



**The author(s) shown below used Federal funding provided by the U.S. Department of Justice to prepare the following resource:**

**Document Title:** Developing Reliable Methods for Microbial Fingerprinting of Soils

**Author(s):** David Foran, Ph.D., Ellen Jesmok, M.S., James Hopkins, M.S.

**Document Number:** 250659

**Date Received:** March 2017

**Award Number:** 2013-R2-CX-K010

**This resource has not been published by the U.S. Department of Justice. This resource is being made publically available through the Office of Justice Programs' National Criminal Justice Reference Service.**

**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.**

# **Developing Reliable Methods for Microbial Fingerprinting of Soils**

**2013-R2-CX-K010**

**David Foran PhD, Ellen Jesmok MS, and James Hopkins MS**

## ABSTRACT

Soil evidence has the potential to be a valuable forensic tool linking a suspect, victim, or item to a crime scene, however, there is currently no reliable and objective method for individualizing soil, as only class characteristics are considered in traditional analysis. In this research, the utility of soil bacterial profiling via next-generation sequencing of the 16S rRNA gene was examined, for the purpose of identifying a soils' origin. Soil was collected from ten different habitat types to establish the general feasibility of differentiating soils based on bacterial profiles. Next, the much more challenging task of differentiating similar habitats was examined by comparing soils from nine woodlots in very close proximity. Factors that can affect bacterial profiles within a site were also considered, by collecting soils over time and space in three habitats. Finally, mock evidentiary items, including cotton t-shirts, a shovel, shoes, socks, and a tire, were exposed to soil to examine its traceability back to the site of origin, both immediately and over time. Soil bacterial profiles were generated using an Illumina MiSeq, which produced approximately 150,000 sequences per soil sample. Initially, five methods for analyzing the sequence data were examined as bacterial profile comparison tools (bacterial abundance charts, pairwise comparisons, nonmetric multidimensional scaling, hierarchical cluster analysis, and the supervised classification technique *k*-Nearest Neighbor). Based on preliminary results, pairwise comparisons and hierarchical cluster analysis were eliminated because they often produced ambiguous results. Abundance charts and nonmetric multidimensional scaling provided simplification and visualization of the massive amounts of data, a clear benefit for explaining complicated scientific results to a jury. *k*-Nearest Neighbor offered an objective, statistics-based assignment of soil to a location, helping to meet the standards suggested in the National Research Council's 2009 report on forensic science. Diverse

and similar habitats were successfully differentiated in both multidimensional space and through supervised classification, which accurately classified soil samples back to their locations of origin 100% and 87.5% of the time respectively. Time and space within a habitat did not affect bacterial profiles enough to hinder location of origin assignment, where samples were correctly classified an average of 96% of the time. Soil collected from evidentiary items exhibited abundance change of certain taxonomic classes, but remained clustered nearest its location of origin, 100% accurately classifying even after a full year of storage. The considerable success in tracing soils back to a location of origin demonstrates the potential of next-generation sequencing of bacteria, in conjunction with a combination of robust statistical techniques, for the individualization of forensic soil samples.

## TABLE OF CONTENTS

<b>ABSTRACT</b> .....	<b>2</b>
<b>EXECUTIVE SUMMARY</b> .....	<b>6</b>
<b>INTRODUCTION</b> .....	<b>26</b>
Forensic Soil Investigation .....	26
Classic Soil Analyses .....	27
Molecular Analyses of Soil Bacteria .....	28
Analysis of Soil Bacteria at the Michigan State University Forensic Biology Laboratory .....	28
Feasibility of Next-Generation Sequencing for Forensic Soil Analysis.....	45
<b>MATERIALS AND METHODS</b> .....	<b>48</b>
Soil Sampling Schemes.....	48
Soil Collection from Diverse Habitats .....	50
Soil Collection from Similar Habitats.....	50
Soil Collection from Three Habitats over Time .....	51
Soil Collection from Three Habitats over Horizontal Space .....	51
Soil Collection from Three Habitats over Vertical Space .....	52
Soil Collection from Evidentiary Items .....	52
DNA Extraction .....	53
PCR Amplification of 16S rRNA Variable Regions 3 – 4 .....	53
DNA Quantification and Equimolar Pooling.....	54
Sequencing Purified PCR Products .....	54
Gene Sequence Data Pretreatment.....	55
Next-Generation Sequencing Analysis Procedures .....	55
<b>RESULTS</b> .....	<b>57</b>
Illumina MiSeq Sequencing Efficiency .....	57
General Results .....	57
Analysis of Soils from Diverse Habitats.....	61
Bacterial Abundance Charts .....	61
Nonmetric Multidimensional Scaling .....	62
k-Nearest Neighbor.....	63
Analysis of Soils from Similar Habitats .....	64

Bacterial Abundance Charts .....	64
Nonmetric Multidimensional Scaling .....	65
k-Nearest Neighbor .....	66
Analysis of Soils from Temporal Study.....	67
Bacterial Abundance Charts .....	67
Nonmetric Multidimensional Scaling .....	68
k-Nearest Neighbor .....	68
Analysis of Soils from Spatial Studies.....	69
Bacterial Abundance Charts .....	69
Nonmetric Multidimensional Scaling .....	71
k-Nearest Neighbor .....	75
Analysis of Preliminary Evidentiary Samples .....	75
Bacterial Abundance Charts .....	75
Nonmetric Multidimensional Scaling .....	76
k-Nearest Neighbor .....	77
Analysis of Secondary Evidentiary Samples .....	78
Bacterial Abundance Charts .....	78
Nonmetric Multidimensional Scaling .....	81
k-Nearest Neighbor .....	81
<b>DISCUSSION .....</b>	<b>85</b>
<b>CONCLUSIONS .....</b>	<b>95</b>
<b>IMPLICATIONS FOR FURTHER RESEARCH .....</b>	<b>97</b>
<b>DISSEMINATION OF RESEARCH FINDINGS .....</b>	<b>98</b>
<b>ACKNOWLEDGEMENTS .....</b>	<b>99</b>
<b>REFERENCES.....</b>	<b>100</b>

## EXECUTIVE SUMMARY

In forensic investigations, soil can prove an invaluable evidentiary source for linking a suspect, victim, or piece of evidence to a crime scene. Potentially found on shoes, tires, shovels, etc., the virtually unlimited types of soil and their geospatial distribution can make such evidence highly probative (Saferstein, 2002). Traditional forensic soil comparisons utilize the physical and chemical characteristics of soil, often requiring large quantities for testing, which may be unavailable in a forensic setting. Additionally, these methods measure class characteristics of soils, leading to only a general association between known and unknown samples, unless the soil contains a rare compound or element. Therefore, the subjectivity in interpretation, and lack of statistical measures (Pye, 2007), limit the value of soil evidence. The recent National Academy of Sciences (NAS) report (National Research Council, 2009) called many of the practices used in forensics into question, soil examination included. This requires a reassessment of the analysis techniques currently used and how they can be improved upon. Furthermore, the *Daubert* ruling accentuated the need for forensic science to develop standardized, peer reviewed methodologies, with recognition of error rates (*Daubert v. Merrell Dow Pharmaceuticals*). Such requirements have led forensic practitioners to utilize better data generation techniques, as well as examine the use of objective statistical measures for analysis. Based on the weakness of current forensic soil analysis techniques and the scrutiny of forensic science practices, clearly there is a need for methods that capture the distinctive characteristics of soil, which will lead to better characterization and identification of this complex form of evidence.

Assessment of soil bacterial populations holds the potential for linking evidentiary and known soil samples, as the breadth of microbial diversity in soil, even if considering only the prokaryotic contribution, is staggering. Several advances have allowed microbiologists to

directly assay the bacterial metagenome, yet, only a few of these techniques have gained a footing in the forensic sciences. In general, the bacterial 16S rRNA locus is assayed, as it contains highly conserved regions that exist in all bacteria to which universal PCR primers can anneal, interspersed with highly variable regions. Methods used to assay the 16S gene include denaturing gradient gel electrophoresis (Muyzer and Smalla, 1998), amplicon length heterogeneity-polymerase chain reaction (Moreno et al., 2006), and, most popularly, terminal restriction fragment length polymorphism (Liu et al., 1997). Although all of these techniques can provide information about a bacterial community, the massive amount of bacteria in soil makes their resolution power low and the interpretation of the data potentially subjective. Recently however, more powerful technologies have come into use that allow for an even better understanding of soil bacterial metagenomics. A promising technique developed in the last ten years is massively parallel sequencing, or next-generation sequencing, which has the ability to generate vast amounts of data in short periods of time, facilitating metagenomic analysis of complex substrates like soil. Many next-generation sequencing platforms exist, each with its own chemistries and detection systems; two of which were used in the current research: Roche 454 pyrosequencing for preliminary studies and Illumina MiSeq sequencing by synthesis for the studies sponsored by this NIJ award.

Hopkins (2014), utilizing 454 technology, tested replicate soil samples from three habitat types near the Forensic Biology Laboratory at Michigan State University. He studied various statistical techniques to analyze the sequence data, each already used by the microbiology community, to ascertain which, if any, had forensic utility. The simplest way to examine the massive amounts of data produced via next-generation sequencing was to visualize the bacterial communities through abundance charts. These charts are generated from taxonomic data,

categorizing and quantifying the bacteria that make up a total profile, allowing visual comparison among several soil samples. Unfortunately, the subjectivity of abundance chart interpretation severely limits their utility as a stand-alone technique in forensic science, and more objective techniques that can produce a measure of similarity/dissimilarity are necessary.

Pairwise comparisons are objective statistical techniques that can be used to determine whether two bacterial profiles differ significantly. Two pairwise comparisons were utilized to examine several replicate soil samples collected from the three habitats, and on some occasions even within-habitat replicates differed significantly. Given this, an even more stringent inspection of pairwise comparisons' utility was performed, wherein a single soil sample was divided into four, DNA was extracted from each, and the 16S genes were sequenced. The four resulting bacterial profiles were compared using both pairwise methods, and were found to differ significantly 5/6 and 4/6 times. This result illustrates the hypersensitivity of pairwise comparisons for soil bacteria analysis, where slight bacterial fluctuations within the same trowel scoop of soil proved to statistically differ, negating its use for forensic applications. Given this, further analysis techniques, which are not overwhelmed by small fluctuations in the bacterial community, were required.

The next method examined was nonmetric multidimensional scaling (NMDS), which is an ordination technique designed to plot data in multidimensional space. Each data point represents a single soil bacterial profile, and the spread of the data approximates the (dis)similarities between samples (Cox, 2001; Borg and Groenen, 2005). The final configuration of data points illustrates the relative similarity among soils (i.e., clustered data points are more similar than distant ones). NMDS plots provide helpful data interpretation because, similar to abundance charts, they offer a visual representation of multiple bacterial profiles. Additionally,

NMDS plots generate information on the relative similarities of several samples through grouping and separation. The weakness of NMDS is that it still lacks the objectivity that is desired for forensic analysis, as cluster identification and general association of samples is open to interpretation.

Another technique tested for assessing relative similarity among bacterial profiles was hierarchical cluster analysis (HCA). HCA is an unsupervised branching method that allows visualization of dissimilarities among samples in a dendrogram (Beebe et al., 1998) where relationships between samples can be inferred. However, incongruous results from dendrograms were obtained when different linkage methods were utilized, therefore the forensic utility of HCA was found to be weak. HCA also provides little information that cannot be gleaned from NMDS plots; thus the latter was preferential when examining relative similarity among multiple soil samples.

A final set of techniques tested for analysis of next-generation sequencing data was supervised classification, which offers an objective analysis of sequence data. These techniques assign bacterial profiles to a location of origin by building models based on groups of known samples, collectively called training sets (Mohri et al., 2012). Unknowns (test sets) are then compared to the model and assigned to the closest group of known samples or, depending on the technique, to no group at all. In this research, the supervised classification technique *k*-Nearest Neighbor (*k*-NN) was used to objectively assess soil samples' locations of origin.

Based on the preliminary studies, a much more detailed set of experiments was undertaken, the results of which are described in this report. The most basic requirement that must be met when assessing the forensic utility of bacterial profiling of soils built through next-generation sequencing is differentiation of diverse soils. Towards this end, soil samples from ten

assorted habitats were collected and tested (detailed below). The second, more challenging test was to determine the differentiability of very similar habitats. Forensic soil samples are often contested as being from one location or another, and these locations may be the same habitat type; therefore, it is requisite that forensic scientists can differentiate soils exposed to very similar environmental factors. In this research, soil samples were collected from nine deciduous woodlots to examine bacterial profile similarity and assess whether highly similar habitats could be differentiated. Scientists must also be aware of factors that may impact the comparison of bacterial profiles from within a site, such as time and space. In both diverse and similar habitat studies soils were collected over time to determine how it influenced bacterial profiling. Additionally, the influence of small spatial scales over the surface of a habitat and in soil depth were tested. Finally, once the general feasibility of soil bacterial profiling was assessed, studies mimicking more realistic forensic scenarios, wherein soil is recovered from evidence, were performed. Soil from a woodlot was placed on several mock evidence items, and aged in both room temperature and 4°C storage. Bacterial profiles from the evidence taken over time were compared to nine woodlots to determine if the evidence samples traced back to their actual origin.

Over the course of these studies, the following soil samples were collected: 5 replicate samples from each of ten diverse habitats, 5 replicate samples from each of nine similar habitats (woodlots), 23 samples from each of three habitats over a full year, 17 samples across the surface of three habitats, 7 samples at various depths within three habitats, and evidentiary samples (88 from t-shirts and 25 from a tire, shoe, sock, shirt, and shovel) exposed to soil in one habitat. In total, 349 soil samples were processed. Following DNA extraction, the 16S rRNA gene variable regions 3 and 4 were amplified with barcoded primers, allowing for pooling of multiple samples.

Sequencing was performed using an Illumina MiSeq, producing approximately 150,000 sequence reads per soil sample. Sequence data was processed in mothur, an open-source software program for analyzing 16S sequences (Schloss, 2009). Sequence libraries were aligned to known bacterial reference files and abundance charts were generated at the taxonomic class level. Sørensen-Dice coefficients were calculated for further analysis through NMDS and *k*-NN.

Bacterial abundance charts of diverse habitats (FIG. 1) were effective in identifying very dissimilar samples (e.g., the dirt road), which differed in specific bacterial class abundance. In multidimensional space (FIG. 2), soil samples from each habitat formed clusters, with some intermingling between habitats. These clusters were resolved when habitats were ordinated as pairs or triads (exemplified in FIG. 3). *k*-NN exhibited an 88% assignment accuracy when all habitats were analyzed together. Misclassifications occurred between the marsh and fallow agricultural field soil samples and between the deciduous woods and yard samples. These were all correctly assigned to their habitat of origin when analyzed as pairs in a *k*-NN model.

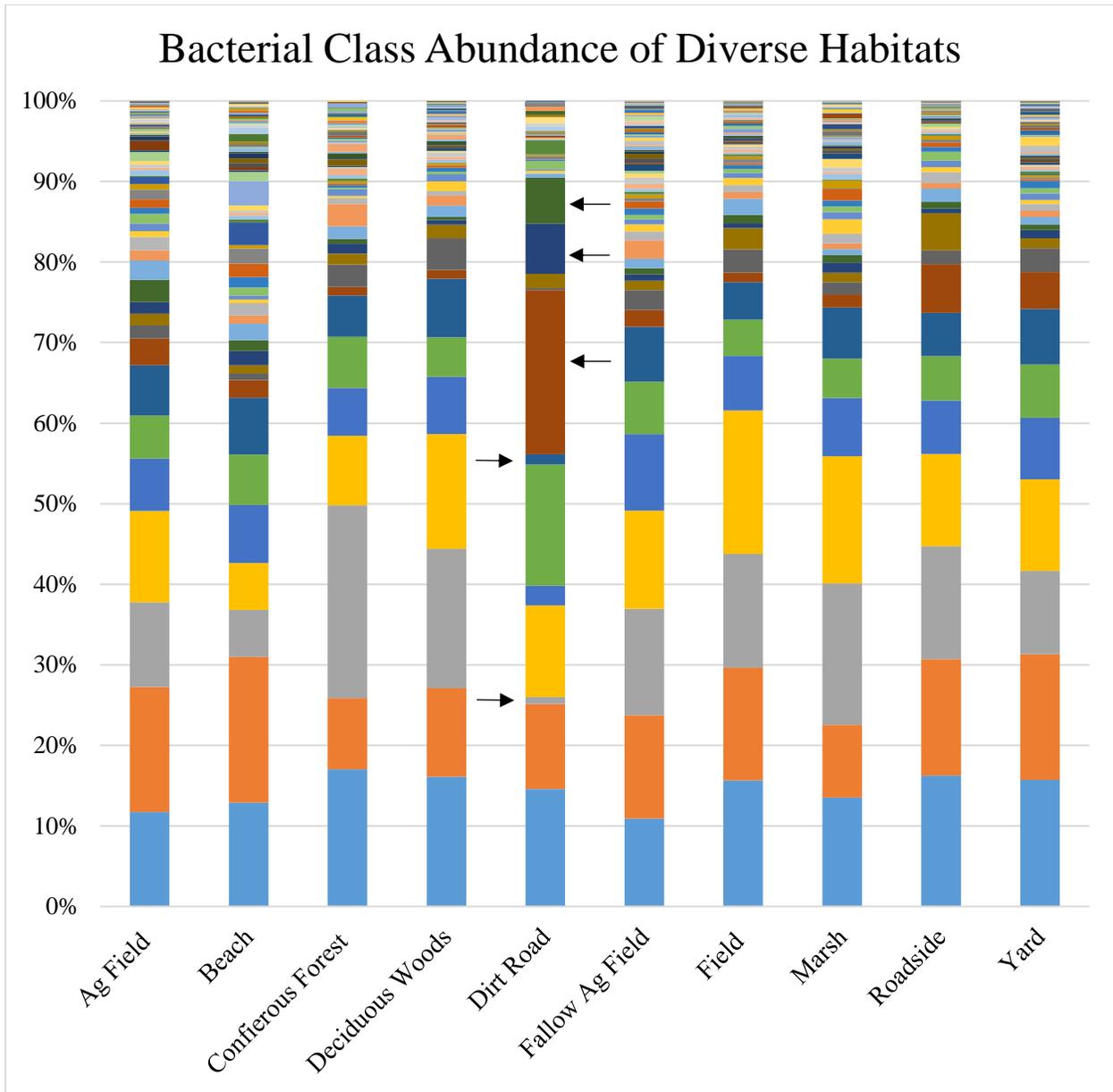


FIG. 1—Average (n=5) bacterial class abundance of ten diverse habitats. The dirt road clearly differed from the other habitats, containing higher levels of *Flavobacteria*, *Clostridia*, and *Bacilli* (designated by arrows in ascending order on the right), along with lower levels of *Acidobacteria* and *Betaproteobacteria* (designated by arrows in ascending order on the left). Ag=Agricultural.

