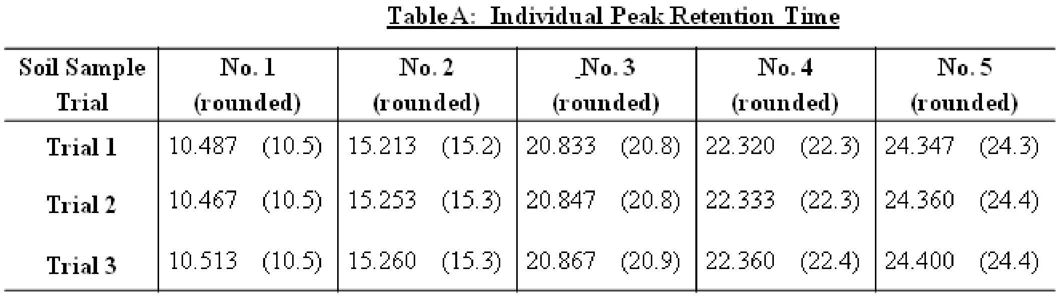


Table A: Individual Peak Retention Time

Soil Sample Trial	No. 1 (rounded)		No. 2 (rounded)		_No. 3 (rounded)		No. 4 (rounded)		No. 5 (rounded)	
Trial 1	10.487	(10.5)	15.213	(15.2)	20.833	(20.8)	22.320	(22.3)	24.347	(24.3)
Trial 2	10.467	(10.5)	15.253	(15.3)	20.847	(20.8)	22.333	(22.3)	24.360	(24.4)
Trial 3	10.513	(10.5)	15.260	(15.3)	20.867	(20.9)	22.360	(22.4)	24.400	(24.4)

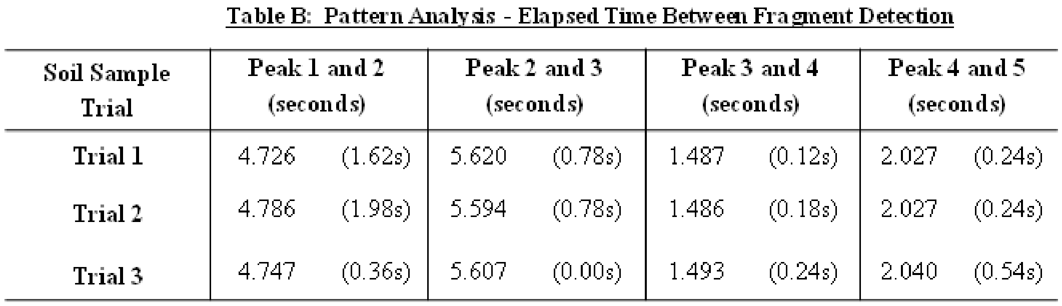
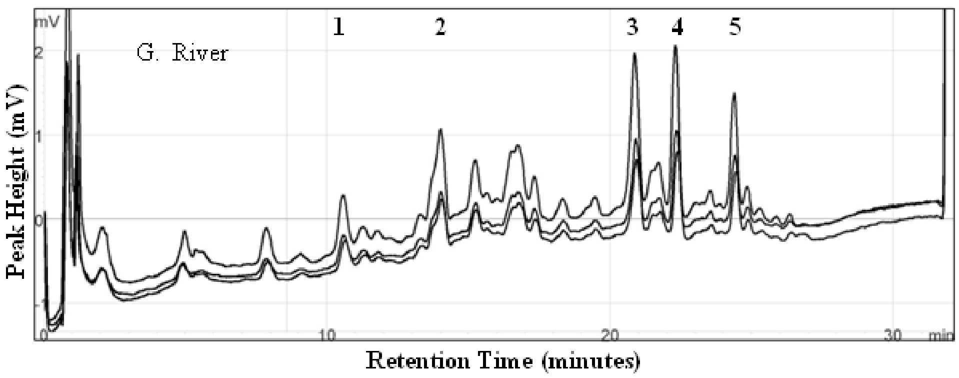