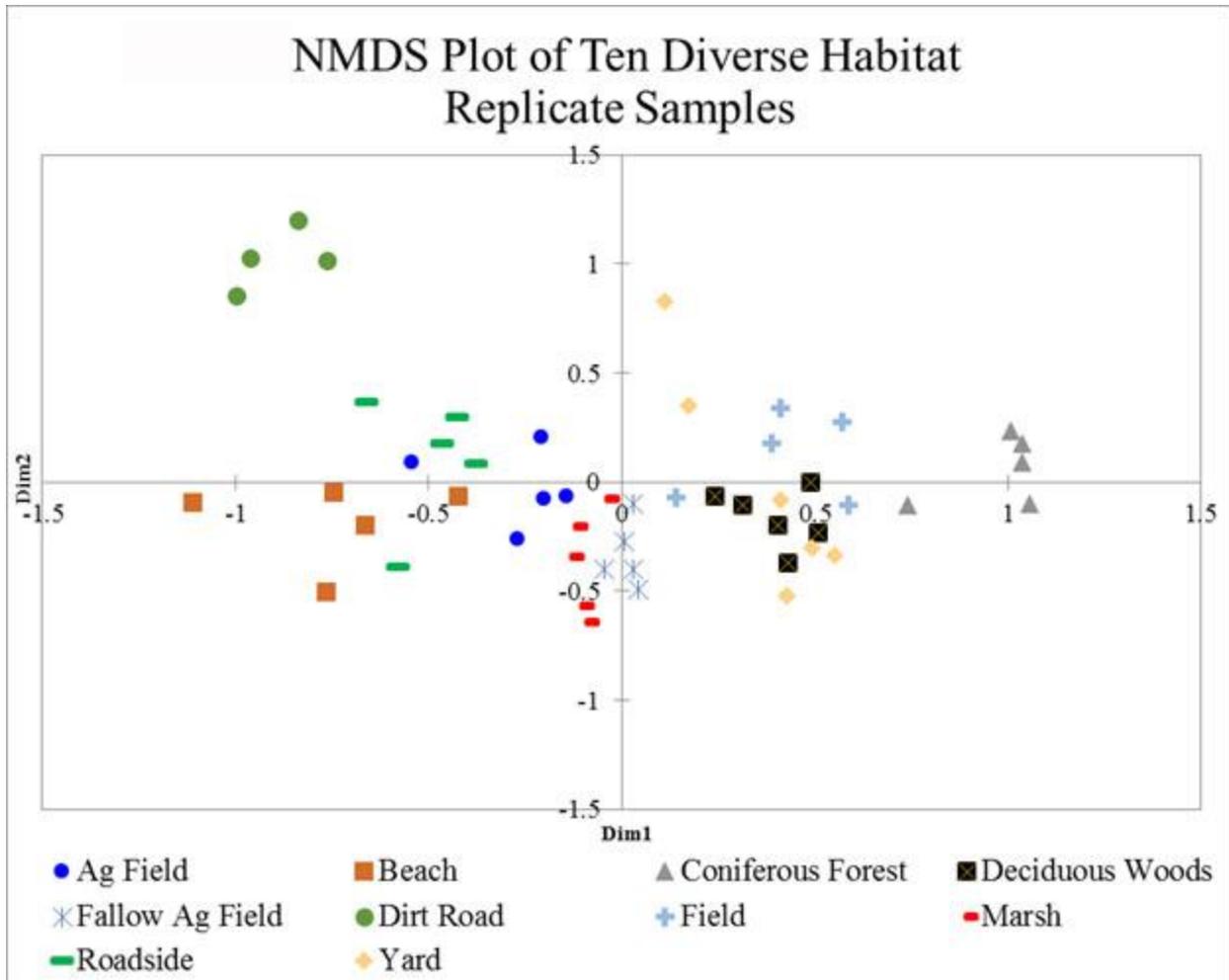


FIG. 2—NMDS plot ordinating samples collected from the ten diverse habitats. Replicate samples from the same habitat formed clusters, but intermingling occurred among some of the habitats. Ag=Agricultural

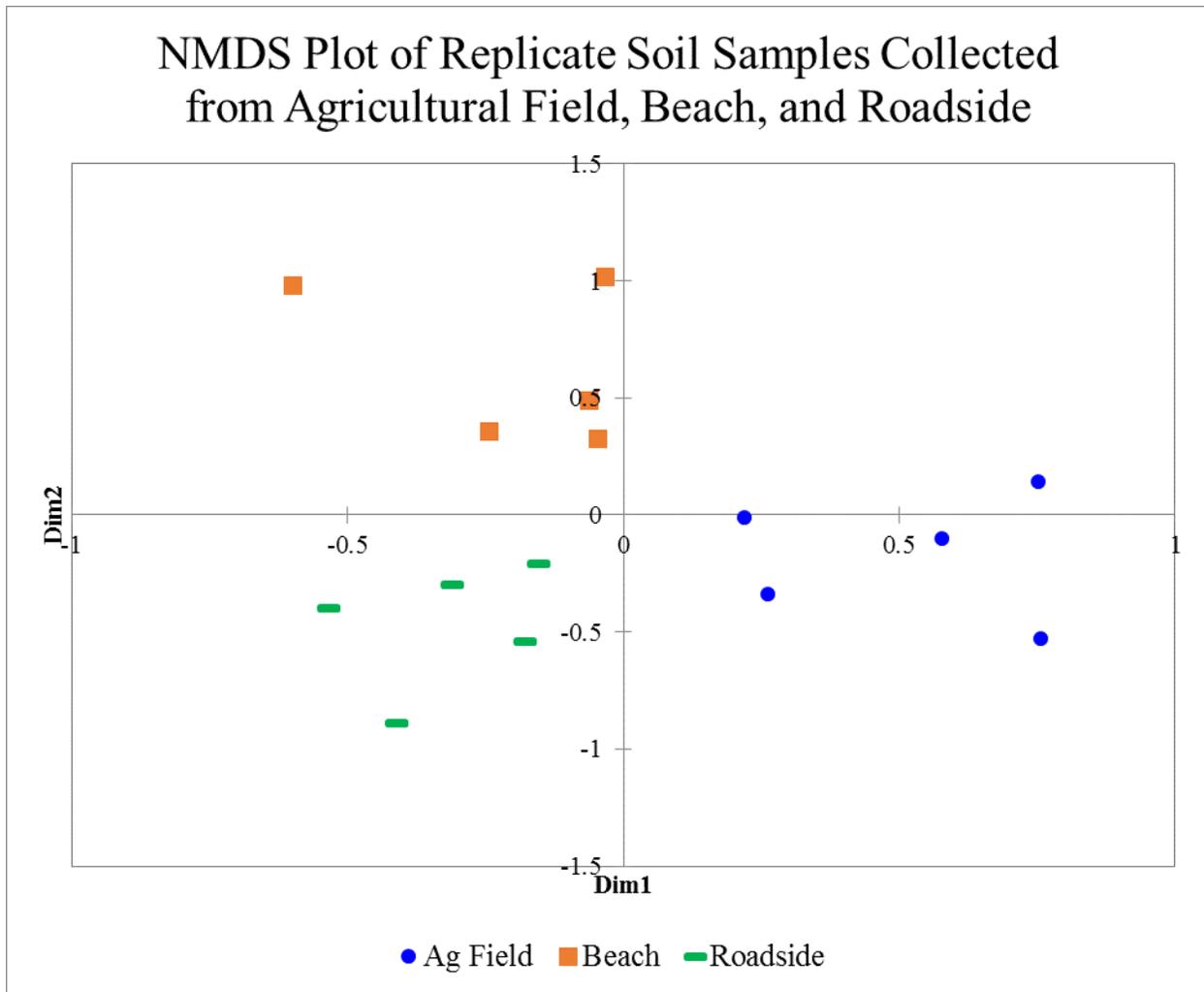


FIG. 3—NMDS plot ordinating soil samples collected from the agricultural (Ag) field, beach, and roadside. Samples from these locations intermingled when all habitats were ordinated together, but were resolved when analyzed as pairs or triads in NMDS plots.

Bacterial abundance charts of similar habitats appeared more similar than the diverse habitats, though diversity differed among the woodlots. Soil samples collected from the same woodlot clustered together in NMDS plots, but intermingling occurred among several of the clusters. Again, when pairs or triads of woodlots were ordinated, separation of the clusters occurred. *k*-NN was accurate in its assignment of the woodlot location samples 87.5% of the time. All samples from woodlot 8 and one sample from woodlot 9 were incorrectly assigned, and when removed from the model, 100% assignment accuracy was achieved.

Bacterial abundance charts of soil samples collected within a habitat over time and across the surface were very similar. Depth samples within a habitat showed variation, where surface samples had clear bacterial class abundance differences compared to the deeper soils. Samples from within each habitat clustered together over time and space with the exception of the depth samples collected from the deciduous woods and yard, which intermingled. When ordinated without the treated yard, the depth sample clusters from these habitats separated. *k*-NN analysis resulted in the assignment of soil samples back to their location of origin 92.3% of the time over the full year, 97.2% of the time for horizontal surface samples, and 100% of the time for depth samples when soils from different depths were used to build the training set. Soil samples collected from the deciduous woods and yard in February and the yard in March were the only temporal samples misassigned, both classifying to the treated yard. Samples more than 90 feet from the nearest training set sample were the only horizontal surface samples misassigned.

Bacterial profiles generated from soil on evidentiary items exhibited abundance changes over time (e.g. FIG. 4 and 5), beginning more rapidly on evidence stored at room temperature. Of note were increases in *Actinobacteria* and *Bacilli* and decreases in *Acidobacteria*, *Sphingobacteria*, *Betaproteobacteria*, and *Spartobacteria* across all evidence types. The most substantial changes occurred within the first 6 months of storage, while less change was evident between samples collected after six months and one year. Soil samples collected off t-shirts over a four month period, and evidentiary items at both six months and one year after exposure, clustered together in multidimensional space (FIG. 6 and 7), away from all woodlots, but in closest proximity to the woodlot of origin. Importantly, while drift away from the woodlot of origin occurred for all soil samples, bacterial profiles never became more similar to other woodlots. *k*-NN accurately assigned soil collected off all evidentiary items back to their location

of origin 100% of the time throughout the full year, regardless of the evidence material or storage temperature.

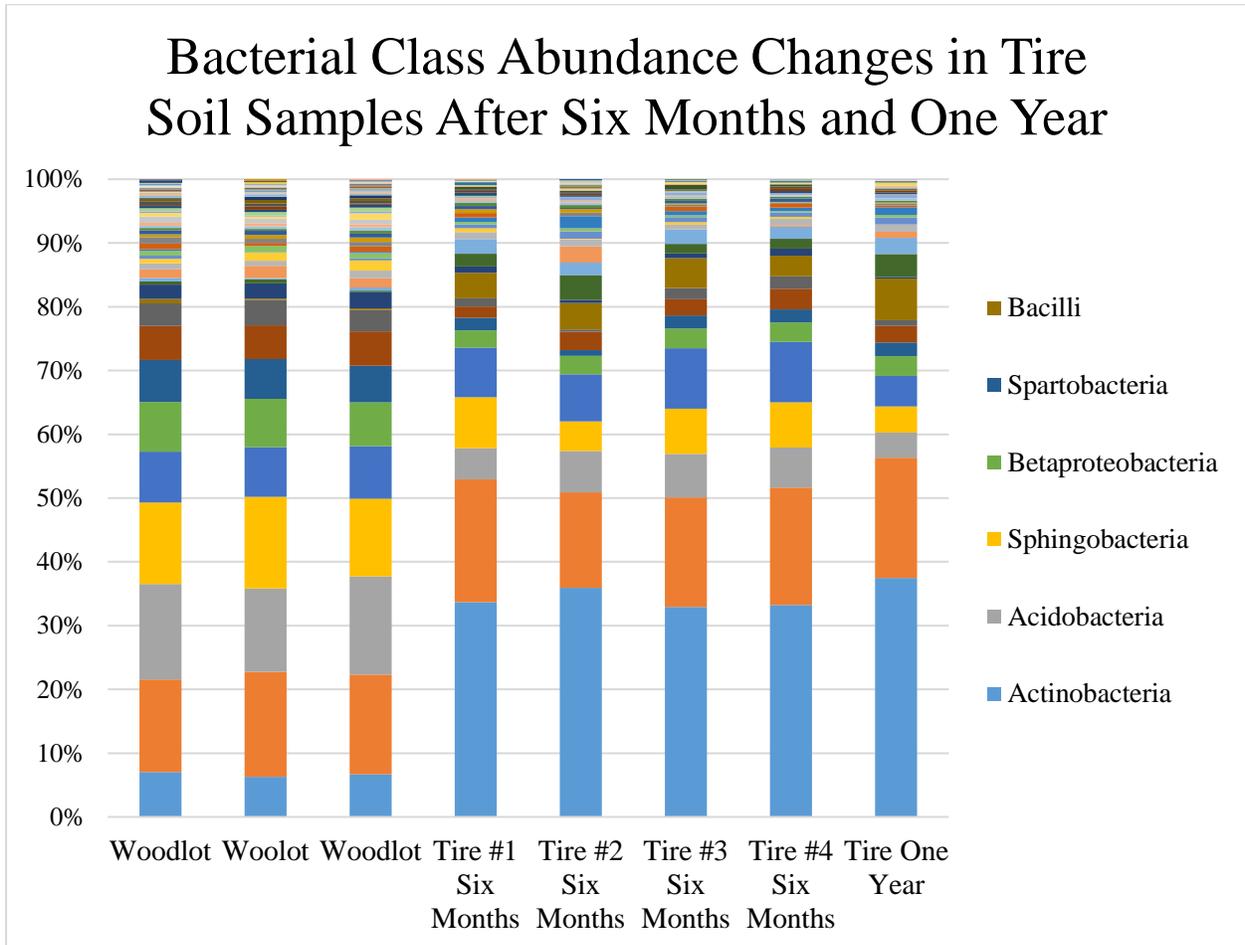


FIG. 4—Bacterial class abundance of soil samples collected from the woodlot of origin (left) and soil samples collected off of the tire after six months and one year in room temperature storage. Evidentiary profiles exhibited notable increase in *Actinobacteria* and *Bacilli* and a decrease in *Acidobacteria*, *Sphingobacteria*, *Betaproteobacteria*, and *Spartobacteria*.

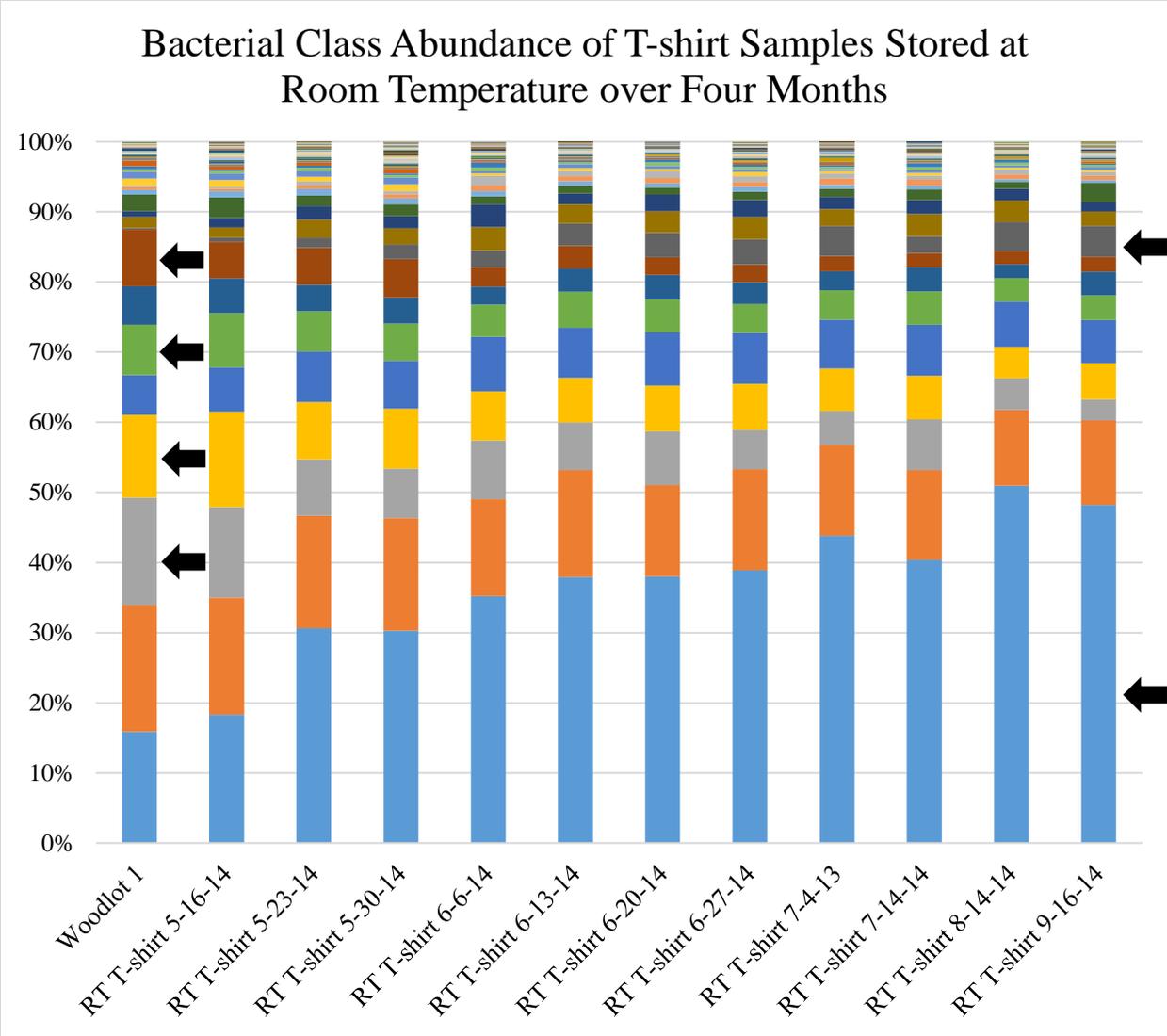


FIG. 5—Bacterial class abundance chart of room temperature t-shirt soil samples collected over the four month sampling period compared to soil collected from the woodlot of origin. Evidentiary soil samples exhibited notable increases in *Actinobacteria* and *Bacilli* (designated by arrows in ascending order on the right of figure) and decreases in *Acidobacteria*, *Sphingobacteria*, *Betaproteobacteria*, and *Spartobacteria* (designated by arrows in ascending order on the left of figure). RT=Room Temperature.

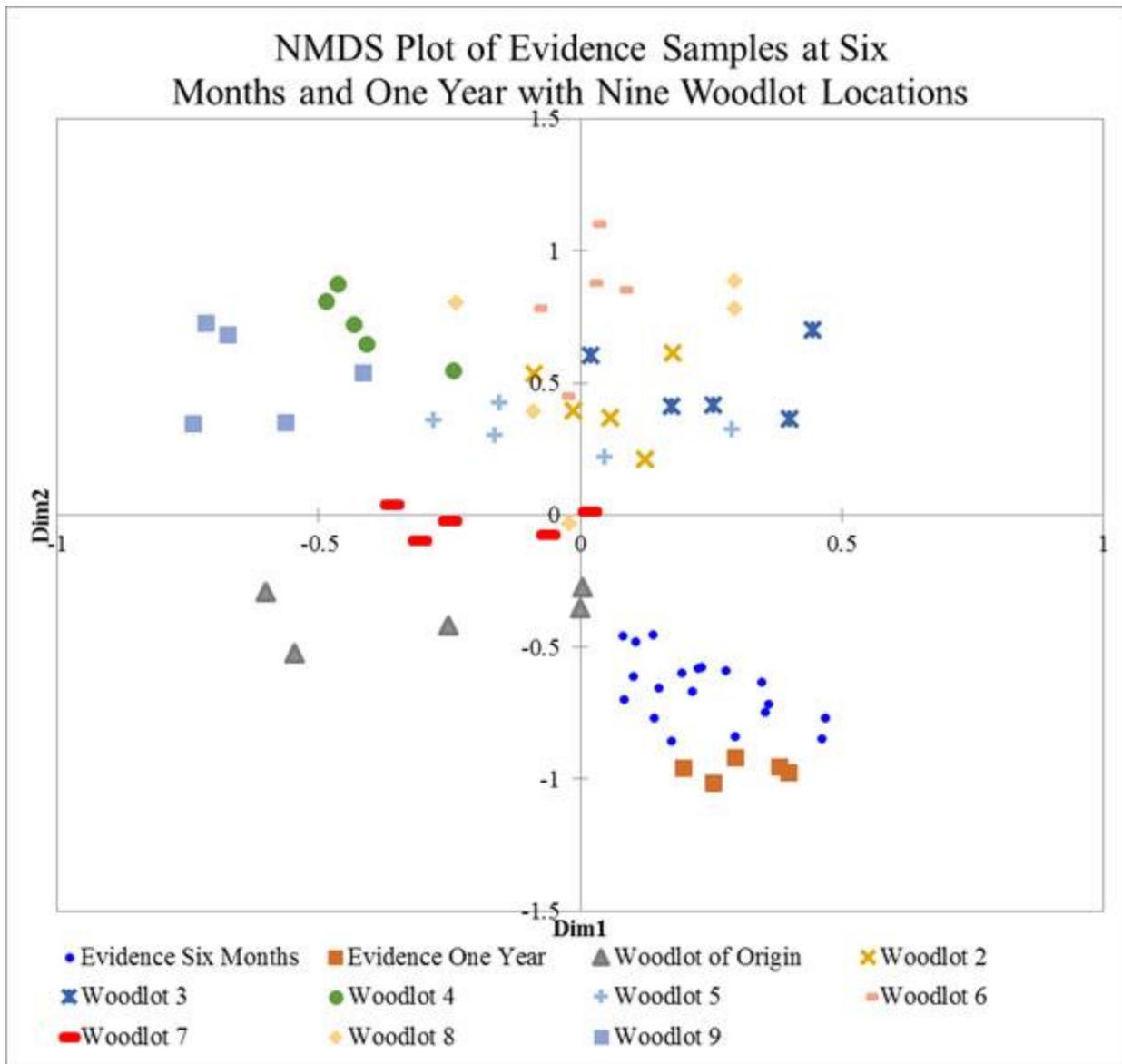


FIG. 6—NMDS plot ordinating evidentiary and nine woodlot soil samples. Evidence samples collected after both six months and one year in storage clustered together, nearest the woodlot of origin, with the year samples plotting slightly further away.

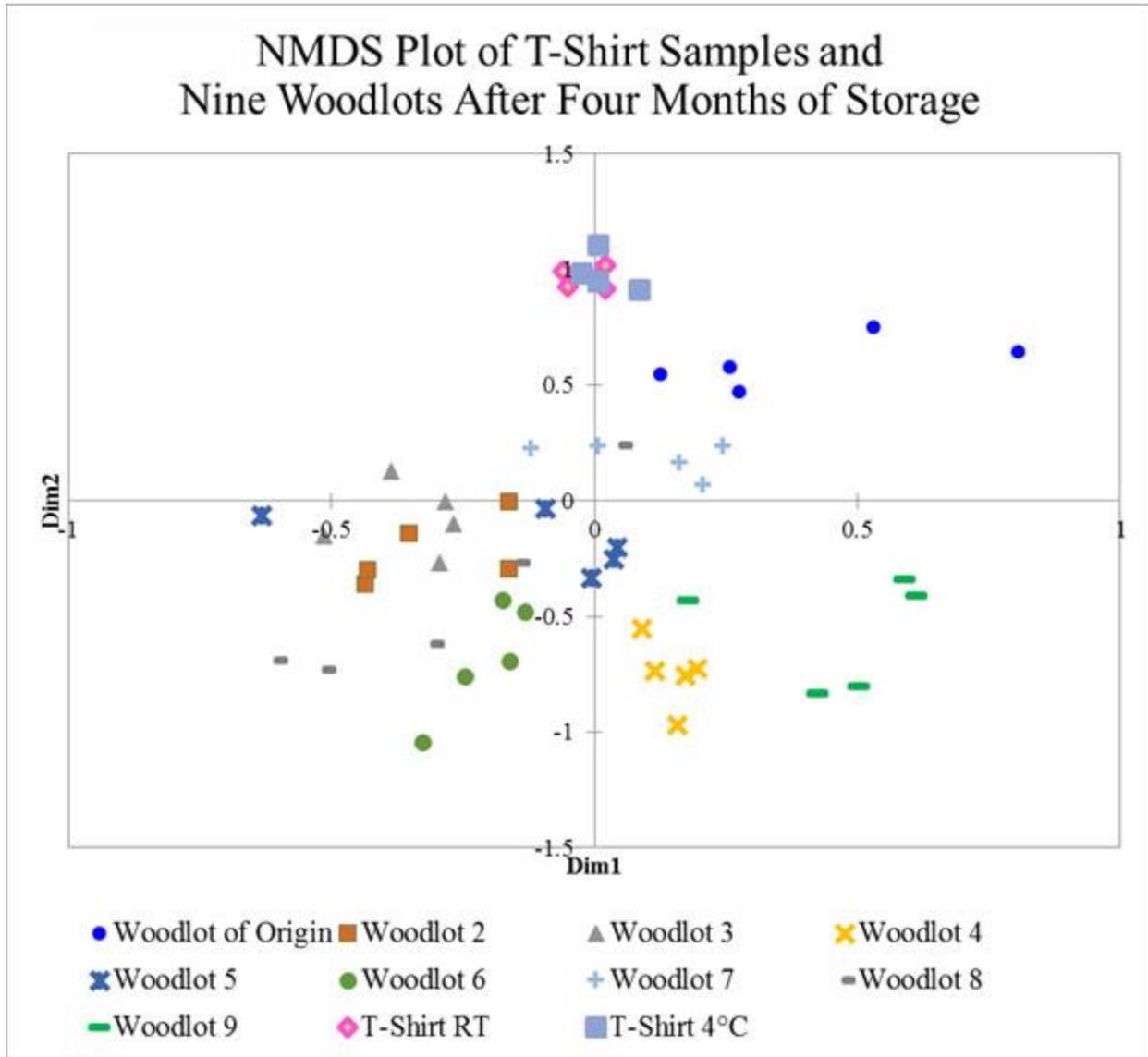


FIG. 7—NMDS plot of nine woodlots and t-shirt evidentiary samples after four months of storage. T-shirt samples from both storage temperatures drifted away from all woodlots, remaining closest to the woodlot of origin cluster. RT=Room Temperature.

A primary goal of this research was to investigate varied techniques for analyzing next-generation sequencing data, which ideally can be both easily interpreted by a jury, and possess objective qualities to meet the recommendations found in the 2009 NAS report (National Research Council). The amount of data produced via next-generation sequencing makes meeting these requirements challenging, as it is impossible to simply look at hundreds of thousands of

DNA sequences and come to a definitive conclusion regarding evidentiary soil's origin.

Therefore, techniques that can sort and display these datasets are vital. In reality, it is unlikely a single analysis technique will encompass both forensic needs, thus it is quite possible more than one technique will be necessary for forensic purposes.

The most straightforward strategy for simplifying the datasets produced in studies such as this is to group them based on bacterial reference DNA sequences, and visualize the groupings using bacterial abundance charts. These charts provide a graphic quantification of what bacteria are present in a profile, which should help facilitate expert witness testimony. For the forensic scientist, comparisons of abundance charts can elucidate separations or groupings in other analysis techniques, adding confidence to results obtained using more objective methods. Unfortunately, we cannot rely solely on visual comparisons of abundance charts to associate soil bacterial profiles, as such comparisons will likely be highly subjective, a well-known shortcoming of many of the forensic sciences (National Research Council, 2009).

Ordination of soil bacterial profiles provided information that abundance charts did not. The strength of NMDS is its ability to group bacterial profiles from a given location while simultaneously distinguishing bacterial profiles that do not belong with that cluster, both of which are easily visualized. However, NMDS clusters are formed based on the relative dissimilarities of all soil samples being compared, meaning a single highly dissimilar sample can force unrelated samples together, resulting in misleading indications of similarity among them. Likewise, when only samples from the same location are analyzed, they will be separated in an NMDS plot even though they are quite similar. Given this, the designation of clusters in NMDS plots is itself subjective; thus, a purely objective analysis method is still required if complete characterization of the data is to be developed.

Supervised classification techniques use measures of dissimilarity to compare bacterial profiles in an objective manner, and have been employed for tracing soils back to their point of origin (Yang et al., 2006). For instance, *k*-NN uses training sets of known samples, resulting in models and classifications that take into account slight temporal and spatial bacterial fluctuations that can be visualized but not objectively described through abundance charts and NMDS. The next-generation sequencing based research presented here, which has much higher resolving power than past techniques, reached an average *k*-NN assignment accuracy of 96.4%, accentuating the utility of supervised classification techniques for forensic application. However, *k*-NN offers no explanation as to how assignments are made, and it is a hard classifier, forcing assignment of all samples whether or not they actually belong in a specified group, which could produce misleading conclusions.

Plainly, there are plusses and minuses associated with each analysis technique, however, based on this research, it seems worthwhile to utilize bacterial abundance charts, NMDS, and *k*-NN concurrently to compare soil bacterial profiles. Clear visual representations that may aid the jury's understanding of soil evidence are generated by abundance charts and NMDS. The two techniques act in a complementary manner wherein the former provides a categorization and quantification of the copious sequences, and the latter produces relative similarity information. Together they can then be used by the expert witness to explain classifications in *k*-NN, which itself delivers the objective assignment of samples to their location of origin. In combination, these three analysis techniques can effectively be used for next-generation sequencing data, providing an avenue for forensic soil analysis to enter the courtroom.

The next major goal of this research was to apply these analysis methods to next-generation sequencing data and determine its feasibility for use in forensic casework. The most

basic step was to determine if soils from diverse habitats could be differentiated. In previous soil studies, researchers were able to differentiate a small number of habitats in multidimensional space or through the presence or absence of specific bacteria (Fierer and Jackson, 2006; Lauber et al., 2009; Lenz and Foran, 2010; Kim et al., 2013; Lauber et al., 2013), but overlap among habitats frequently occurred. In the current research, similar distinguishability and overlap existed when examining soils from many habitats simultaneously, but this was resolved when fewer habitats were compared in both NMDS plots and via *k*-NN analysis. The increased differentiability may have been the result of greater resolution of next-generation sequencing technology, where data production and bacterial profile generation is far more extensive and objective than previous methods. Improved resolution may also have been achieved through the use of more robust analysis techniques. Supervised classification was not attempted in past forensic soil research, and soil sample association might have improved had it been employed. Regardless of the reason for increased differentiation, higher resolution next-generation sequencing and supervised classification greatly improved our capacity to differentiate soils from diverse habitats.

Once it was apparent that soils from environmentally diverse locations could be differentiated based on 16S bacterial profiles, the next step was to determine if the same was true for similar habitats at locations in very close proximity. This presents a much greater challenge, as similar habitats are likely to share many of the same environmental characteristics that affect bacterial communities. This challenging task was accomplished in the current research through differentiation of nine woodlot sites. Although soils collected from two of these woodlots showed substantial intra-location variation, clusters of all locations were resolved by ordinating fewer woodlots in NMDS plots. However, this variation within a woodlot could not be overcome

in *k*-NN analysis, where one woodlot had no samples correctly classified. The reason for this is unclear, yet despite these few exceptions, the very similar habitats were largely differentiated and correctly classified in this study, validating the high resolution of the bacterial profile data produced via MiSeq sequencing.

With the strength of next-generation sequencing of soil bacteria for differentiating diverse and similar habitats established, factors that may affect bacterial profiles within a location were considered. It is essentially impossible to collect known soil samples exactly when a crime occurs; therefore, temporal changes in bacterial makeup must be assessed. Past studies examining change over time (through terminal restriction fragment length polymorphism [Meyers and Foran, 2008] and pyrosequencing [Lauber et al., 2013; Hopkins, 2014] of the 16S locus) have shown considerable differences in bacterial profiles collected across seasons, resulting in intermingling of habitats during analysis. Relatively little temporal change was evident in the current study, where bacterial profiles remained stable across seasons, likely due to the superior resolving power and data production of newer next-generation sequencing platforms.

It is also unlikely that known forensic soil samples will be collected from the exact spot to which the evidence item was exposed, but instead could be collected feet, yards, or greater distances away, stressing the importance of considering spatial variability of soil within a location. Microenvironmental factors such as foliage, pH, nutrient supply, etc. have been proposed for differences in bacterial profiles over small distances (Ettema and Wardle, 2002; Eichorst et al., 2007), although any number of factors might come into play. Bacterial variation was reflected in soil samples collected in the current research, where surface soil across a habitat and soil from various depths did not always cluster tightly in NMDS plots. Despite these within-site

bacterial profile differences, samples were still correctly classified using *k*-NN analysis, emphasizing the importance of collecting many knowns for building training sets.

The final studies detailed here combined similar habitats, temporal changes, and mock evidence to assess traceability of soils in more realistic forensic scenarios. Most importantly, soil on mock evidence traced back to its location of origin 100% of the time even after a full year of storage, highlighting the success of the sequencing technique and supervised classification for forensic soil analysis. Further, similar abundance changes occurred across all stored evidentiary items, where specific classes of bacteria increased or decreased over time. Predictable transformation of bacterial profiles on evidence has the potential to act as a biological clock for how long soil has been removed from a habitat, providing a valuable tool in criminal investigations. Although evidence was only exposed to soil from a woodlot, it is possible these same bacterial class abundance changes will take place in soil from other habitats, as the taxonomic classes that changed are common in soils. Overall, evidence studies from this research show that soil bacterial profiles can be effectively used to trace evidence back to a location of origin regardless of the material it existed on and the length of time it was stored.

Together, the application of next-generation sequencing and three data analysis methods to compare soils collected from diverse and similar habitats, within a habitat, and from evidentiary items, have revealed the exceptional potential of this novel forensic soil analysis technique. The vast majority of soil samples in this research successfully traced back to their location of origin, through the comparison of bacterial profiles. More importantly, the results were both objective and easy to interpret for the forensic scientist and layperson. Based on this, forensic soil evidences' value will be greatly increased through implementation of these methods

by crime laboratories, where they can objectively link evidence, a victim, or a suspect to a crime scene.

## INTRODUCTION

### Forensic Soil Investigation

In forensic investigations, soil can prove an invaluable evidentiary source for linking a suspect or victim to a crime. Potentially found on shoes, tires, shovels, or other objects, the virtually unlimited types of soil and their geospatial distribution can make such evidence highly probative (Saferstein, 2002). These properties for linking an individual with a geographic location, though explored as far back as the works of Sir Arthur Conan Doyle (see Alden, 2014), were rarely used in a forensic context until the early 20<sup>th</sup> century.

Among the first documented criminal cases involving soil evidence were a silver theft on the Prussian railroad in 1856 (Ritz et al., 2008) and the investigation into the death of Eva Disch in 1908 (Bressan, 2010). In the first, sand had been used to replace silver in a barrel being transported along the railways. Professor Ehrenberg, a scientist from Berlin, was asked to examine several samples of sand collected from each railway stop, identify visual consistencies, and determine where the switch had taken place. One station was identified, and the guilty railroad employee was arrested. The death investigation involved a young woman who was strangled in a bean field in Frankfurt, Germany. Crucial evidence found at the scene included a soiled handkerchief with particles of hornblende, snuff, and coal (Bergslien, 2012). After Disch's identity had been established, local authorities identified Karl Laubach as a suspect. Investigators enlisted Georg Popp, a chemist, to examine the soil found on Laubach's clothing. Popp identified two distinct layers of sediment in the cuffs of the pants worn by Laubach the day Disch was murdered. One was consistent with the soil at the crime scene. The other contained mica, which was consistent with the path between the scene and Laubach's home. Combining these two pieces of evidence and challenging Laubach resulted in a confession to the murder.

While the admissibility of Ehrenberg's and Popp's analyses in these cases would be questionable today, forensic scientists have continued developing more precise, accurate, and acceptable methodologies for the examination of soil evidence. However, the recent National Academy of Sciences (NAS) report (National Research Council, 2009) has called many of the practices used in forensics into question, soil examination included, requiring a reassessment of what analysis techniques are currently used and how they can be improved. Additionally, the *Daubert* ruling has enforced the need for forensic science to develop accepted and peer reviewed procedures, with established error rates (*Daubert v. Merrell Dow Pharmaceuticals*). These requirements have encouraged forensic scientists to employ more robust data generation techniques, as well as incorporate the use of objective statistical measures for analysis.

### **Classic Soil Analyses**

Expanding upon the work of Ehrenberg, Popp, and others, forensic geologists in the 20<sup>th</sup> century aimed to utilize the diverse characteristics of soils to classify, compare, and identify them (Saferstein, 2002). A collection of tests exist to analyze attributes of soil, dividing into four categories that focus mainly on its physical properties: general, microscopic, non-microscopic, and chemical (Saferstein, 2002). One shortcoming of these techniques is that they often require large amounts of soil for testing, which may be unrealistic in a forensic scenario. Each, unless the soil contains a rare compound or element, measures class characteristics, leading to only a general association between known and unknown samples. Further, the subjectivity in interpretation, e.g., matching soil color to a Munsell color chart, and lack of statistical significance measures (Pye, 2007) seriously limit the value of soil evidence. Similar types of soils may not be differentiable from each other using these techniques, increasing the possibility of a false association among evidentiary samples. Based on these limitations, it is clear that there

is a need for techniques that capture the unique characteristics of soil for better characterization and identification of this complex medium.

### **Molecular Analyses of Soil Bacteria**

It has been estimated that  $4 \times 10^7 - 2 \times 10^9$  prokaryotic cells are present in one gram of soil, representing up to 18,000 different genomes, which may in turn be an underestimate (Daniel, 2005). The potential breadth of microbial diversity in soil, considering only the prokaryotic contribution, is staggering. Recent advances have allowed forensic scientists to assay the bacterial metagenome with the goal of using bacterial communities to link evidentiary and known samples. Several techniques exist for assaying bacterial communities in soil; however, only a few have gained a footing in the forensic sciences. The commonly used techniques include denaturing gradient gel electrophoresis (DGGE [Muyzer and Smalla, 1998]), amplicon length heterogeneity-polymerase chain reaction (ALH-PCR [Moreno et al., 2006]), and, most popularly, terminal restriction fragment length polymorphism (T-RFLP [Liu et al., 1997; MacDonald et al., 2008]). Although all of these techniques can provide information about a bacterial community, the massive amount of bacteria in soil makes their resolution power low and the interpretation of the data potentially subjective.

### **Analysis of Soil Bacteria at the Michigan State University Forensic Biology Laboratory**

For the last eight years, forensic biologists at Michigan State University have been studying various methodologies for identifying soil samples based on their microbial populations. The goals, through utilization of T-RFLPs, were to characterize how bacterial populations differ among habitats and within the same habitat over time and space. These initial research questions needed to be addressed to fully document how feasible bacterial profiling could be in a forensic context.

First, Meyers and Foran (2008) examined spatiotemporal considerations. Soil samples were collected from five habitats: an agricultural field, a marsh edge, a yard, a deciduous forest, and a sandy woodlot approximately 100 miles distant from the other sites. Samples were taken from a central location once every month for one year, with auxiliary samples taken ten feet distant from the center point in each cardinal direction every three months. This sampling scheme aimed to address changes in bacterial composition month-to-month and over short distances. DNAs were extracted and the entirety of the 16S rRNA gene was amplified, incorporating an end-labeled primer. Amplicons were digested with *MspI* and capillary electrophoresed. Normalized similarity indices were calculated for each electropherogram and analyzed using single factor ANOVA as well as multivariate ANOVA. The authors found that similarities were highest among collections within habitats, although substantial differences in bacterial profiles were sometimes seen over very short distances. The extent of temporal change was much higher, where within habitat collections fluctuated throughout the year and among habitat differences were not significant.

Lenz and Foran (2010) sought to differentiate among the same five habitats using T-RFLPs, through a more focused approach. *Rhizobia* DNAs were amplified using *recA* gene specific primers, with amplicons subjected to *RsaI*, *MspI*, or *DpnII* digestion and capillary electrophoresis. Relationships among the samples' T-RFLP profiles were evaluated with an ordination technique known as multidimensional scaling (discussed below). In two-dimensional multivariate space, the deciduous forest and sandy woodlot could almost always be differentiated regardless of the restriction enzyme used, while the other three habitats were heavily comingled when all five were plotted together. Accurate differentiation of sites, except for the agricultural field, was accomplished when pairwise comparisons were projected into two dimensions. The

analysis of ‘unknown’ samples met with variable success depending on the endonuclease used and the habitats being compared; however, the correct association was seen more often than not. These results further support the idea that bacterial communities can be used to differentiate unrelated habitats. The use of multidimensional scaling does not allow attribution of statistical significance, although it can represent the underlying patterns within these data, with useful information displayed in the ordination plots.

Taken together, these studies have shown that T-RFLP analysis is a viable tool for the study of microbial populations in soil. Recently however, more powerful technologies have come into use that allow for an even greater understanding of soil bacterial metagenomics. A promising technique developed in the last ten years is massively parallel sequencing (next-generation sequencing). Introduced in 2005, it is an alternative to Sanger sequencing (Margulies et al., 2005; reviewed by Shokralla et al., 2012). Next-generation sequencing platforms have the ability to generate vast amounts of data in short periods of time, and do not require the creation of clone libraries (MacLean et al., 2009), which facilitates metagenomic analysis of complex substrates like soil. A great number of next-generation sequencing platforms exist, each with its own chemistries and detection systems; however, they can be broken down into two major groups: PCR-based sequencing, which includes Roche 454, Illumina MiSeq, and Applied Biosystems SOLiD, and single-molecule-based technologies, which includes Helicos BioSciences HeliScope and PacBio RS SMRT (MacLean *et al.*, 2009; Metzker, 2010; Shokralla *et al.*, 2012). Hopkins (2014) utilized 454 pyrosequencing to compare replicate soil samples collected in each of three habitats near the Forensic Biology Laboratory at Michigan State University. He studied various statistical techniques to analyze the sequence data, each already used by the microbiology community, to ascertain which had the most forensic utility.

The simplest way to examine the massive amounts of data produced by next-generation sequencing platforms is to visualize the bacterial communities through abundance charts (FIG. 1). These charts are generated from taxonomic data, categorizing and quantifying the bacteria that make up a total profile, allowing subjective comparison among several samples. Microbiologists have used abundance charts to characterize bacterial profile changes after environmental events such as drying and rewetting of soils (Barndard et al., 2013) or exposure to strong chemicals (Sutton et al., 2013). For these types of comparisons, charts are generated at the phylum level, where changes brought on by harsh environmental conditions are evident. For forensic purposes, soil samples being compared are not necessarily subjected to vastly different environmental conditions; therefore, a more informative level (such as the taxonomic class level) may be more probative. Finally, the subjectivity of abundance chart interpretation severely limits their utility as a stand-alone technique in forensic science and more objective techniques that can produce a measure of similarity/dissimilarity are necessary.

Pairwise comparisons are objective statistical techniques that can be used to determine whether two bacterial profiles differ significantly, which has the potential to provide forensic utility. It is impossible to statistically compare a sample size of two; however, by resampling profiles, datasets can be built to calculate a p-value for the degree of random similarity between two bacterial profiles. One such method, the LIBSHUFF statistic, was introduced by Singleton et al. (2001) for the statistical comparison of 16S rRNA clone libraries using the approximation form of the Cramér-von Mises test statistic for curve fitting and Monte Carlo simulations to calculate significance. An updated version of LIBSHUFF,  $\int$ -LIBSHUFF, developed by Schloss et al. (2004), uses the exact and integral form of the Cramér-von Mises test statistic, which measures the number of singleton sequences in one bacterial profile compared to another. This



A second pairwise comparison technique, the Unique Fraction Metric (UniFrac), introduced by Lozupone and Knight (2005), uses phylogenetic distance to compare two bacterial profiles through tree building (FIG. 2). UniFrac also uses a Monte Carlo procedure to calculate the statistical significance of the difference between samples by randomizing the sequences at the ends of the branches while keeping the tree constant. A p-value is determined through the percent of random trees that share the same or greater fraction of unique branch lengths with the original tree.

When multiple pairwise comparisons are made using either of these methods, a Bonferroni correction can be applied to account for the large number of tests conducted. This correction aims to preserve a family-wise error rate, while reducing the probability of a false positive (statistical significance) (Kaltenbach, 2012). As the number of tests increases, the p-value can become incredibly small to the point where no meaningful results can be determined; therefore, the Bonferroni correction should only be used for a small number of tests (Kaltenbach, 2012), presenting a limitation for forensic soil analysis where a large number of samples may be analyzed.