Table B: Pattern Analysis - Elapsed Time Between Fragment Detection

Soil Sample Trial	Peak 1 and 2 (seconds)		Peak 2 and 3 (seconds)		Peak 3 and 4 (seconds)		Peak 4 and 5 (seconds)	
Trial 1	4.726	(1.62s)	5.620	(0.78s)	1.487	(0.12s)	2.027	(0.24s)
Trial 2	4.786	(1.98s)	5.594	(0.78s)	1.486	(0.18s)	2.027	(0.24s)
Trial 3	4.747	(0.36s)	5.607	(0.00s)	1.493	(0.24s)	2.040	(0.54s)

Figure S27F – Validation using Field soil



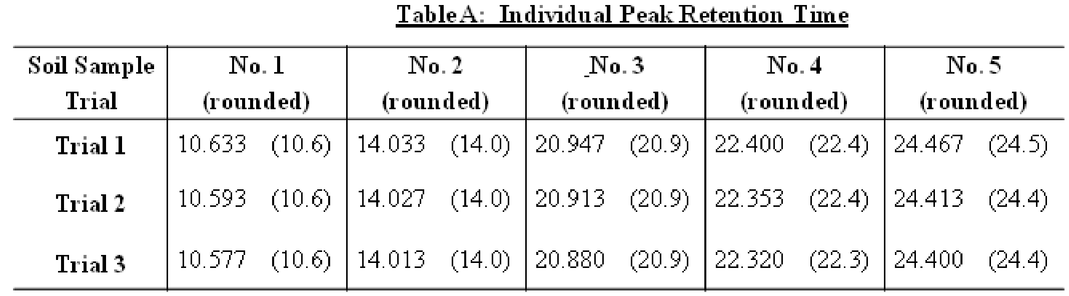


Table A: Individual Peak Retention Time

Soil Sample Trial	No. 1 (rounded)		No. 2 (rounded)		_No. 3 (rounded)		No. 4 (rounded)		No. 5 (rounded)	
Trial 1	10.633	(10.6)	14.033	(14.0)	20.947	(20.9)	22.400	(22.4)	24.467	(24.5)
Trial 2	10.593	(10.6)	14.027	(14.0)	20.913	(20.9)	22.353	(22.4)	24.413	(24.4)
Trial 3	10.577	(10.6)	14.013	(14.0)	20.880	(20.9)	22.320	(22.3)	24.400	(24.4)

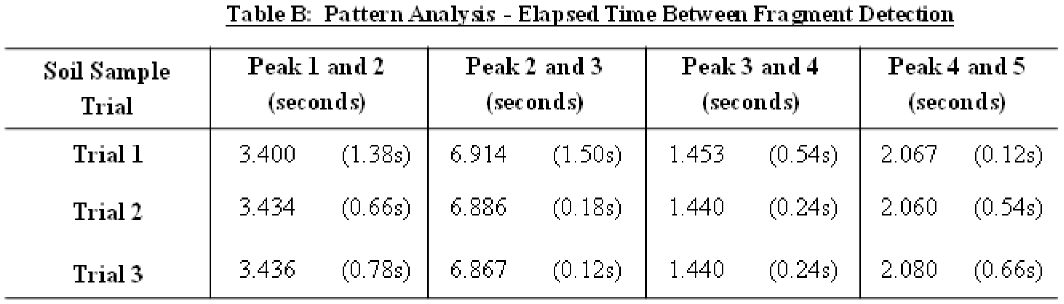


Table B: Pattern Analysis - Elapsed Time Between Fragment Detection

Soil Sample Trial	Peak 1 and 2 (seconds)		Peak 2 and 3 (seconds)		Peak 3 and 4 (seconds)		Peak 4 and 5 (seconds)	
Trial 1	3.400	(1.38s)	6.914	(1.50s)	1.453	(0.54s)	2.067	(0.12s)
Trial 2	3.434	(0.66s)	6.886	(0.18s)	1.440	(0.24s)	2.060	(0.54s)
Trial 3	3.436	(0.78s)	6.867	(0.12s)	1.440	(0.24s)	2.080	(0.66s)