Hopkins (2014) used both of these procedures to examine soil biological replicates, and found that on some occasions even the replicates were assessed as significantly different. Given this, during the current research an examination of pairwise comparisons' utility was performed to assess their sensitivity. A single soil sample was divided into four, DNA extracted, and the 16S genes sequenced. The four resulting bacterial profiles were compared using both  $\beta$ -LIBSHUFF and (unweighted) UniFrac, which were found to differ significantly 5/6 times and 4/6 times respectively. This result illustrates the hypersensitivity of pairwise comparisons, where slight bacterial fluctuations in soil collected in the same trowel scoop proved to differ

significantly. Forensic soil samples will rarely be collected at the same time or spot, let alone from the same sampling bag, thus pairwise comparisons of next-generation sequencing data are not feasible for forensic soil analysis, as their sensitivity can cause two very similar samples to be declared different. In this regard, objective analysis techniques that are not overwhelmed by slight fluctuations in the bacterial community are still necessary.

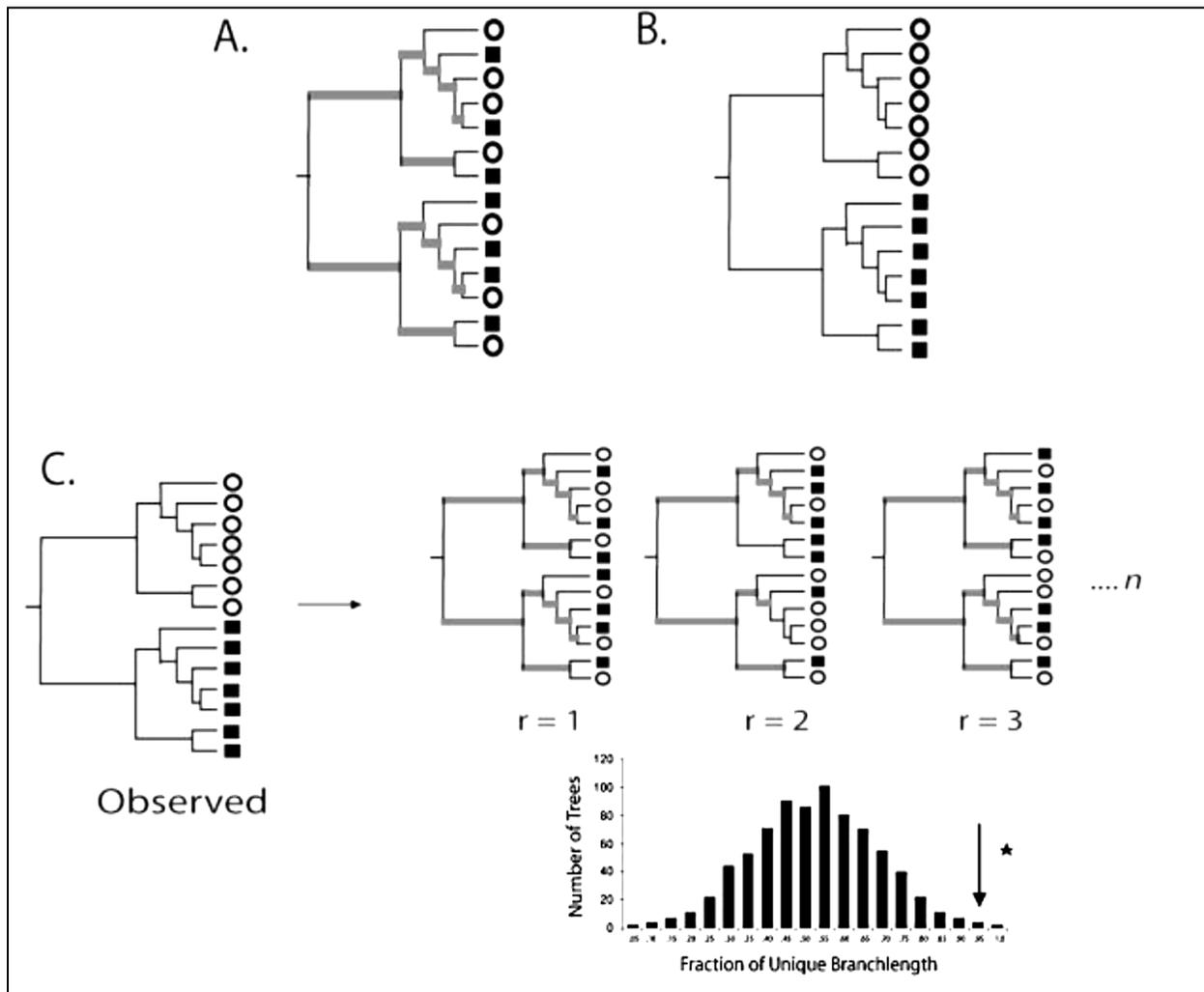


FIG. 2—Illustration of how UniFrac compares two bacterial profiles (represented by circles and squares). First, a tree is built to assess the evolutionary relationship of phylogenies in each profile. Phylogenies can either be intermingled (A) or completely unique (B). C demonstrates the Monte Carlo iterations for assessment of significance between the circle and square communities in B. The histogram is a composite of all random trees with the star indicating the unique branch

length of the original tree and the arrow indicating the chosen p-value (e.g., 0,05), or proportion of random trees that had an equal or greater fraction of unique branch length (bold branches) as the original tree. In this example, the p-value is less than the unique branch length of the original tree, indicating that the two communities are significantly different. Taken from Lozupone and Knight (2005).

For further analysis of next-generation sequencing data, a measure of dissimilarity between two bacterial profiles can be calculated, such as  $\beta$ -diversity (also encompassed by UniFrac), which was defined by Whittaker (1960) as “the extent of change of community composition, or degree of community differentiation, in relation to a complex-gradient of environment, or a pattern of environments”. Two commonly used  $\beta$ -diversity indices, the Bray-Curtis dissimilarity index (1957), and the Sørensen-Dice coefficient, described independently by Sørensen (1948) and Dice (1945), were used by Hopkins (2014) to investigate the diversity in bacterial communities among several soil samples. The Bray-Curtis index measures the structural dissimilarity between communities, i.e., not only is shared membership considered, but also the number of individuals in each bacterial group. On the other hand, Sørensen-Dice calculates community membership differences by only assessing shared membership between the bacterial profiles. This concept is illustrated in Table 1, with the two communities being compared, A and B, having three species in common. The Bray-Curtis value calculated from the data in this table would reflect the large difference in individuals of species 2 because it considers the number of individuals of each population. Sørensen-Dice does not detect this difference, and would calculate a difference of zero from the data. Note that comparing A to B and B to A results in the same dissimilarity measurement, so the final matrix is square symmetric.

TABLE 1—*Example communities for the explanation of Bray-Curtis and Sørensen-Dice dissimilarity calculations.*

Community	Number of Individuals of Species 1	Number of Individuals of Species 2	Number of Individuals of Species 3
A	5	45	19
B	5	1	18

Similar to pairwise comparisons, Bray-Curtis dissimilarity index calculations may be too sensitive for the data produced via next-generation sequencing. Small fluctuations in bacterial abundance within a habitat will change Bray-Curtis values, making two soil samples appear different when they come from the same location. Sørensen-Dice coefficients, unaffected by abundance differences, have been suggested as a better dissimilarity measure for genetic comparison of very similar samples (Dalirsefat et al., 2009), and thus were selected for the research presented here.

Once dissimilarities are calculated, further statistical analysis of soil samples can be performed. Techniques currently used by microbiologists to this end include nonmetric multidimensional scaling (e.g., Fierer and Jackson, 2006; Lenz and Foran, 2010), hierarchical cluster analysis (e.g., Pye et al., 2006; Heath and Saunders, 2006), and supervised classification (e.g., Yang et al., 2006). These techniques provide additional visualization or analysis of bacterial profile data with both subjective and objective results.

Nonmetric multidimensional scaling (NMDS) is an ordination technique designed to plot data in a low, multidimensional space, where each data point represents a single soil sample and the spread of the data approximates the original (dis)similarities (Borg and Groenen, 2005; Cox, 2001). The final configuration of data points illustrates the relative similarity among samples (i.e., closer data points are more similar than distant ones) (FIG. 3). In NMDS, all data points are

randomly plotted in a given number of dimensions. The points are then systematically adjusted in relation to each other to reduce the amount of stress, which is a measure of how accurately the plot represents the data. When the global minimum stress is achieved, further iterations are discontinued. A stress diagram, or Scree plot, is a measure of the badness-of-fit of the NMDS configuration to the given proximities (Borg and Groenen, 2005). The lower the stress the better the configuration fits the data. In general, stress decreases as more dimensions are added to the plot; however, as the number of dimensions increases, the interpretability of the NMDS plot becomes more difficult. The number of dimensions is identified so that the stress is low and the plot is understandable; generally two dimensions is adequate. There is no globally accepted level of stress for a NMDS plot, and thus acceptance is at the discretion of the analyst, although it is usually selected as the ‘elbow’ in the stress diagram, which can be seen at two dimensions in FIG. 4.

Similar to Scree diagrams, Shepard diagrams (FIG. 5) are an indicator of the badness-of-fit for the final configuration of the data. These plots measure distance and disparities of data points being plotted, where a plot with a perfect stress of zero would have the distance and disparities directly on top of one another (Borg and Groenen, 2005). In cases where stress is nonzero, the vertical distance between each disparity and distance is the error of representation for that pair. The comparison of these points allows for the identification of outliers and possible sources of high stress. The larger the deviation of distances from disparities, the worse NMDS is at explaining the original proximities, and the larger the stress. Forensically, NMDS plots can provide helpful data interpretation. Similar to abundance charts, they offer a visual representation of multiple bacterial profiles. Additionally, NMDS plots provide information on the relative similarities of several samples through grouping and separation. Nonetheless, NMDS still lacks

the objectivity that is desired for forensic analysis, as cluster identification and general association of samples is open to interpretation.

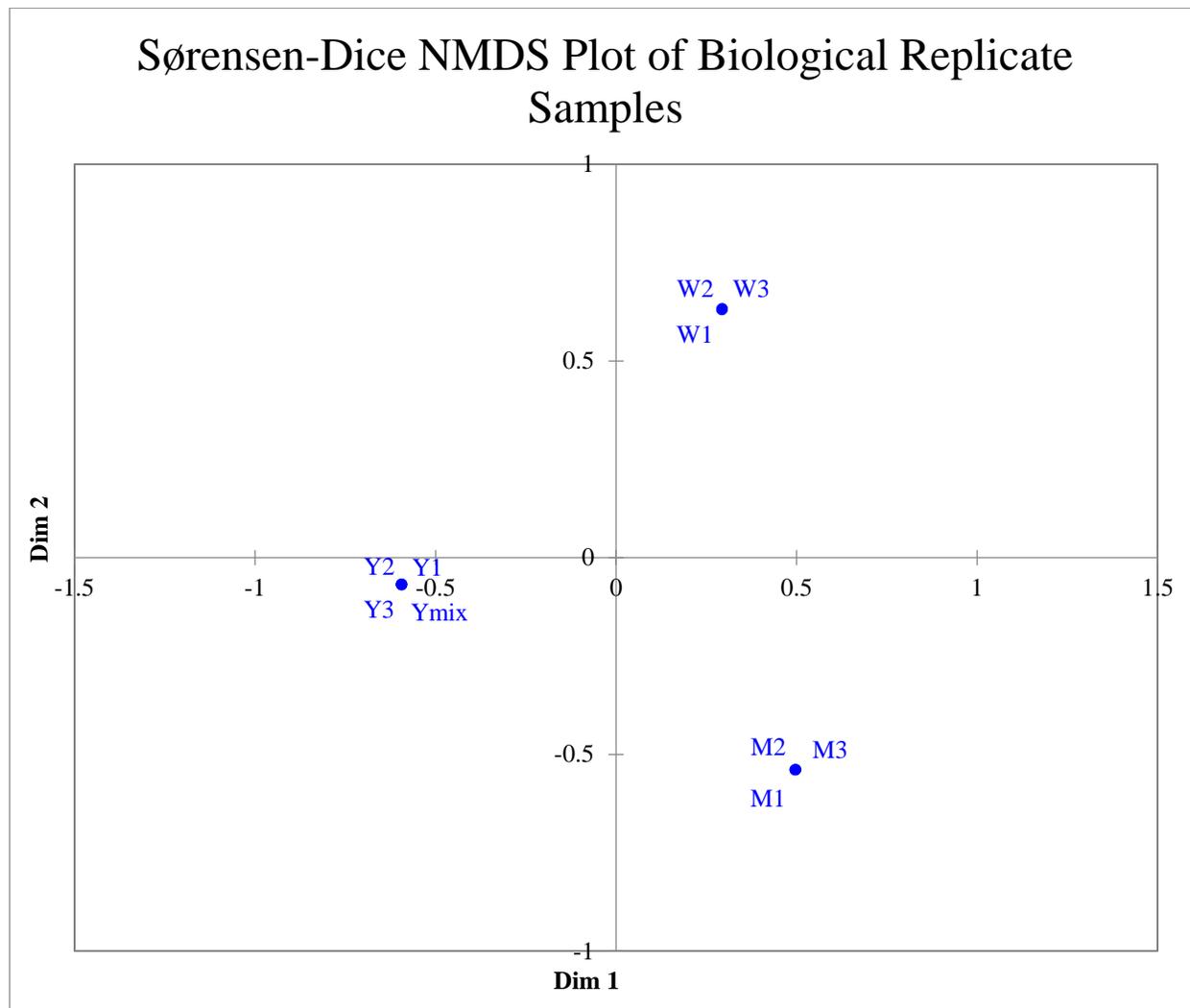


FIG. 3—NMDS plot taken from Hopkins (2014) orienting biological replicate soil samples in two dimensions based on Sørensen-Dice coefficients. Replicate samples fall directly on top of one another and apart from other habitats, indicating their relative similarity.

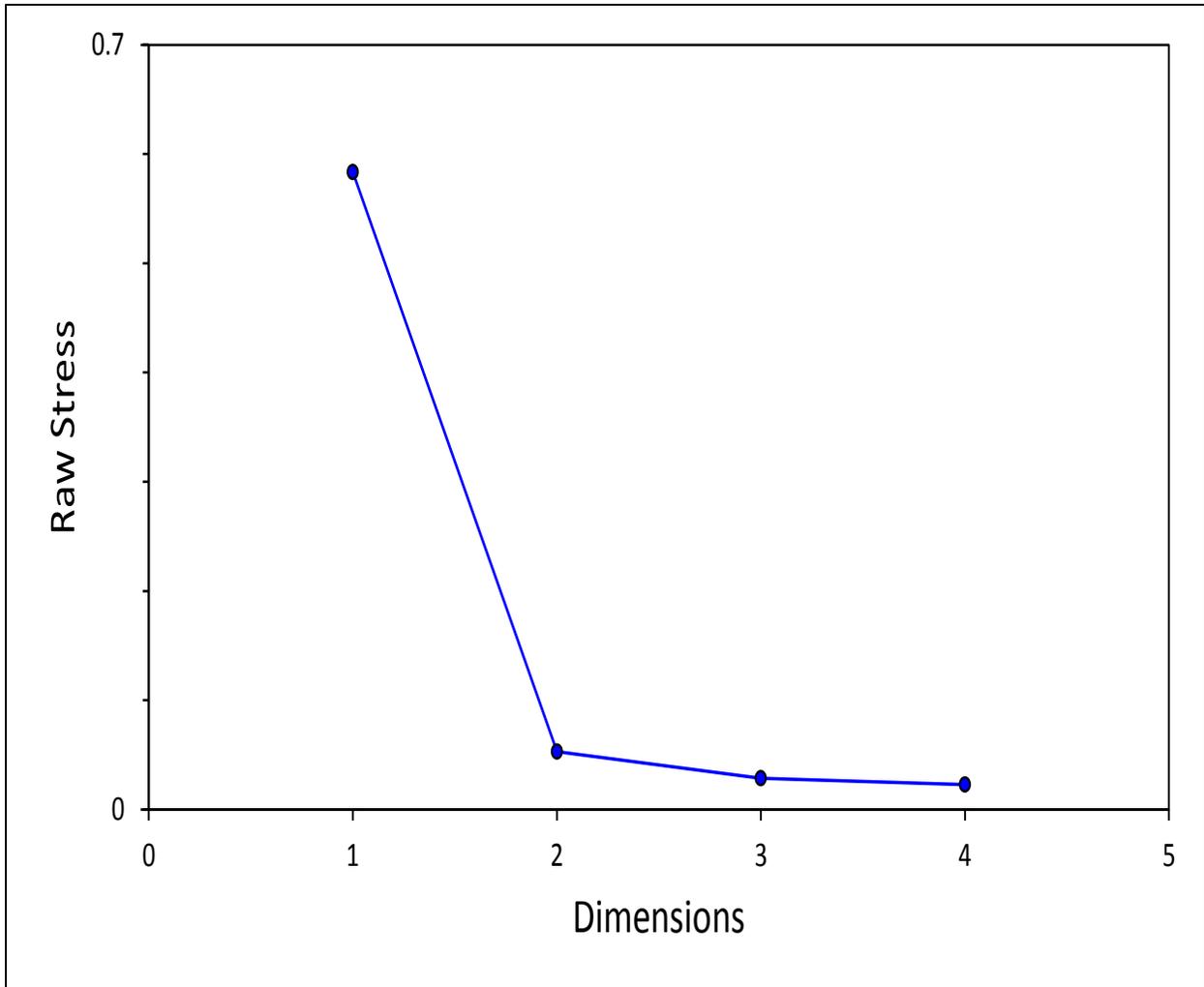


FIG. 4—Typical Scree diagram for multidimensional scaling. Stress is high in one dimension followed by a substantial decrease at two dimensions. Stress continues to decrease into higher dimensionality, though not appreciably, creating the elbow, after which little additional information is gained. Taken from Hopkins (2014).

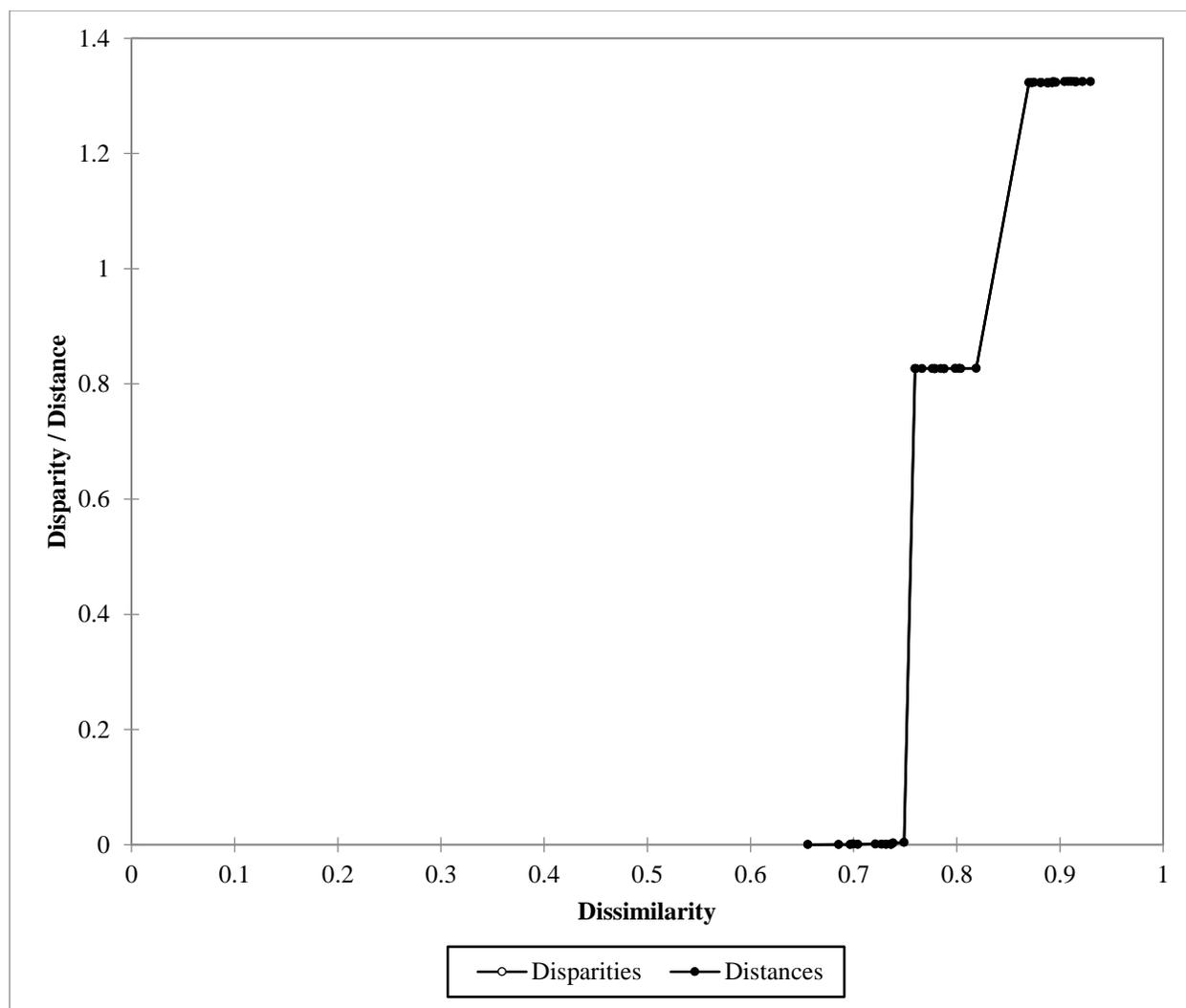


FIG. 5—A Shepard diagram with low stress and good association of disparities and distances. The closer the association of the filled circles (representing approximated distances) and open circles (representing disparities) the better multidimensional scaling is representing the data. Taken from Hopkins (2014).

Another technique for assessing relative similarity among multiple bacterial profiles is hierarchical cluster analysis (HCA), an unsupervised cluster method that allows visualization of distances among samples in a dendrogram (FIG. 6) (Beebe *et al.*, 1998) that allows inference of the relationship between samples (FIG. 7). Three linkage methods, single, complete, and unweighted pair-group method using arithmetic averages (UPGMA) can be utilized, each differing in how distance is calculated between clustered and unclustered samples (Beebe *et al.*,

1998), and thus sometimes resulting in differing dendograms. Another potential disadvantage of HCA is a phenomena known as ‘chaining’ (exemplified in FIG. 8). This occurs when single samples branch together in sequence, artifactually producing difficult to interpret clusters in single linkage clustering. Chaining was often seen in the research performed by Hopkins (2014) on bacterial profiling of soil. Further, due to the subjectivity resulting from differing dendograms depending on linkage method, the utility of HCA for forensic purposes is weak. HCA also provides little information that cannot be gleaned from NMDS plots; therefore, NMDS is preferential to HCA when examining relative similarity among multiple samples for forensic soil analysis.

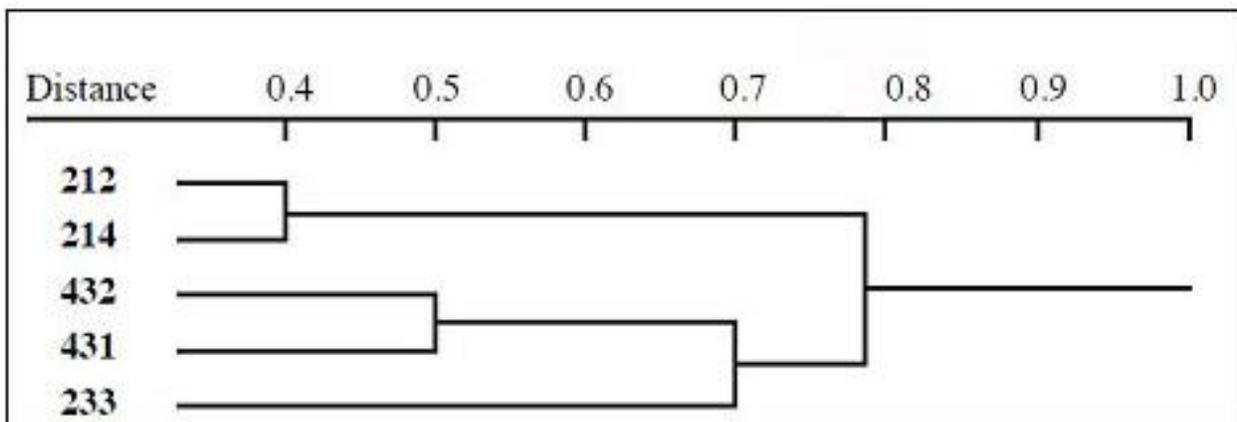


FIG. 6—A representative dendrogram of five samples. The axis along the top represents the distance between the samples in each cluster in multivariate space. Taken from Legendre and Legendre (2012).

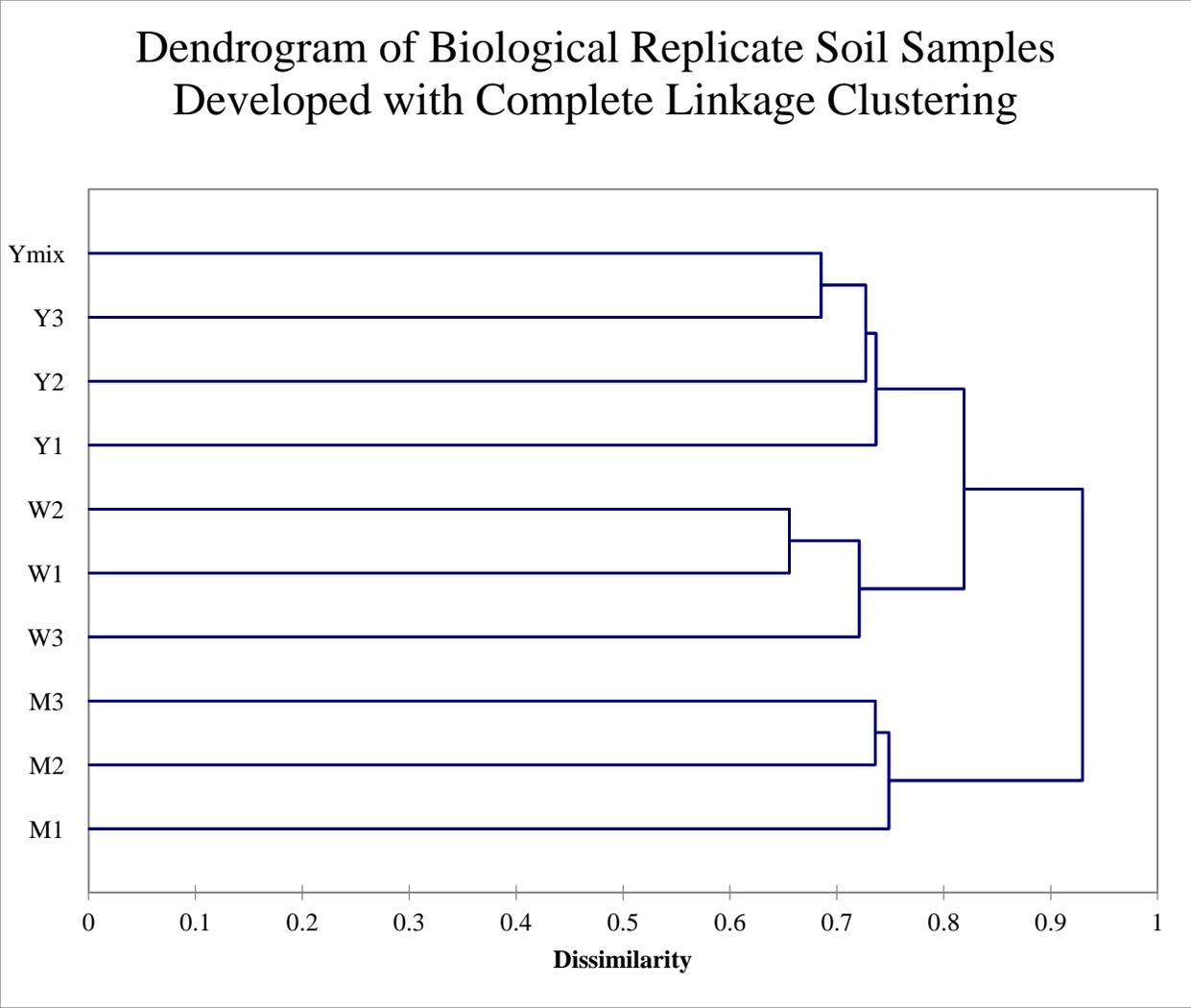


FIG. 7—Example dendrogram taken from Hopkins (2014) showing clusters of biological replicate soil samples. Three clusters were formed where replicate samples were more similar to each other than to other habitats. Y=Yard, W=Woods, M=Marsh.

## Dendrogram of Soil Samples Collected Across the Surface of a Treated Yard Developed with Single Linkage Clustering

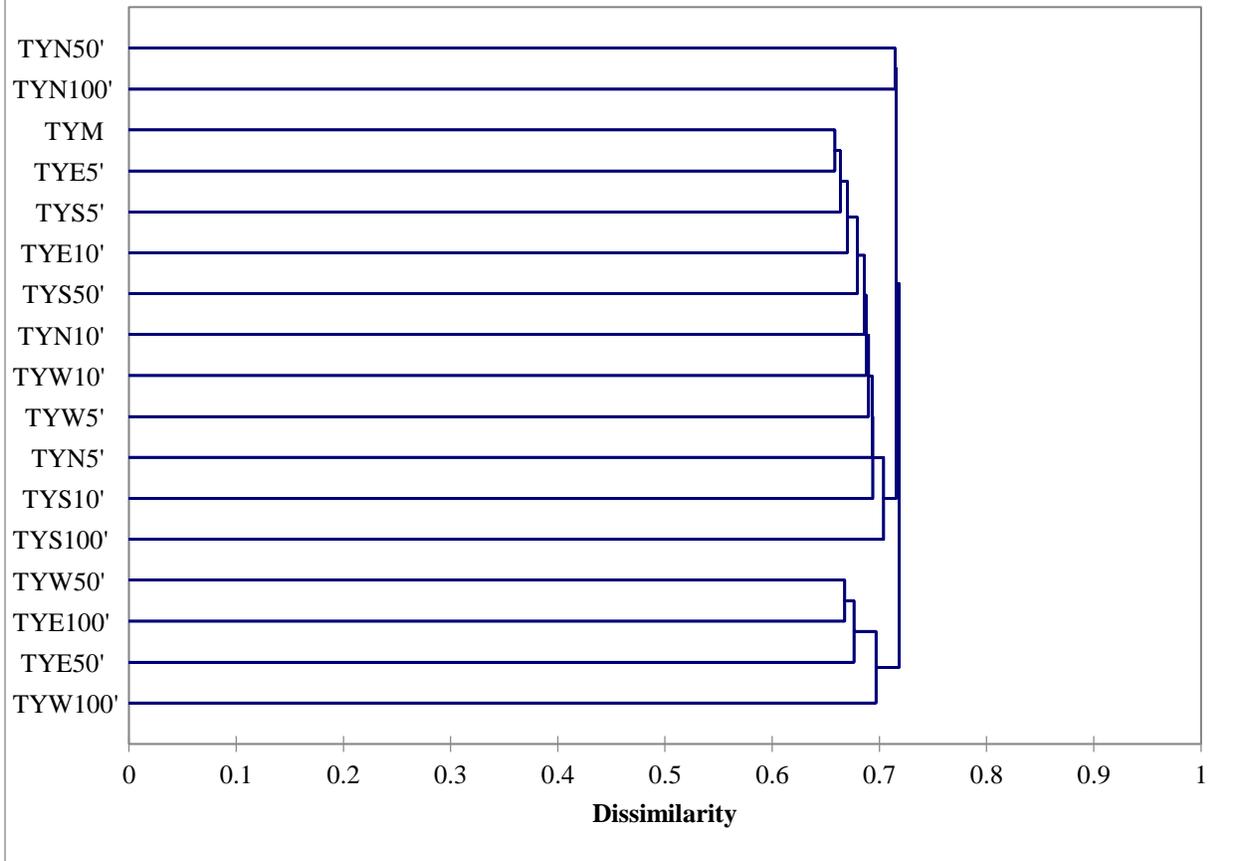


FIG. 8—Example HCA dendrogram illustrating chaining of samples. Dissimilarity in the middle cluster is difficult to interpret due to the rapid, sequential branching of single samples. Taken from Hopkins (2014). TY=Treated Yard.

A final set of techniques that offers an objective analysis of sequence data is supervised classification. These techniques allow an objective assignment of bacterial profiles to a location of origin by building models based on groups of known samples, collectively called training sets (Mohri et al., 2012). Unknowns are then compared to the model and assigned to the closest group or, depending on the technique, to no group at all. Yang et al. (2006) used supervised classification to assign soil microbial communities to their location of origin with approximately

90% accuracy, based on length differences in 16S rRNA variable regions 1, 2, 3, and 9. Their methodology does not hold nearly the resolving power of next-generation sequencing, yet it still highlights the utility of supervised classifiers for bacterial profile analysis.

A popular supervised classification technique is *k*-Nearest Neighbor (*k*-NN). It is considered baseline because it is a relatively simple method that is helpful for establishing the strength of supervised classification analysis for a complex data set such as those produced via next-generation sequencing (Lavine and Davidson, 2006). For *k*-NN analysis, the questioned sample is compared to training sets of known samples. First, the training sets are validated through *a priori* assignment of class membership by individually classifying each training set sample based on the remaining training set members (Pirouette user guide, version 4.0). The unknowns are then classified to a specific group of knowns based on the number of nearest neighbors (*k*) they share with the training sets, with the majority vote determining the final class membership of a sample (Beebe et al., 1998; Pirouette user guide, version 4.0). The higher the *k*, the more confidence can be placed in the classification of unknown samples. For forensic soil analysis, *k*-NN has two very important qualities: it delivers a definitive assignment of unknown samples to a location of origin, and also can capture bacterial fluctuation by building training sets from multiple known samples. However, while supervised classification methods offer several advantages for analysis of bacterial profiles, an explanation of how it generates results (such as a visualization of the data points) is not possible, potentially precluding a simple explanation of how relationships among samples were reached.

Based on Hopkins (2014) and our preliminary studies on the analysis of next-generation sequencing data, the viability of abundance charts, NMDS, and supervised classification for forensic analysis was examined in detail. Abundance charts offer a valuable simplification of the

massive amounts of complex sequence data produced, allowing the user to visualize taxonomic members and their frequency within a bacterial profile. NMDS compares multiple profiles at once to help visualize relative bacterial similarity. Finally, an objective measure of similarity among profiles can be obtained through supervised classification, where an assignment to a particular group is the output. Together, these techniques provide the elements necessary for forensic casework.

### **Feasibility of Next-Generation Sequencing for Forensic Soil Analysis**

As next-generation sequencing technology has evolved in the field of microbiology, a trend has occurred involving platforms that produce far more sequence reads per sample along with substantial bacterial identity information, at a reasonable price. Sequencing by synthesis, (e.g., Illumina MiSeq sequencing) produces sequence reads with a low error rate while maintaining high throughput (Loman et al., 2012). The MiSeq platform sequences several million groups of amplified DNA in tandem, using fluorescently tagged nucleotides. This platform is advantageous over many, including the 454 pyrosequencing methodology used in our preliminary studies, in that read lengths can be up to 500 base pairs, and hundreds of thousands of sequences are produced from each soil sample. This is not to say that other next-generation sequencing platforms would not meet the forensic community's needs, however the MiSeq results presented here seemed superior to our early pyrosequencing results, as others have reported (e.g., Loman et al., 2012). Utilizing this next-generation platform for soil bacterial profile generation meant that its utility for producing data applicable to forensic soil analysis could be assessed so that it can be employed for forensic casework.

Sensabaugh (2009) detailed three essential criteria if any microbial technique is to gain a footing in the forensic sciences: the technique must be robust and repeatable, differentiation of

two locations must be consistent over time and space, and objective statistical measures must be implemented to assess differences between samples. All of these were addressed in the research presented here.

First, soils from various habitats were collected and compared in order to assess the abilities of next-generation sequencing to differentiate locations, which set the baseline for next-generation sequencing of soil sample differentiation. The second, more challenging step was to examine differentiability of very similar habitats, in our case deciduous woodlots. The literature shows that microbiologists generally pool soil samples from ecologically similar sites for microbiological comparisons so as to obtain a generalized profile, however forensic scientists may need to differentiate ecologically similar locations in casework. For instance, if soil is collected from an evidentiary item purported to have been used in a yard, and the crime scene is a different yard, the two bacterial profiles must be differentiable in order to accurately trace the evidence item back to its location of origin. As Sensabaugh pointed out, factors within a habitat also need to be considered, as forensic samples are rarely collected at the exact spot and virtually never collected at the same time an item was exposed to soil. If a crime is committed in a particular area, but the exact location of an evidentiary item's exposure to soil is unknown, knowledge of profile variation over the habitat is necessary. Additionally, if this same soil is left on an item for a week or month before being found by investigators, knowledge of bacterial profile change must be considered to accurately link the item to a location. The effects of both time and space on soil bacterial profiles were examined in this research.

The final requirement outlined by Sensabaugh for the establishment of any new technique in forensic science is the implementation of effective statistical measures that can meet the demands of forensic science. The 2009 NAS report noted the highly subjective nature of analysis

techniques used in several forensic disciplines, resulting in differing expert witness opinion and lower evidentiary value. The report called for more objective analysis methods that produce consistent results. In addition to the need for objectivity, simplification of data is valuable to aid the expert witness. Next-generation sequencing produces huge amounts of data that will be difficult for a layperson to understand and interpret without simplified results and visualizations. A single analysis technique may not meet all criteria needed by a forensic scientist, thus the three techniques outlined above (bacterial abundance charts, NMDS, and supervised classification) were all used to see which, if any, worked for forensic next-generation sequencing data analysis.

In addition to meeting Sensabaugh's requirements for microbial techniques, additional studies mimicking more realistic forensic scenarios were performed to ensure next-generation sequencing of soil bacteria and robust analysis methods were effective. Studies to assess the traceability of soil samples collected from various materials over time and at realistic storage temperatures was performed. The results of this research heightened our confidence that these techniques will perform well in actual forensic casework, where many factors play a role in bacterial profile similarity. Profiling of soil bacteria via next-generation sequencing, the assessment of multiple factors that may influence bacterial profile comparison, and the implementation of various analysis techniques, together allowed for the development of a powerful, new forensic soil analysis method. What follows is a detailed description of the studies performed to meet Sensabaugh's recommendations for the microbial profiling of forensic soil samples.

## MATERIALS AND METHODS

### Soil Sampling Schemes

Table 2 summarizes all soil sample habitat types, GPS coordinates, the study or studies that those samples were included in, and the number of samples collected at each location (n). All soil samples (except those for the vertical space study) were gathered in the same manner: approximately 100 g of surface soil was collected with a garden trowel rinsed with RO water between collections. Three scoops of soil from a 1X1 foot area were homogenized in an 18oz Whirl-Pak® bag (Nasco, Fort Atkinson, WI). Soils were stored at -20°C until extraction. If soils could not be frozen within an hour of collection the bag was kept on ice.

TABLE 2—*Summary of soil samples collected for all studies.*

Sample	GPS Coordinates	Study	n
Marsh	42°42'32.0"N 84°30'53.4"W	Diverse Habitat	5
Fallow Agricultural Field	42°45'06.4"N 84°39'42.8"W	Diverse Habitat	5
Beach	42°45'13.9"N 84°24'16.5"W	Diverse Habitat	5
Coniferous Forest	42°41'11.9"N 84°38'05.1"W	Diverse Habitat	5
Field	42°42'38.9"N 84°31'15.4"W	Diverse Habitat	5
Corn Agricultural Field	42°42'33.5"N 84°28'17.5"W	Diverse Habitat	5
Dirt Roadside	42°48'17.2"N 84°09'33.5"W	Diverse Habitat	5*
Roadside	42°48'03.4"N 84°11'10.1"W	Diverse Habitat	5
Deciduous Woods	42°42'33.7"N 84°31'01.3"W	Diverse Habitat, Temporal, Horizontal and Vertical Space	51
Yard	42°42'39.0"N 84°30'53.5"W	Diverse Habitat, Temporal, Horizontal and Vertical Space	51
Treated Yard 1	42°43'26.6"N 84°28'02.5"W	Temporal, Horizontal Space	44
Treated Yard 2	42°43'44.0"N 84°28'23.4"W	Vertical Space	7
Woodlot 1	42°42'33.7"N 84°31'00.6"W	Similar Habitat, Evidentiary	5
Woodlot 2	42°44'28.2"N 84°27'09.8"W	Similar Habitat, Evidentiary	5
Woodlot 3	42°41'03.3"N 84°31'26.1"W	Similar Habitat, Evidentiary	5
Woodlot 4	42°40'57.2"N 84°28'05.6"W	Similar Habitat, Evidentiary	5
Woodlot 5	42°43'38.9"N 84°30'08.8"W	Similar Habitat, Evidentiary	5
Woodlot 6	42°44'38.9"N 84°28'57.9"W	Similar Habitat, Evidentiary	5
Woodlot 7	42°42'50.8"N 84°28'38.5"W	Similar Habitat, Evidentiary	5
Woodlot 8	42°42'00.8"N 84°31'35.0"W	Similar Habitat, Evidentiary	5
Woodlot 9	42°41'25.6"N 84°27'41.2"W	Similar Habitat, Evidentiary	5
Tire	Stored at room temperature	Year Evidentiary	5
Shoe	Stored at room temperature	Year Evidentiary	5
Sock	Stored at room temperature	Year Evidentiary	5
Shirt	Stored at room temperature	Year Evidentiary	5
Shovel	Stored at room temperature	Year Evidentiary	5
T-shirts	Stored at room temperature	Short-term Evidentiary	44
T-shirts	Stored at 4°C	Short-term Evidentiary	44

\*One sample from this set produced less than 3000 sequence reads and was excluded from further processing.

### *Soil Collection from Diverse Habitats*

Soils were collected from ten diverse habitats in the Greater Lansing area every three months for one year in 2013 and 2014. FIG. 9 is a map of the sampling locations. Four of these habitats were in the Fenner Nature Center, a 134 acre park, magnified by the box in FIG. 9.