Figure S27G – Validation using CT River freshwater sediment

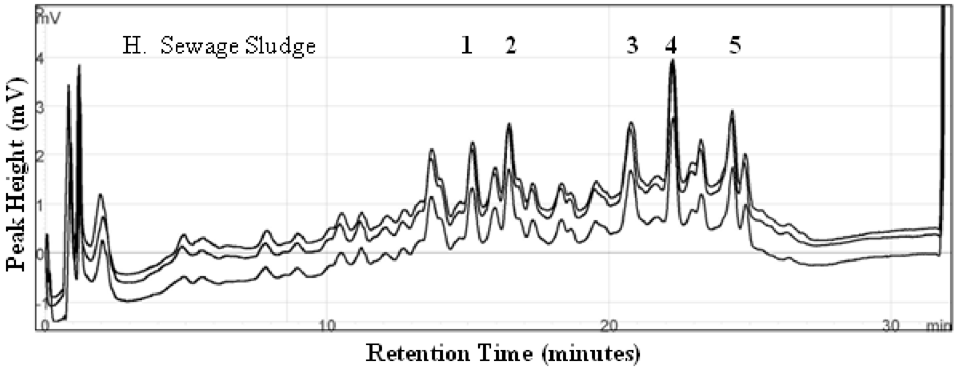


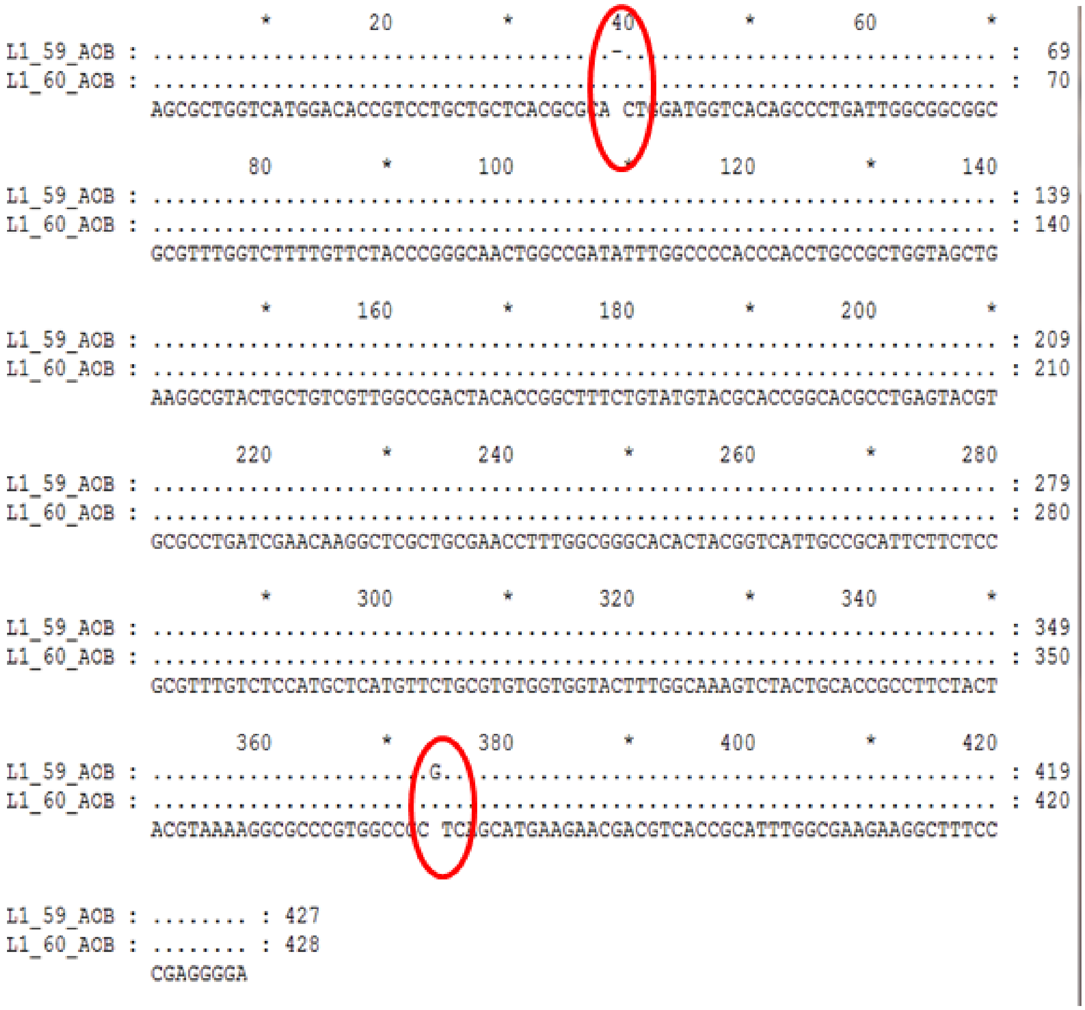
Table A: Individual Peak Retention Time

Soil Sample Trial	No. 1 (rounded)		No. 2 (rounded)		No. 3 (rounded)		No. 4 (rounded)		No. 5 (rounded)	
Trial 1	15.153	(15.2)	16.433	(16.4)	20.727	(20.7)	22.227	(22.2)	24.333	(24.3)
Trial 2	15.160	(15.2)	16.460	(16.5)	20.780	(20.8)	22.273	(22.3)	24.387	(24.4)
Trial 3	15.187	(15.2)	16.473	(16.5)	20.787	(20.8)	22.273	(22.3)	24.373	(24.4)

Table B: Pattern Analysis - Elapsed Time Between Fragment Detection

Soil Sample Trial	Peak 1 and 2 (seconds)		Peak 2 and 3 (seconds)		Peak 3 and 4 (seconds)		Peak 4 and 5 (seconds)	
Trial 1	1.280	(0.54s)	4.294	(0.90s)	1.500	(0.42s)	2.106	(0.06s)
Trial 2	1.300	(0.66s)	4.320	(0.66s)	1.493	(0.00s)	2.114	(0.42s)
Trial 3	1.286	(0.18s)	4.314	(0.30s)	1.486	(0.42s)	2.100	(0.42s)