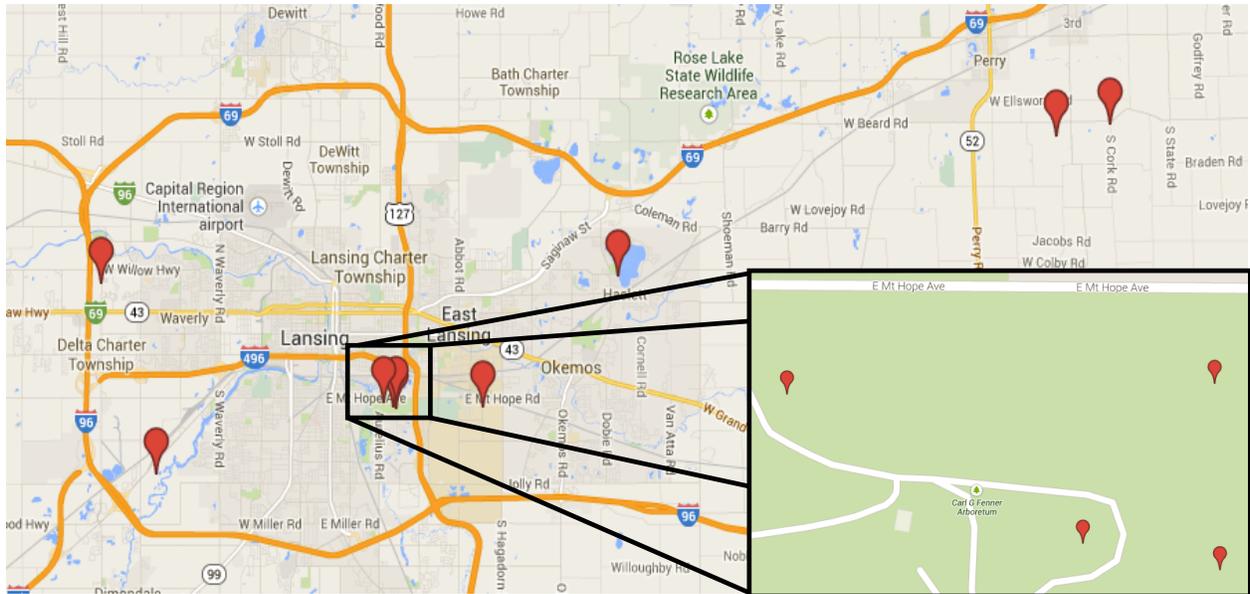


FIG. 9—Map of soil sampling locations for diverse habitat studies. The group of sites on the left are the four sampling locations at the Fenner Nature Center which are magnified on the right.

### *Soil Collection from Similar Habitats*

Soil samples were collected from nine woodlots in the Greater Lansing area once every two weeks for eight weeks in 2014. FIG. 10 represents the sampling locations.

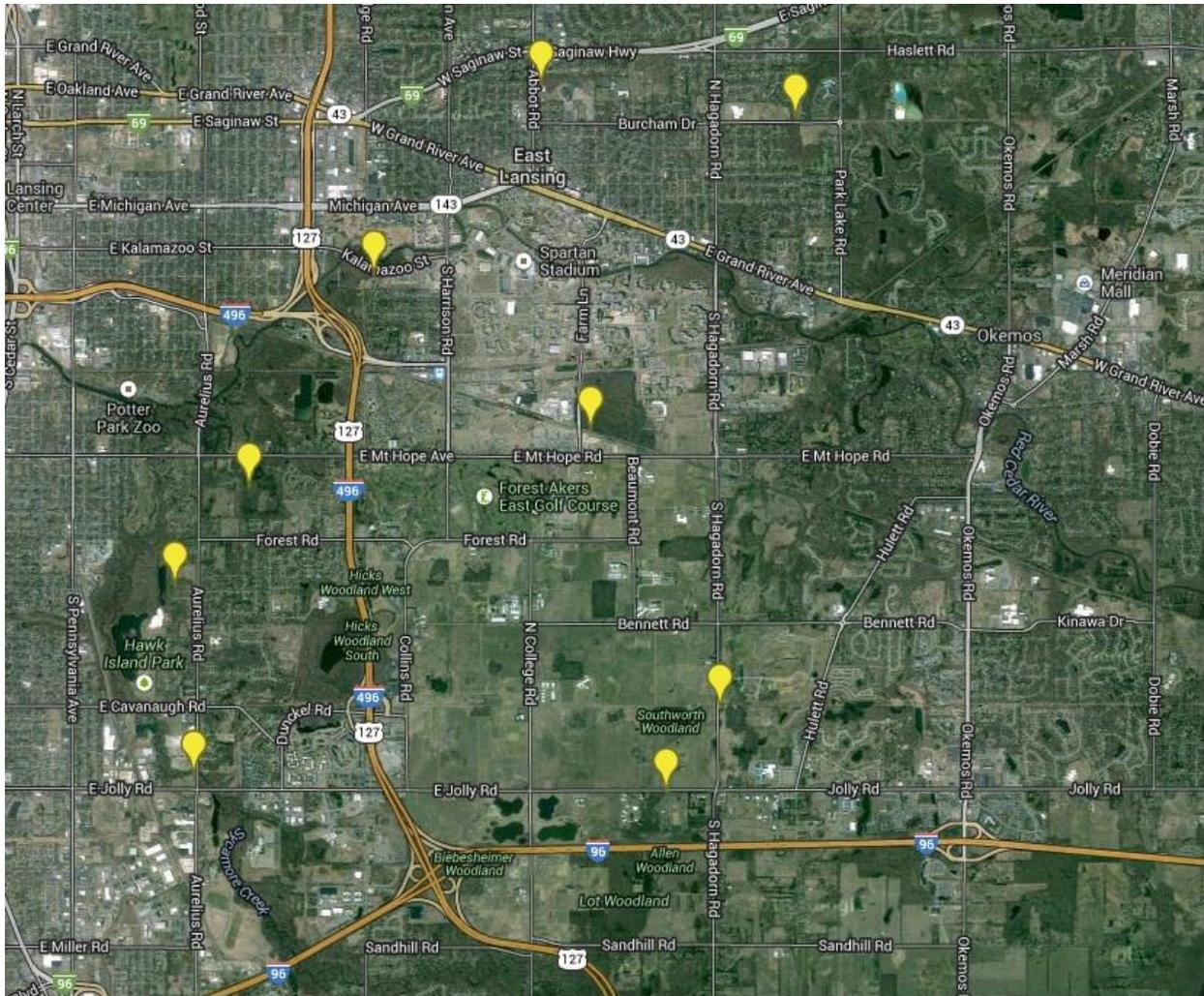


FIG. 10—Map of soil sampling locations for similar habitat collections. All sampling locations are within six miles of one another in the Greater Lansing area.

### *Soil Collection from Three Habitats over Time*

Soils on the surface of a deciduous woods, yard, and treated yard on the Michigan State University campus (treated with pesticides and fertilizer) were collected once a day for four days, once a week for two months, and once a month for the remainder of the year starting in August 2013.

### *Soil Collection from Three Habitats over Horizontal Space*

Surface soil samples were collected at three habitats, the Fenner Nature Center yard, the Fenner Nature Center deciduous woods, and the treated yard on the Michigan State University

campus in March 2014. At each site a main soil sample was collected as were four additional samples collected at 5, 10, 50, and 100 feet in each cardinal direction, resulting in 17 samples per location.

#### *Soil Collection from Three Habitats over Vertical Space*

Soil samples were collected at the Fenner Nature Center yard, the Fenner Nature Center deciduous woods, and a second treated yard on the Michigan State University campus using a soil corer and mud auger (AMS, Inc. American Falls, ID) that were rinsed with RO water between samplings. The corer was removed from the ground, and a tape measure was used to measure the various depths. Approximately 100 g of soil was removed from the corer at each depth and placed in a sampling bag. A surface soil sample was taken with additional samples collected at 1, 2, 5, 10, 20, and 60 inches below it in April 2014. In the treated yard, 25 inches was the maximum depth reached due to obstruction.

#### *Soil Collection from Evidentiary Items*

In a preliminary mock evidence study, surface soil collected from woodlot 1 was deposited on a shoe, shovel, shirt, and sock, which were placed in brown paper bags. Additional soil was collected from the site and transported to the laboratory to be placed on a tire. The items were stored at room temperature for one year. After six months, three soil samples were collected from different areas of each evidentiary item as well as one homogenized sample containing soil from several parts of the object. After the full year, one homogenized soil sample was collected from each item.

A second evidentiary study investigating the traceability of soil over shorter time periods, as well as bacterial population shift at different storage conditions, was also conducted. Eight, new white cotton t-shirts (Hanes®, Winston Salem, NC) were exposed to soil in a 2X2

feet area of woodlot 1 by rubbing the t-shirt against itself with dirt in-between. After soil exposure, the shirts were placed in numbered brown paper evidence bags. T-shirts 1 – 4 were stored in a room temperature incubator (24°C), while t-shirts 5 – 8 were stored in a laboratory refrigerator (4°C) for the duration of the study. On day zero and once a week for eight weeks, small (ca. 1 cm<sup>2</sup>) soil covered portions of each shirt were collected. Additional portions were collected once a month for 2 months following this eight week period. On day zero and every two weeks for eight weeks, soil was collected from the woodlot of origin (woodlot 1) and from the eight other woodlots described in the similar habitat study.

### **DNA Extraction**

Pipette tips and tubes were UV irradiated in a Spectrolinker XL-1500 UV Crosslinker (Spectronic Corporation, Lincoln, NE) for 5 min (~ 2.5 J/cm<sup>2</sup>). DNA was extracted from soil samples using a PowerSoil® DNA Isolation Kit (MoBio, Carlsbad, CA) with two minor modifications of the manufacturer's protocol: an additional wash was conducted after step 16 by adding 500 µL of 70% ethanol and centrifuging for 30 s at 10,000 x g, and DNA was eluted using 100 µL of solution C6 that had been heated to 55°C. Reagent blanks were processed with every extraction.

### **PCR Amplification of 16S rRNA Variable Regions 3 – 4**

Bacterial 16S rRNA gene variable regions V3 and V4 were amplified with a forward primer [357F (Kozich et al., 2013)] and one of 96 bar-coded reverse primers (806R) from the Caporaso et al. (2010) primer set, producing an amplicon of approximately 450bp. Fifteen microliter PCR reactions contained final concentrations of 1X AmpliTaq Gold buffer (Life Technologies, Carlsbad, CA), 2.5 mM MgCl<sub>2</sub>, 0.2 mM nucleotide triphosphates, 0.4 µg/µL bovine serum albumin, 1U AmpliTaq Gold (Life Technologies), 1 µL of template DNA, and 1

$\mu\text{L}$  of 10  $\mu\text{M}$  forward and reverse primers. DNAs were denatured on an Applied Biosystems® 2720 thermal cycler (Life Technologies) for 10 min at 94°C, followed by 35 cycles of 94°C for 30 s, 60°C for 45 s, and 72°C for 60 s, and a final extension of 10 min at 70°C. Four microliters of the PCR product were electrophoresed on a 1% agarose gel followed by ethidium bromide staining and UV visualization.

### **DNA Quantification and Equimolar Pooling**

The PCR products were quantified using a Quant-iT™ PicoGreen® dsDNA Assay Kit (Life Technologies) following the manufacturer's protocol, and pooled so that each bacterial sample was at an equal concentration (~6 ng/ $\mu\text{L}$ ). DNAs were purified using Agencourt® AMPure® XP (Beckman Coulter, Brea, CA) beads. The bottle containing the beads was vortexed briefly and the beads were added in a 0.6:1 ratio to the pooled DNAs. This mixture was vortexed and incubated at room temperature for 15 min. The beads were bound to a MagnaRack™ (Life Technologies) for a minimum of 5 min. The supernatant was aspirated and discarded. Undisturbed beads were washed with 500  $\mu\text{L}$  70% ethanol for 30 s. The supernatant was again aspirated and the beads were washed an additional time. Beads were dried on the magnet for 30 min at 37°C. DNA was eluted by adding 100  $\mu\text{L}$  of 10 mM Tris, pH 8 and vortexing the tubes for 10 s. The tubes were returned to the magnet and beads were bound for at least five min. Supernatant was aspirated away and transferred to a 1.5 mL micro centrifuge tube.

### **Sequencing Purified PCR Products**

The pooled amplified bacterial DNAs were sequenced on an Illumina MiSeq (Illumina, San Diego, CA) following the manufacturer's protocols using a paired end 250bp v2 Reagent Kit (Illumina). Base calling was performed with Real Time Analysis software v1.18.54 (Illumina),

and the output was demultiplexed and converted to FastQ files with Bcl2fastq Conversion Software v1.8.4 (Illumina).

### **Gene Sequence Data Pretreatment**

Sequencing data were processed using open-source mothur software following standard operating procedures on the mothur webpage (Schloss et al., 2009; [www.mothur.org](http://www.mothur.org); [www.mothur.org/wiki/MiSeq\\_SOP](http://www.mothur.org/wiki/MiSeq_SOP)). MiSeq libraries were subsampled to 3000 sequences per soil sample<sup>1</sup>. Sequences were then organized into operational taxonomic units (OTUs) at a 97% similarity cutoff, and compared to the SILVA database.

### **Next-Generation Sequencing Analysis Procedures**

OTUs were used to calculate Sørensen-Dice coefficients within mothur for all soil samples, and the resulting square symmetric matrices were used as the input for NMDS, which was run in XLSTAT Pro (Addinsoft, New York, NY), and *k*-NN, which was run in Pirouette 4.0 (Infometrix, Inc. ©, Bothell, WA). For *k*-NN analysis of diverse and similar habitat soil samples, the jackknife resampling method (Tukey, 1958) was employed, testing each sample against the other four collected from these sites. Training and test sets for *k*-NN analysis are described in Table 3. Soil bacterial sequences were also classified using the SILVA bacterial reference alignment (Quast et al., 2013), and abundance charts were created at the taxonomic class level in Excel (Microsoft, Redmond, WA).

---

<sup>1</sup> Subsampling is a necessary step in sequence processing due to the computational limits when handling massive amounts of data such as those produced via next-generation sequencing. To examine the effects of subsampling on sequence libraries, the diverse habitat soil samples (n=49) were subsampled down to 3000 sequences four times and analyzed to assess congruity. All subsampling events resulted in similar measures of dissimilarity and orientation in NMDS plots, exemplifying that subsampling had little biasing effect on next-generation sequencing results.

TABLE 3—*Training and test sets for k-NN analysis of soil samples.*

Study	Training Set	Test Set
Diverse Habitats*	N=4 per habitat	N=1
Similar Habitats*	N=4 per woodlot	N=1
	Center, 5ft N, 5ft S, 5ft W, 5ft	All other distance soil samples
Within Habitat Horizontal Space	Center point plus 5ft E, 10ft N, 50ft W, and 100ft S	All other distance soil samples
	Center 100ft N, 100ft S, 100ft W, 100ft E	All other distance soil samples
Within Habitat Vertical Space	Surface, 2 in, 10 in, and 60 in	All other depth soil samples
Preliminary Evidentiary	Woodlot samples collected over eight wk period	Various evidentiary soil samples after six months and one year
T-shirt Evidentiary	Woodlot samples collected over eight wk period	T-shirt evidentiary soil samples over four months

\*Analyzed via the Jackknife resampling method (Tukey, 1948) in which each of the five soil samples was systematically left out and tested against the other four samples.

## RESULTS

### Illumina MiSeq Sequencing Efficiency

MiSeq sequencing resulted in large data sets, producing approximately 150,000 sequence reads per sample. The processing of sequence libraries in mothur culminated in the removal of 94 – 97% of total sequences, most often during the subsampling portion of sequence analysis. Only the dirt road sample from February 2014 did not produce the requisite number of sequences and was excluded from analysis.

### General Results

Each soil sample contained approximately 50 bacterial taxonomic classes (range: 42 – 58), with the exception of samples collected from the dirt road, which had reduced diversity (range: 24 – 29). Aside from the dirt road samples, the bacterial profiles contained the same major bacterial classes, making up a large part of their abundance; differences among habitats tended to be in the least abundant classes<sup>2</sup>.

The Scree diagrams for all NMDS plots had high stress in one dimension with a decrease (elbow) at two and more dimensions (e.g. FIG. 11). Shepard diagrams for NMDS exhibited close association of distances and disparities, affirming the low stress at two dimensions (e.g. FIG. 12).

---

<sup>2</sup> Note that in the abundance charts that follow, bacterial classes are graphed in order of abundance for that sample set, thus class order can differ among charts.

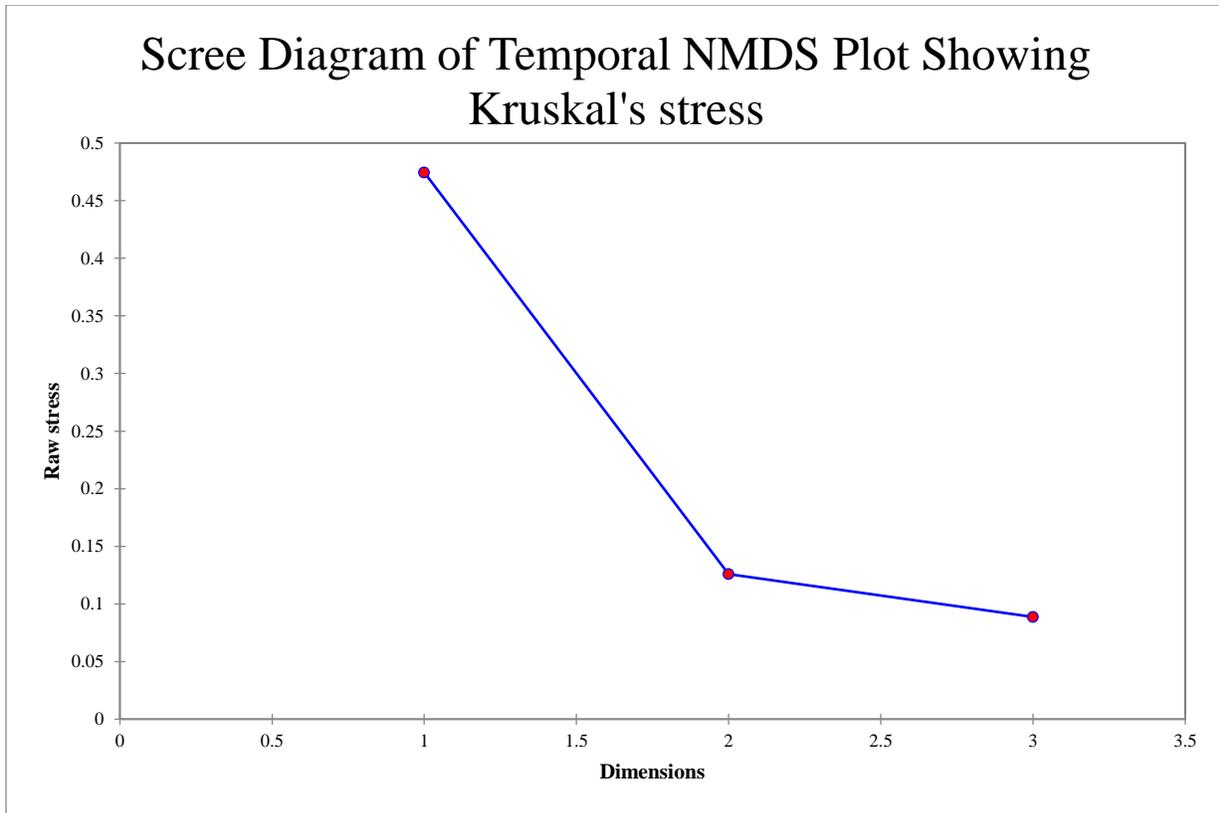


FIG. 11—Scree diagram generated with NMDS plot of temporal samples showing the characteristic elbow signifying a substantial decrease in stress from one to two dimensions, and a general leveling off with additional dimensions. All scree diagrams in this research looked similar.

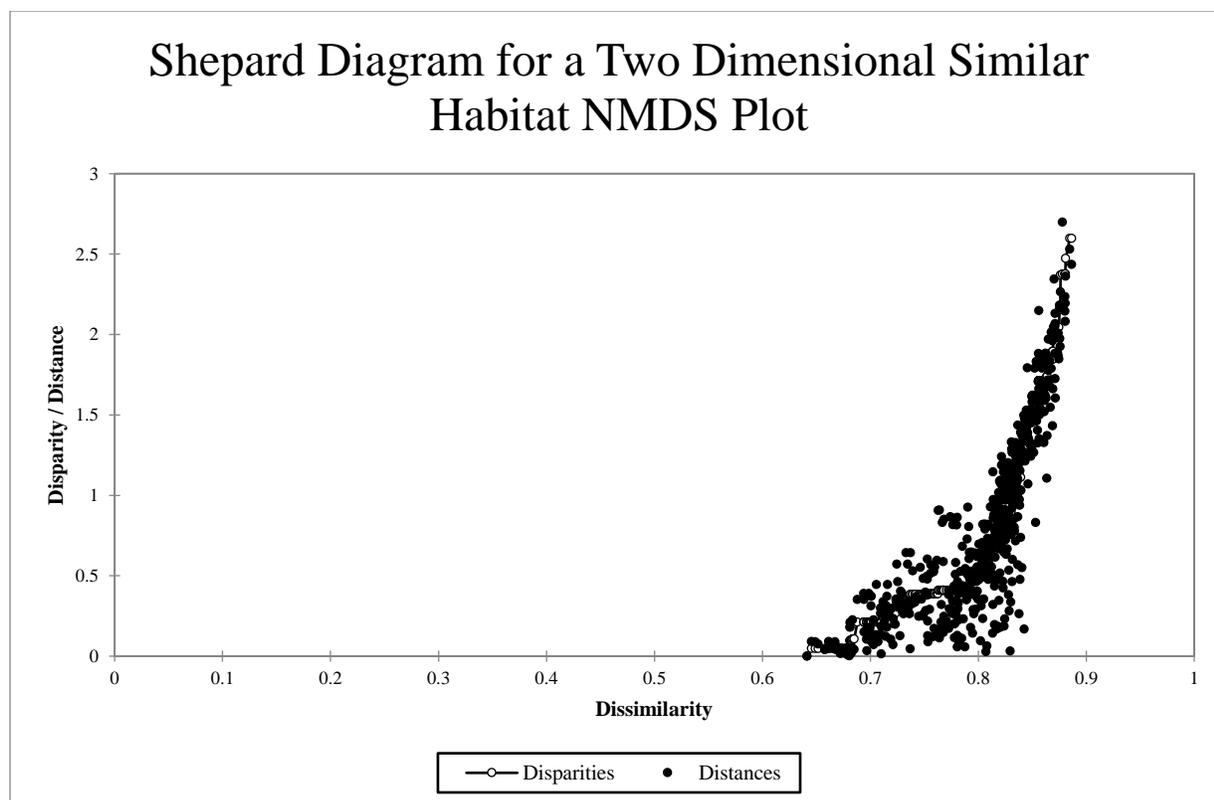


FIG. 12—Shepard diagram generated with similar habitat NMDS plot. All distances fall close to corresponding disparities, indicating good correlation of the two in the final configuration. All Shepard diagrams in this research looked similar.

In general, NMDS plots produced clusters of soil samples collected from the same location in all studies. Intermingling of clusters was common when many unrelated samples were ordinated together, however, these clusters were resolved when the locations containing overlapping members were ordinated in pairs or triads.

*k*-NN analysis resulted in the accurate assignment of soils back to their point of origin for nearly all samples. Table 4 summarizes the *k*-NN results, outlining the training sets used and any soil sample misclassifications. In some cases, soil classification accuracy changed depending on the training set used, but this did not result from simply reducing the number of training set members, as five were always used (with the exception of one vertical space training set, made up of only four samples).

TABLE 4— *k*-NN summary findings. Multiple training sets were examined for all studies with the exception of the evidentiary studies.

Study	Training Set	Accuracy	Misassigned Samples
Diverse Habitats	All Habitat Samples*	88%	Marsh with Fallow Ag <sup>†</sup> Field and Deciduous Woodlot with Yard
	Marsh and Fallow Ag Field*	100%	-
	Deciduous Woodlot and Yard*	100%	-
Similar Habitats	All Location Samples*	87.5%	All samples from Woodlot 8 and one sample from Woodlot 9
	Woodlots 1 – 7	100%	-
Horizontal Space	Center, 5ft N, 5ft S, 5ft W, 5ft E	94.4%	Yard 100ft N, Woods 100ft S
	Center, 5ft E, 10ft N, 50ft W, and 100ft S	97.2%	Deciduous Woods 100ft N
	Center, 100ft N, 100ft S, 100ft W, 100ft E	94.4%	Yard 5ft E, Yard 10ft N
Vertical Space	Surface, 2 in, 10 in, 60 in	100%	-
	Surface, 1 in, 2 in, 5 in, 10 in	83%	Deciduous Woods 60”
Preliminary Evidentiary	Woodlots 1 – 9	100%	-
Short-term Evidentiary	Woodlots 1 – 9	100%	-

\* Analyzed via Jackknife Method (Tukey, 1948)

<sup>†</sup>Ag=Agricultural

## Analysis of Soils from Diverse Habitats

### Bacterial Abundance Charts

Diverse habitat profiles (FIG. 13) appeared fairly similar, with the exception of the dirt road, which had substantially lower levels of *Acidobacteria* and *Betaproteobacteria*, and higher levels of *Flavobacteria*, *Gamaproteobacteria*, *Clostridia*, and *Bacilli* relative to the other habitats.

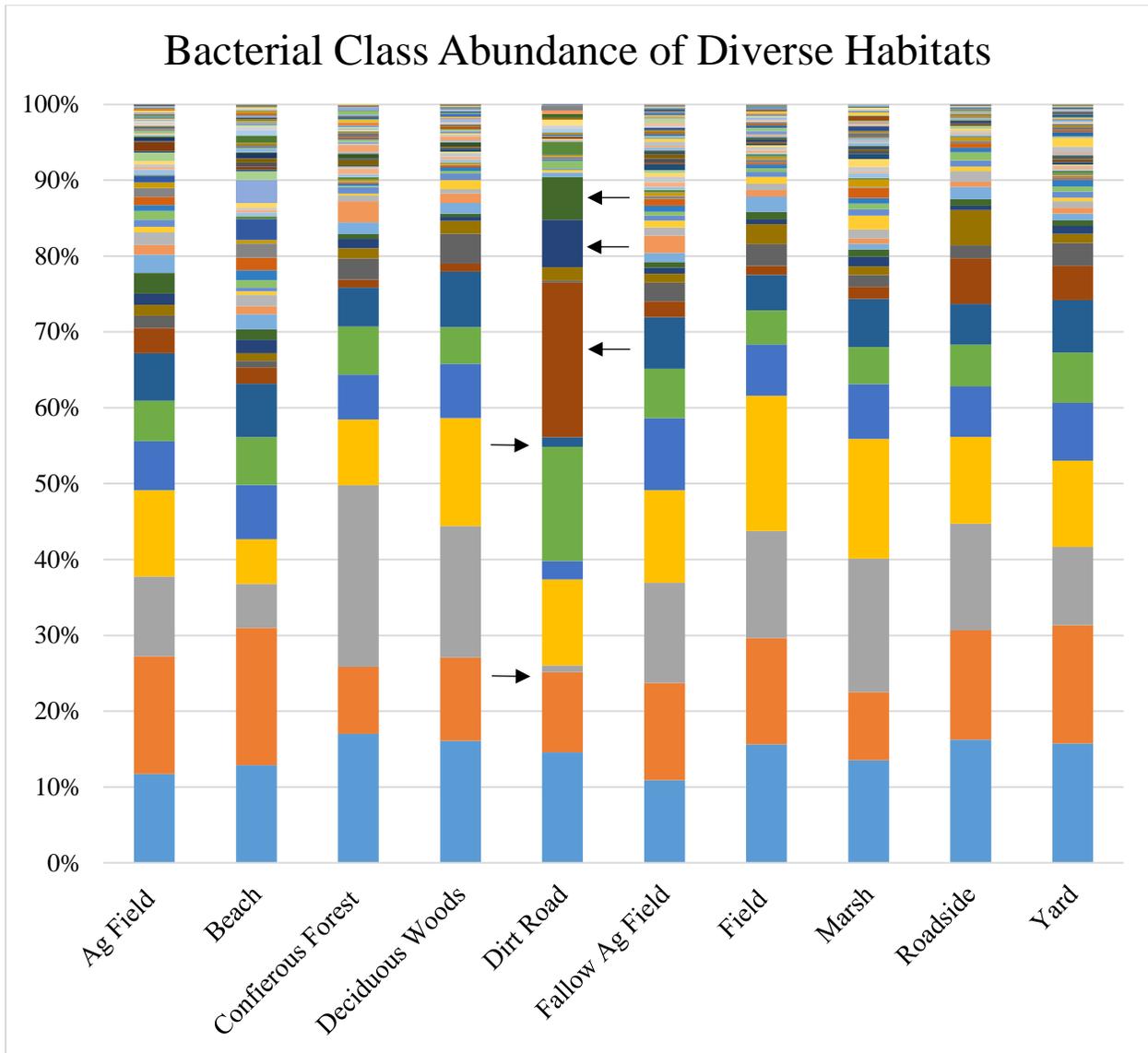


FIG. 13—Average (n=5) bacterial class abundance of five replicates from ten diverse habitats. The dirt road clearly differed from the other habitats, containing higher levels of *Flavobacteria*,

*Clostridia*, and *Bacilli* (designated by arrows in ascending order on the right), along with lower levels of *Acidobacteria* and *Betaproteobacteria* (designated by arrows in ascending order on the left). Ag=Agricultural.

*Nonmetric Multidimensional Scaling*

Samples collected within a habitat clustered together in NMDS plots (FIG. 14), but some intermingling occurred among the ten habitats. When three habitats were oriented at a time, the clusters separated in all cases (e.g., FIG. 15).

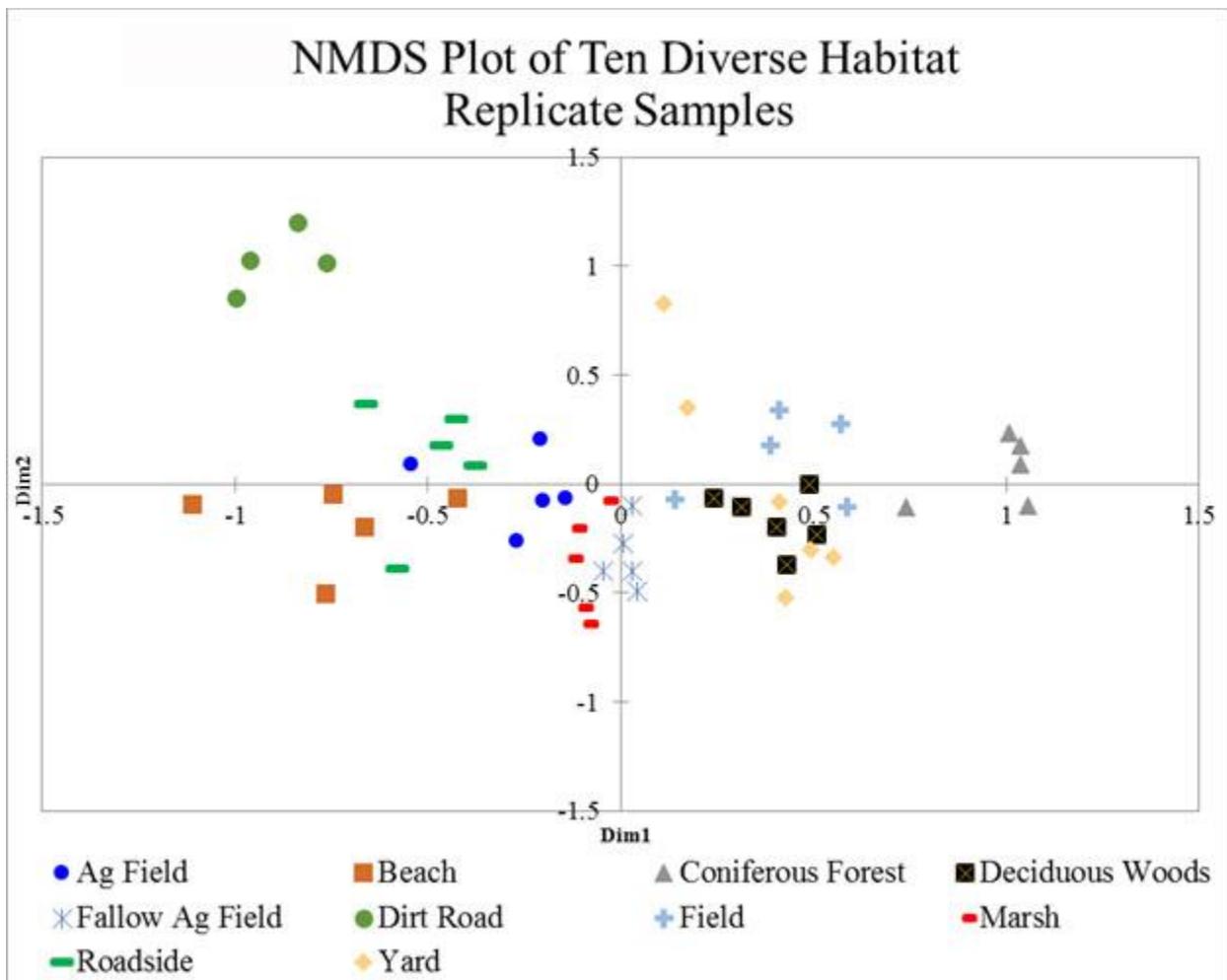


FIG. 14—NMDS plot ordinating samples collected from the ten diverse habitats. Replicate samples from the same habitat formed clusters, but intermingling occurred among some of the habitats. Ag=Agricultural.

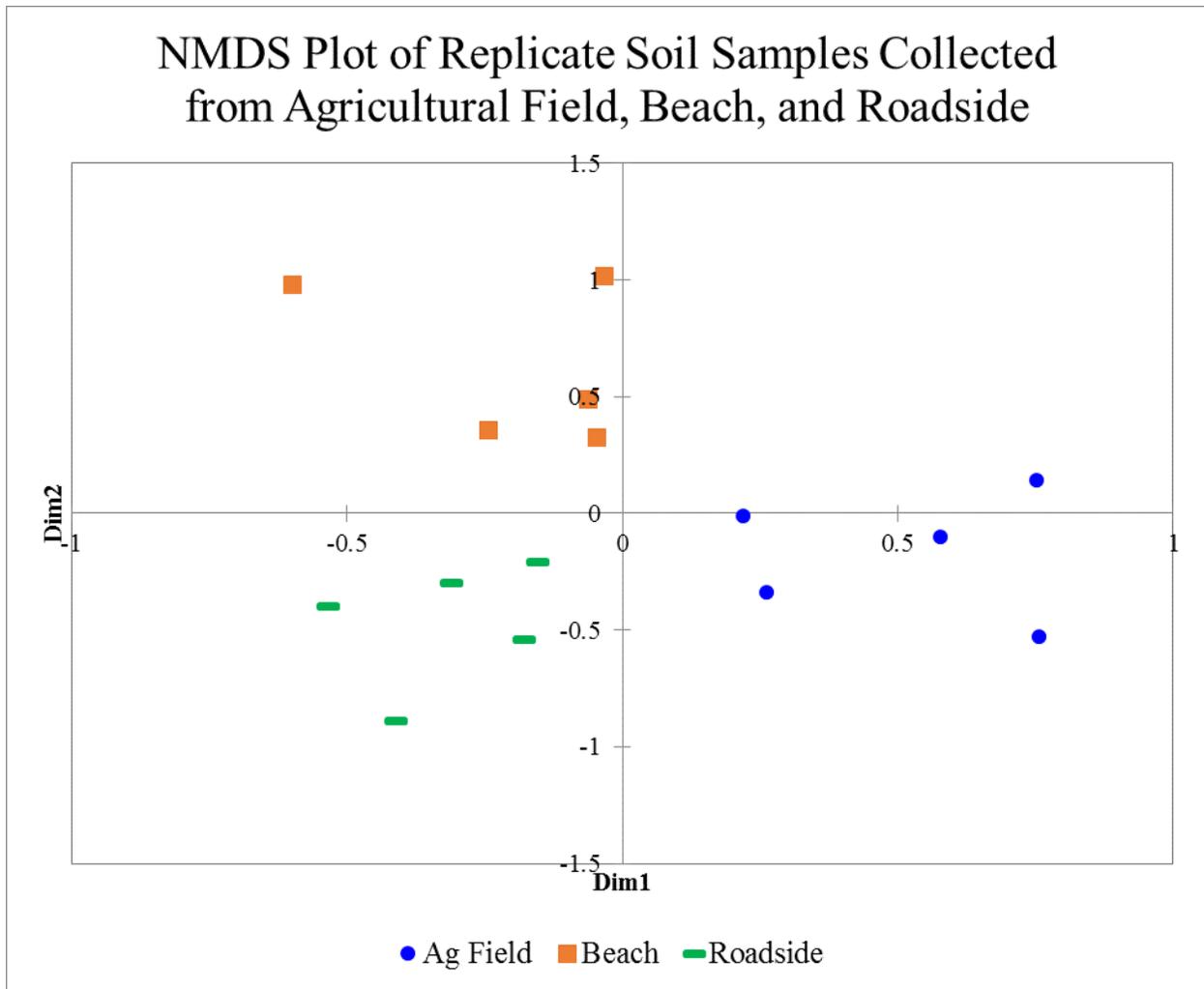


FIG. 15—NMDS plot ordinating soil samples collected from the agricultural (Ag) field, beach and roadside. Samples from these locations intermingled when all habitats were ordinated together, but were resolved when analyzed as pairs or triads in NMDS plots.

*k-Nearest Neighbor*

*k*-NN exhibited an 88% assignment accuracy when all habitats were analyzed together (Table 4). Misclassifications occurred between the marsh and fallow agricultural field soil samples and between the deciduous woods and yard samples. These were all correctly assigned to their habitat when analyzed as pairs in a *k*-NN model.

## Analysis of Soils from Similar Habitats

### *Bacterial Abundance Charts*

The nine woodlot samples abundance charts (FIG. 16) appeared more similar than did the diverse habitats (FIG. 13). However, diversity differed among woodlots, ranging from 40 – 59 bacterial classes per soil sample.

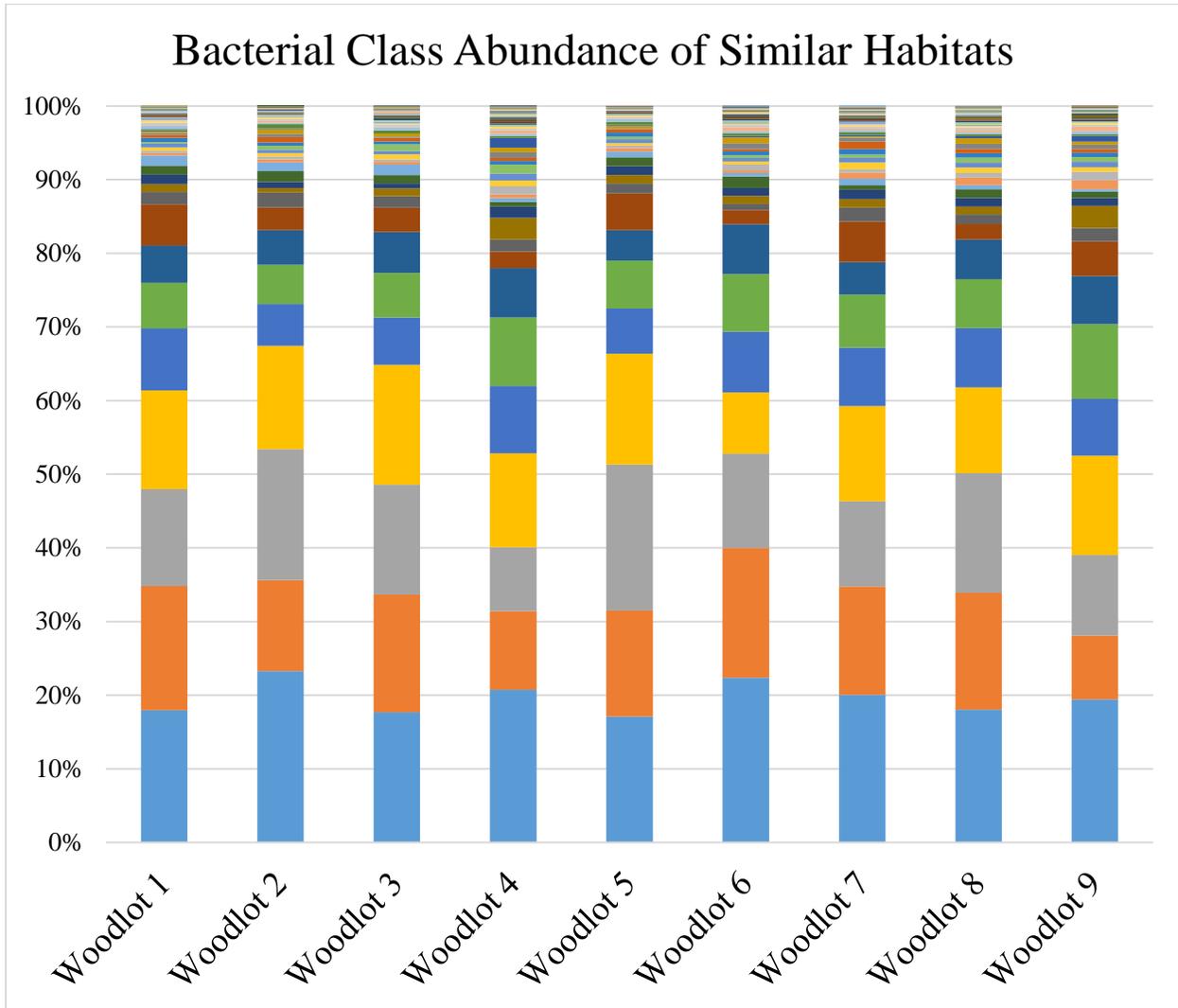


FIG. 16—Average (n=5) bacterial class abundance of woodlot locations. The soils appeared very similar, sharing the majority of bacterial classes.

*Nonmetric Multidimensional Scaling*

Samples collected from the same woodlot clustered together in NMDS plots, but intermingling occurred among several of the clusters (FIG. 17). The most substantial overlap involved woodlot 8, whose samples were interspersed with several other clusters. By ordinating samples in pairs or triads, separation of woodlots occurred in all cases (e.g. FIG. 18).