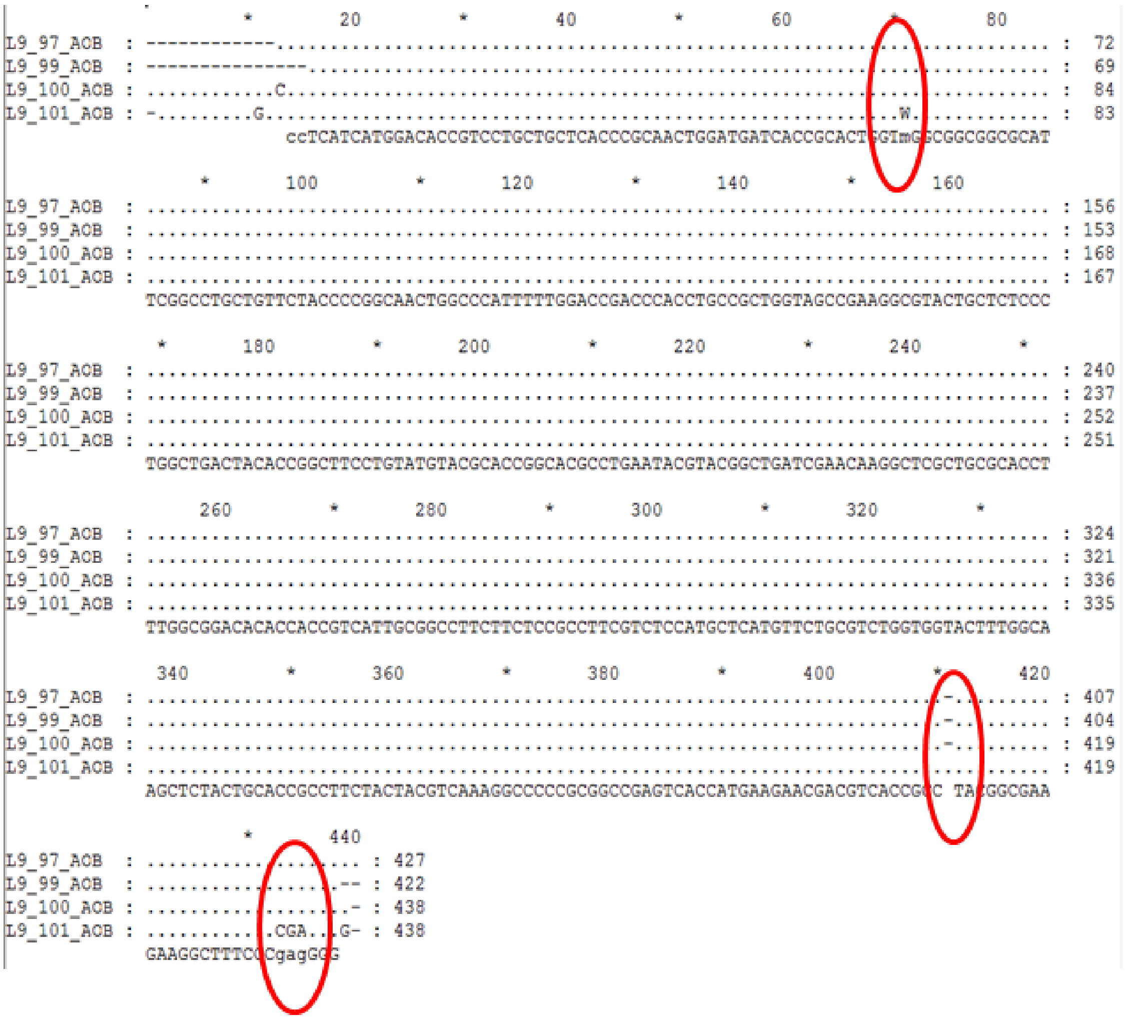
Figure S27H – Validation using Sewage Treatment sludge



```
                     *        20         *        40         *        60         *
L1_59_AOB : ......................................-............................... :  69
L1_60_AOB : ...................................................................... :  70
            AGCGCTGGTCATGGACACCGTCCTGCTGCTCACGCGCA CTGGATGGTCACAGCCCTGATTGGCGGCGGC

                    80         *       100         *       120         *       140
L1_59_AOB : ...................................................................... : 139
L1_60_AOB : ...................................................................... : 140
            GCGTTTGGTCTTTTGTTCTACCCGGGCAACTGGCCGATATTTGGCCCCACCCACCTGCCGCTGGTAGCTG

                     *       160         *       180         *       200         *
L1_59_AOB : ...................................................................... : 209
L1_60_AOB : ...................................................................... : 210
            AAGGCGTACTGCTGTCGTTGGCCGACTACACCGGCTTTCTGTATGTACGCACCGGCACGCCTGAGTACGT

                   220         *       240         *       260         *       280
L1_59_AOB : ...................................................................... : 279
L1_60_AOB : ...................................................................... : 280
            GCGCCTGATCGAACAAGGCTCGCTGCGAACCTTTGGCGGGCACACTACGGTCATTGCCGCATTCTTCTCC

                     *       300         *       320         *       340         *
L1_59_AOB : ...................................................................... : 349
L1_60_AOB : ...................................................................... : 350
            GCGTTTGTCTCCATGCTCATGTTCTGCGTGTGGTGGTACTTTGGCAAAGTCTACTGCACCGCCTTCTACT

                   360         *       380         *       400         *       420
L1_59_AOB : .......................G.............................................. : 419
L1_60_AOB : ...................................................................... : 420
            ACGTAAAAGGCGCCCGTGGCCCC TCAGCATGAAGAACGACGTCACCGCATTTGGCGAAGAAGGCTTTCC

L1_59_AOB : ........ : 427
L1_60_AOB : ........ : 428
            CGAGGGGA
```

Supplementary Figure S28 – Sequence alignment *AOB* group [*amoA* gene]: Lawn 1

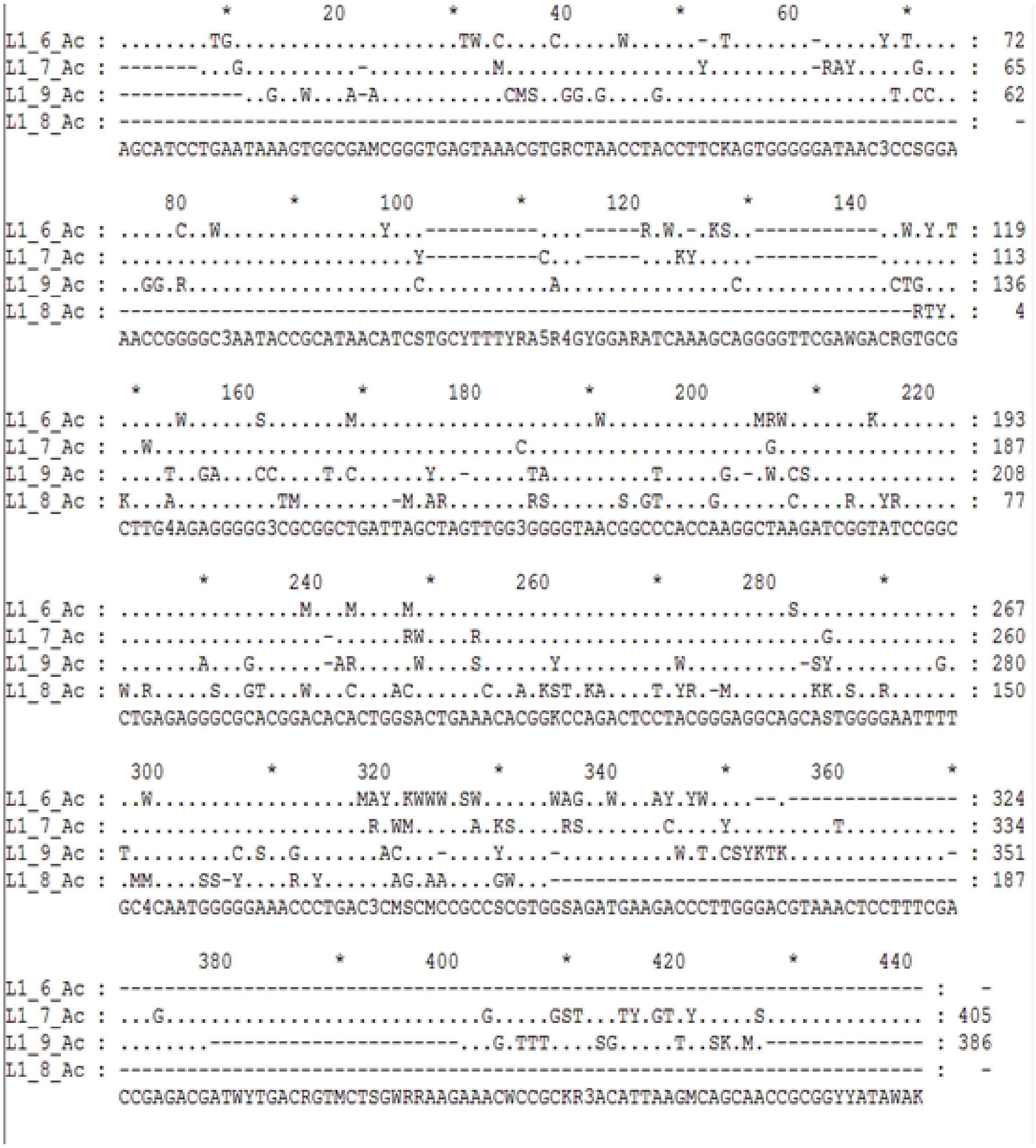
Figure shows a sequence alignment using Clustal. Ovals highlight positions that did not match between the two sequences.



```
                      *         20         *         40         *         60         *         80
L9_97_ACB  : ------------........................................................................ :   72
L9_99_ACB  : ---------------..................................................................... :   69
L9_100_ACB : ............C....................................................................... :   84
L9_101_ACB : -.........G...........................................................W............. :   83
                          ccTCATCATGGACACCGTCCTGCTGCTCACCCGCAACTGGATGATCACCGCACTGGTmGGCGGCGGCGCAT

                  *       100         *       120         *       140         *       160
L9_97_ACB  : .................................................................................... :  156
L9_99_ACB  : .................................................................................... :  153
L9_100_ACB : .................................................................................... :  168
L9_101_ACB : .................................................................................... :  167
             TCGGCCTGCTGTTCTACCCCGGCAACTGGCCCATTTTTGGACCGACCCACCTGCCGCTGGTAGCCGAAGGCGTACTGCTCTCCC

              *       180         *       200         *       220         *       240         *
L9_97_ACB  : .................................................................................... :  240
L9_99_ACB  : .................................................................................... :  237
L9_100_ACB : .................................................................................... :  252
L9_101_ACB : .................................................................................... :  251
             TGGCTGACTACACCGGCTTCCTGTATGTACGCACCGGCACGCCTGAATACGTACGGCTGATCGAACAAGGCTCGCTGCGCACCT

                  260         *       280         *       300         *       320         *
L9_97_ACB  : .................................................................................... :  324
L9_99_ACB  : .................................................................................... :  321
L9_100_ACB : .................................................................................... :  336
L9_101_ACB : .................................................................................... :  335
             TTGGCGGACACACCACCGTCATTGCGGCCTTCTTCTCCGCCTTCGTCTCCATGCTCATGTTCTGCGTCTGGTGGTACTTTGGCA

              340         *       360         *       380         *       400         *       420
L9_97_ACB  : ..........................................................................-......... :  407
L9_99_ACB  : ..........................................................................-......... :  404
L9_100_ACB : ..........................................................................-......... :  419
L9_101_ACB : .................................................................................... :  419
             AGCTCTACTGCACCGCCTTCTACTACGTCAAAGGCCCCCGCGGCCGAGTCACCATGAAGAACGACGTCACCGGC TACGGCGAA

                      *       440
L9_97_ACB  : .................... : 427
L9_99_ACB  : ..................-- : 422
L9_100_ACB : ...................- : 438
L9_101_ACB : ............CGA...G- : 438
             GAAGGCTTTCCCgagGGG
```

Supplementary Figure S29 – Sequence alignment *AOB* group [*amoA* gene]: Lawn 9

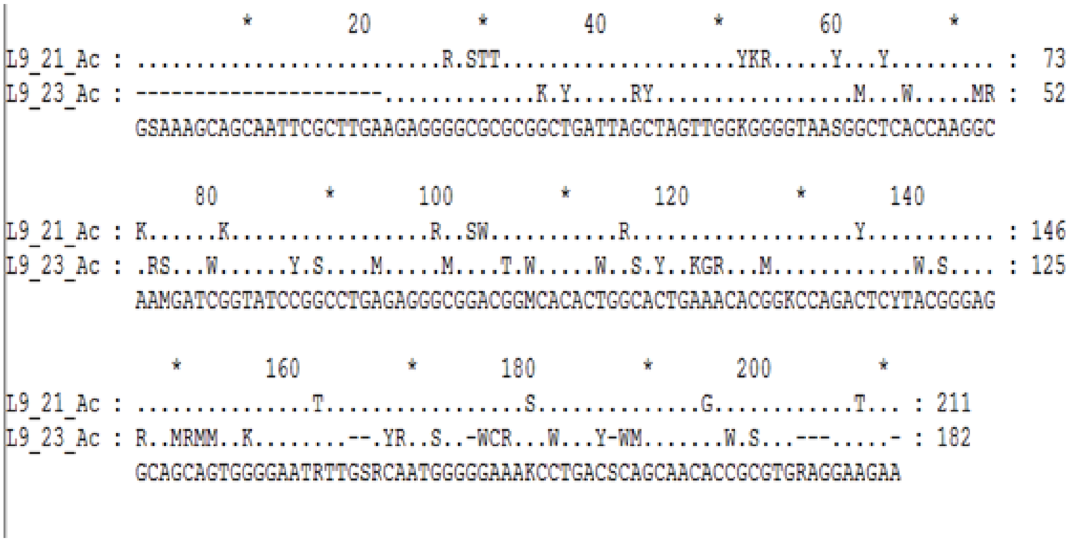
Figure shows a sequence alignment using Clustal. Ovals highlight positions that did not match between the two sequences.



```
                   *        20         *        40         *        60         *
L1_6_Ac : ........TG....................TW.C....C.....W......-.T.......-.....Y.T.... :  72
L1_7_Ac : -------...G..........-...........M.................Y.........-RAY.....G... :  65
L1_9_Ac : -----------..G..W...A-A...........CMS..GG.G....G....................T.CC.. :  62
L1_8_Ac : -------------------------------------------------------------------------- :   -
          AGCATCCTGAATAAAGTGGCGAMCGGGTGAGTAAACGTGRCTAACCTACCTTCKAGTGGGGGATAAC3CCSGGA

              80         *       100         *       120         *       140
L1_6_Ac : .....C..W..............Y...----------....-----R.W.-.KS..-----------..W.Y.T : 119
L1_7_Ac : ..........................Y----------C...-----...KY.....-----------....... : 113
L1_9_Ac : ..GG.R....................C...........A...............C.............CTG... : 136
L1_8_Ac : ----------------------------------------------------------------------RTY. :   4
          AACCGGGGC3AATACCGCATAACATCSTGCYTTTYRA5R4GYGGARATCAAAGCAGGGGTTCGAWGACRGTGCG

           *       160         *       180         *       200         *       220
L1_6_Ac : .....W......S.......M.....................W.............MRW.......K....... : 193
L1_7_Ac : ..W................................C.....................G................ : 187
L1_9_Ac : ....T..GA...CC....T.C......Y..-.....TA.........T.....G.-.W.CS............. : 208
L1_8_Ac : K...A..........TM.......-M.AR.......RS......S.GT....G......C....R..YR..... :  77
          CTTG4AGAGGGGG3CGCGGCTGATTAGCTAGTTGG3GGGGTAACGGCCCACCAAGGCTAAGATCGGTATCCGGC

                 *       240         *       260         *       280         *
L1_6_Ac : ................M...M...M..................................S.............. : 267
L1_7_Ac : ..................-......RW....R..............................G........... : 260
L1_9_Ac : .......A...G......-AR.....W....S......Y..........W..........-SY.........G. : 280
L1_8_Ac : W.R.....S..GT...W...C...AC......C..A.KST.KA....T.YR.-M.......KK.S..R...... : 150
          CTGAGAGGGCGCACGGACACACTGGSACTGAAACACGGKCCAGACTCCTACGGGAGGCAGCASTGGGGAATTTT

           300         *       320         *       340         *       360         *
L1_6_Ac : ..W..................MAY.KWWW.SW......WAG..W...AY.YW....--.--------------- : 324
L1_7_Ac : .......................R.WM....A.KS....RS.......C....Y.........T.......... : 334
L1_9_Ac : T.........C.S..G.......AC...-....Y....-..........W.T.CSYKTK..............- : 351
L1_8_Ac : .MM....SS-Y....R.Y......AG.AA....GW...------------------------------------ : 187
          GC4CAATGGGGGAAACCCTGAC3CMSCMCCGCCSCGTGGSAGATGAAGACCCTTGGGACGTAAACTCCTTTCGA

                 380         *       400         *       420         *       440
L1_6_Ac : ----------------------------------------------------------------------- :   -
L1_7_Ac : ...G.............................G....GST...TY.GT.Y.....S.............. : 405
L1_9_Ac : ........----------------------...G.TTT....SG.....T..SK.M.-------------- : 386
L1_8_Ac : ----------------------------------------------------------------------- :   -
          CCGAGACGATWYTGACRGTMCTSGWRRAAGAAACWCCGCKR3ACATTAAGMCAGCAACCGCGGYYATAWAK
```

Supplementary Figure S30 – Sequence alignment *Acidobaceria* [16S rRNA gene]: Lawn 1

Figure shows a sequence alignment using Clustal.



```
                    *        20         *        40         *        60         *
L9_21_Ac : ..........................R.STT....................YKR.....Y...Y......... :  73
L9_23_Ac : ---------------------.............K.Y.....RY.................M...W.....MR :  52
           GSAAAGCAGCAATTCGCTTGAAGAGGGGCGCGCGGCTGATTAGCTAGTTGGKGGGGTAASGGCTCACCAAGGC

                80         *       100         *       120         *       140
L9_21_Ac : K......K.................R..SW...........R...................Y........... : 146
L9_23_Ac : .RS...W......Y.S....M.....M....T.W.....W..S.Y..KGR...M............W.S.... : 125
           AAMGATCGGTATCCGGCCTGAGAGGGCGGACGGMCACACTGGCACTGAAACACGGKCCAGACTCYTACGGGAG

              *       160         *       180         *       200         *
L9_21_Ac : ...............T.................S..............G............T... : 211
L9_23_Ac : R..MRMM..K........--.YR..S..-WCR...W...Y-WM.......W.S...---.....- : 182
           GCAGCAGTGGGGAATRTTGSRCAATGGGGGAAAKCCTGACSCAGCAACACCGCGTGRAGGAAGAA
```

Supplementary Figure S31 – Sequence alignment *Acidobaceria* [16S rRNA gene]: Lawn 9

Figure shows a sequence alignment using Clustal.