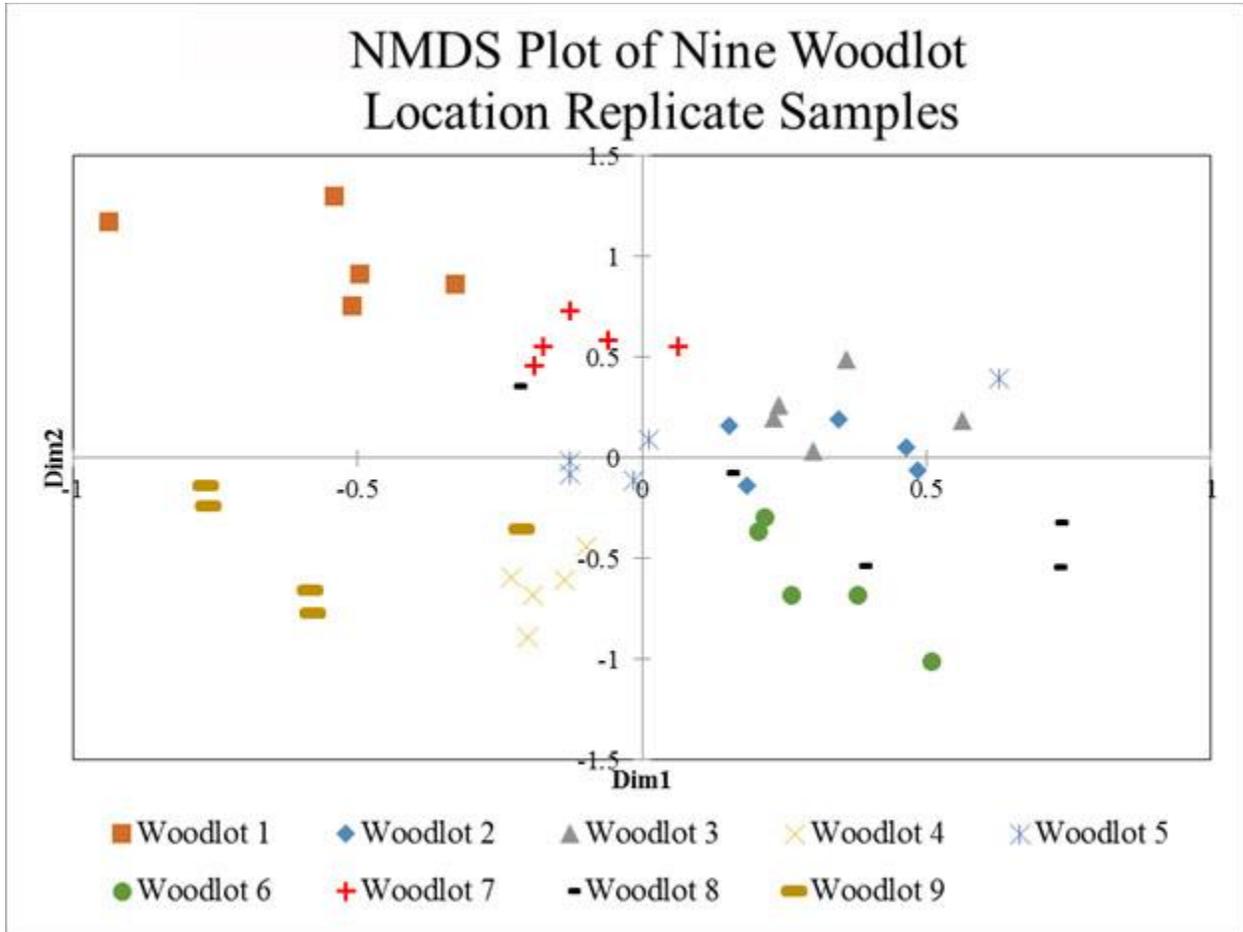


FIG. 17—NMDS plot ordinating soil samples collected from the nine woodlots. Replicate samples from the same location formed clusters, but intermingling occurred among some of the location clusters. Woodlot 8 replicate samples clustered relatively poorly, intermingling with several other woodlot samples.

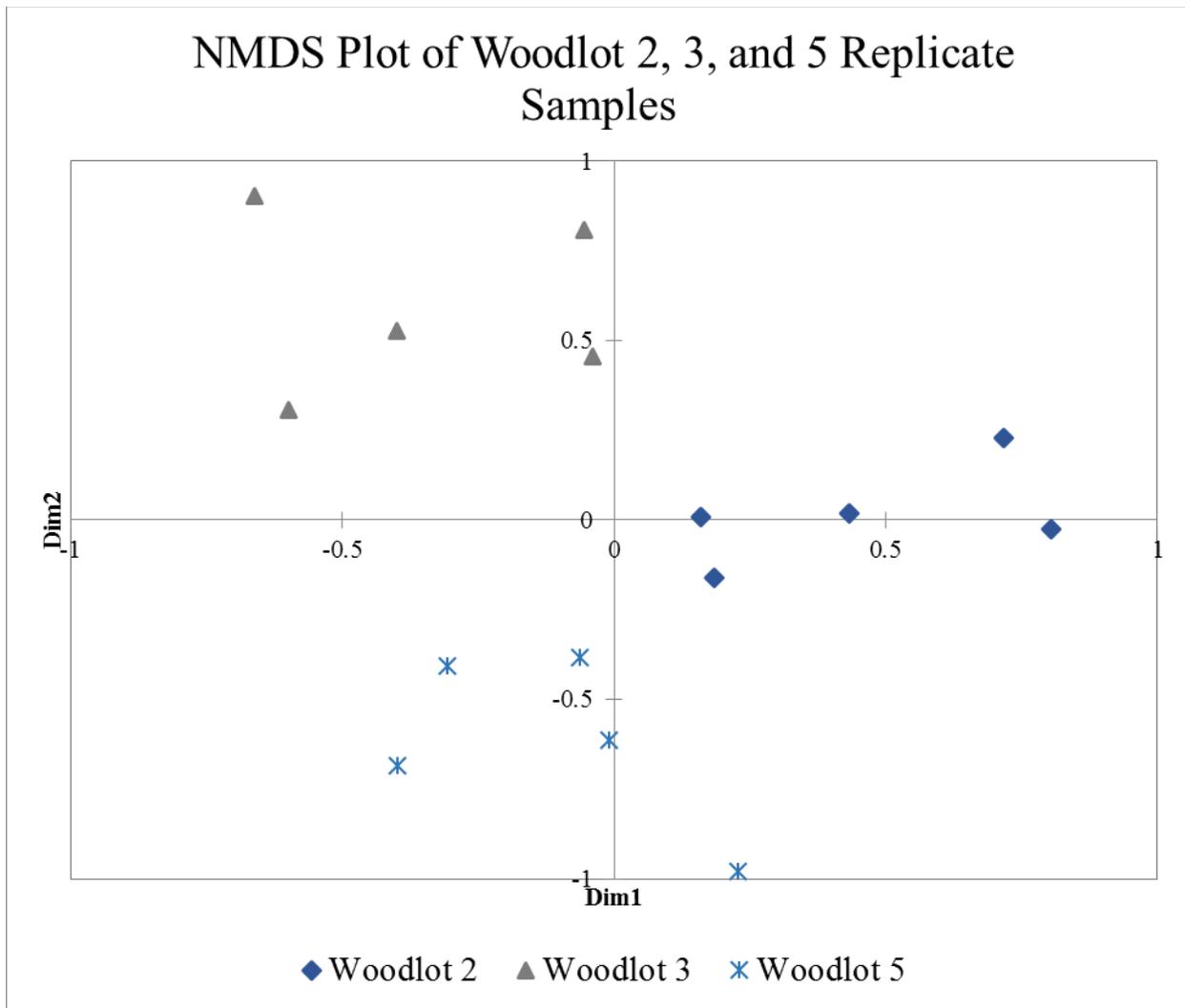


FIG. 18—NMDS plot of the replicate woodlot 2, 3, and 5 soil samples. These samples were intermingled when all woodlots were ordinated together (FIG. 17), but were resolved when they were analyzed alone in a NMDS plot.

*k-Nearest Neighbor*

k-NN was accurate in its assignment of the woodlot location samples 87.5% of the time (Table 4). All samples from woodlot 8 and one sample from woodlot 9 were incorrectly assigned, and when removed from the model, 100% assignment accuracy was achieved.

## Analysis of Soils from Temporal Study

### Bacterial Abundance Charts

Samples collected over time within each habitat appeared very similar in abundance charts (e.g. FIG 19). Fluctuations of bacterial class abundance over time in each habitat were evident, but not extreme.

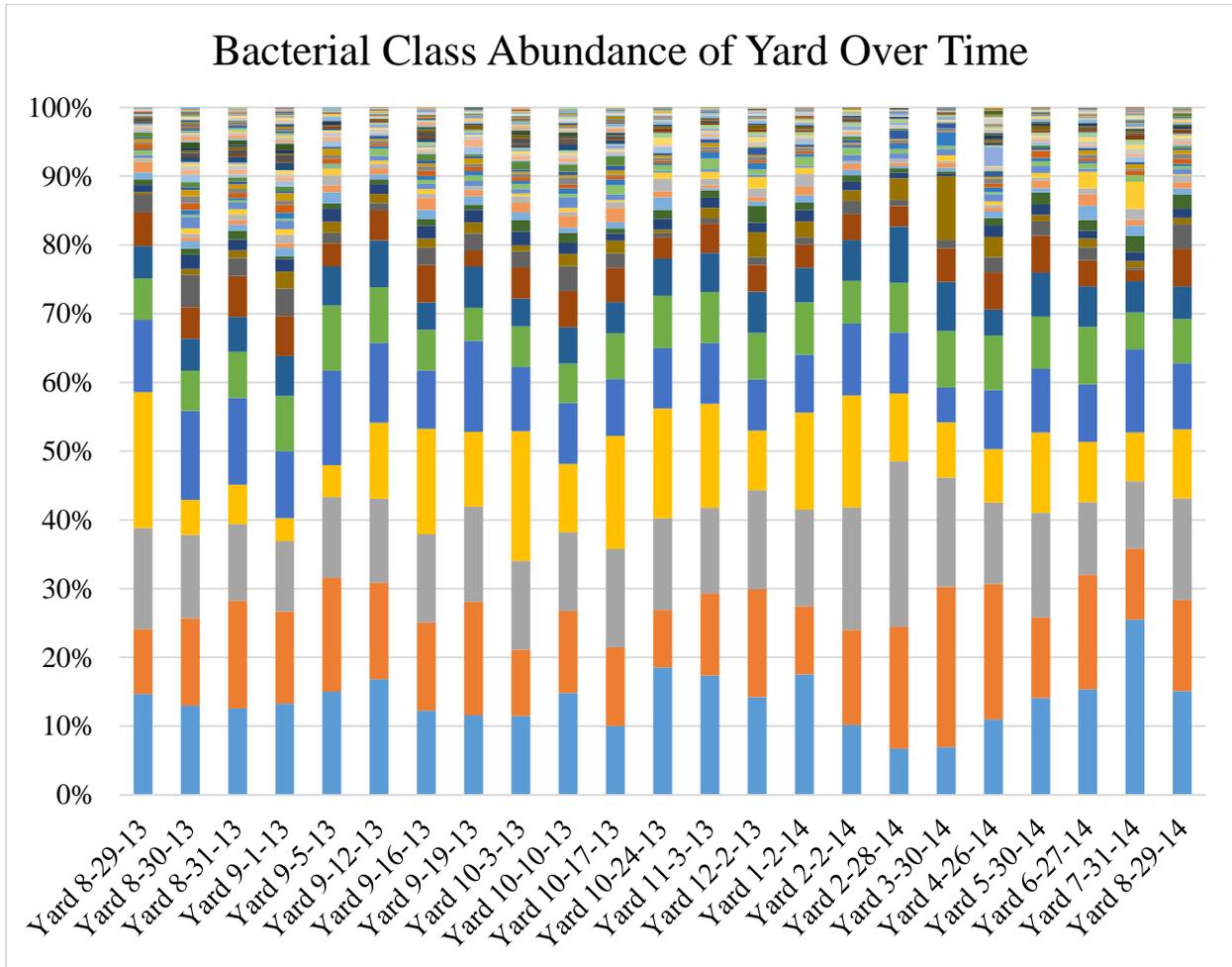


FIG. 19—Yard bacterial class abundance over one year (left to right). Slight fluctuations in bacterial class abundance were evident, but samples shared major bacterial classes throughout the year.

*Nonmetric Multidimensional Scaling*

Samples formed clusters based on habitat of origin in multidimensional space (FIG. 20). February and March samples collected from the deciduous woods and yard fell the furthest from the main habitat clusters.

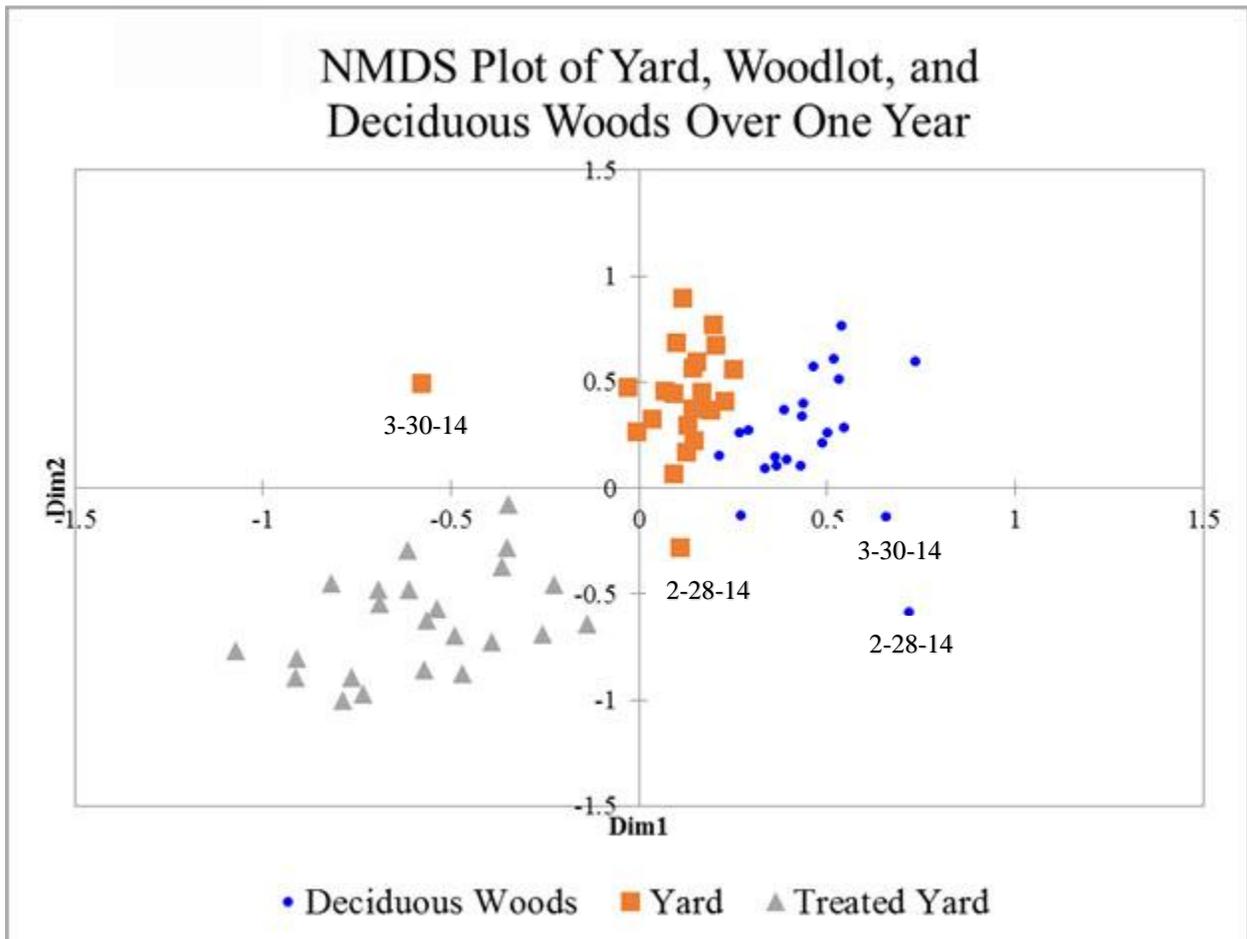


FIG. 20—NMDS plot of temporal samples from three habitats. Samples from each habitat formed clusters, however, samples collected in February and March fell the furthest away for the yard and deciduous woods (labeled below point with date of collection).

*k-Nearest Neighbor*

*k*-NN was accurate in its assignment of 92.3% of soil samples back to their site of origin over the full year (Table 4). Soil samples collected from the deciduous woods and yard in

February and the yard sample collected in March were the only samples misassigned, both classifying to the treated yard.

## **Analysis of Soils from Spatial Studies**

### *Bacterial Abundance Charts*

Surface soil bacterial profiles within and between habitats contained similar bacteria, again sharing the majority of taxonomic classes (e.g. FIG. 21). Fluctuations of specific bacterial class abundance across the surface of a habitat were evident, but profile difference was not extreme.

## Bacterial Abundance Chart of Deciduous Woods Samples Collected at Short Horizontal Distances

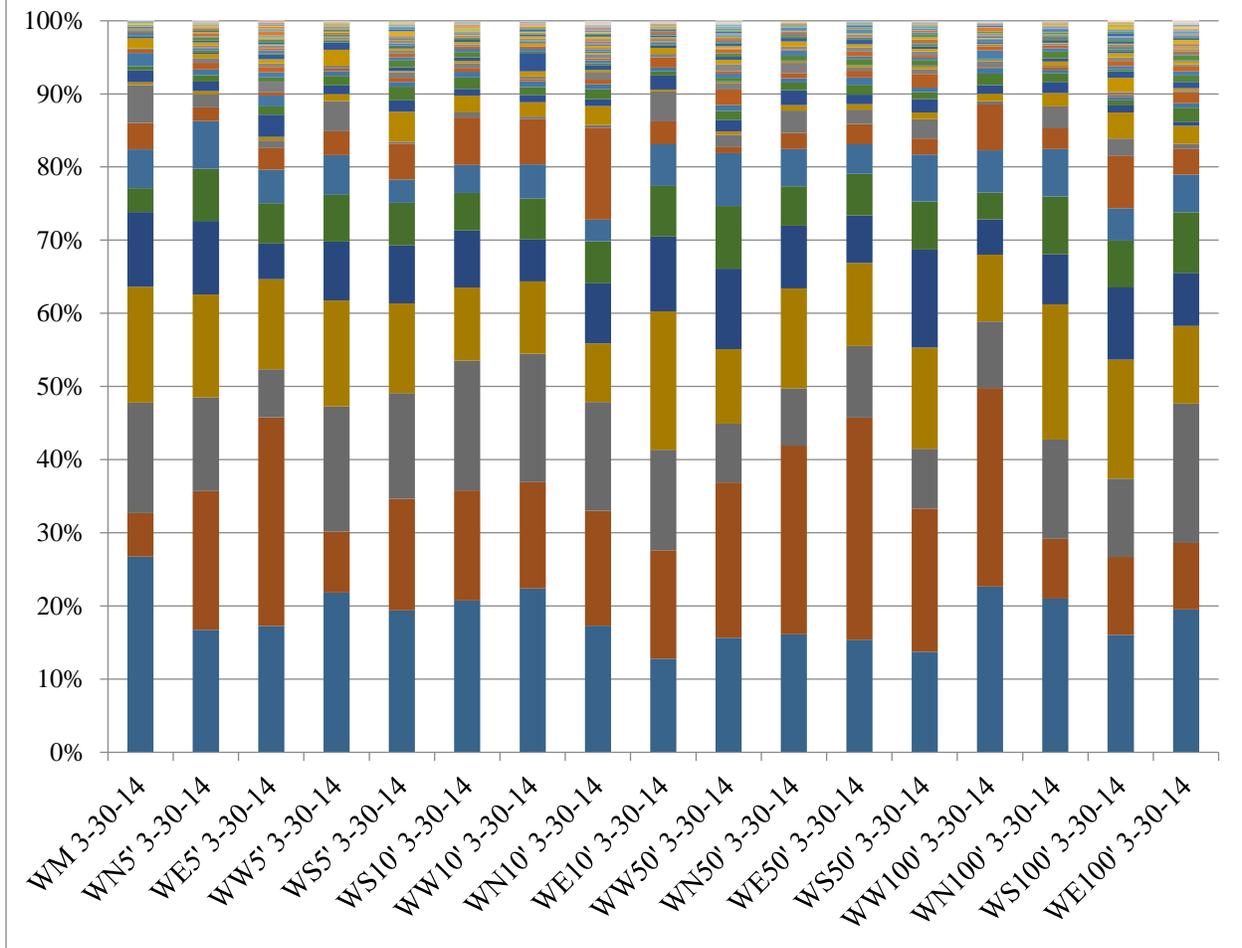


FIG. 21—Deciduous woods bacterial class abundance of samples collected on the surface, at a center point and 5, 10, 50, and 100 feet in the cardinal directions. Samples shared the majority of bacterial classes, but fluctuation of some classes was evident.

Abundance charts generated from the depth samples revealed class differences as deeper samples were tested (e.g. FIG. 22), although diversity remained constant. The most substantial class abundance differences in all habitats were higher amounts of *Clostridia* and *Nitrospira* as depth increased.

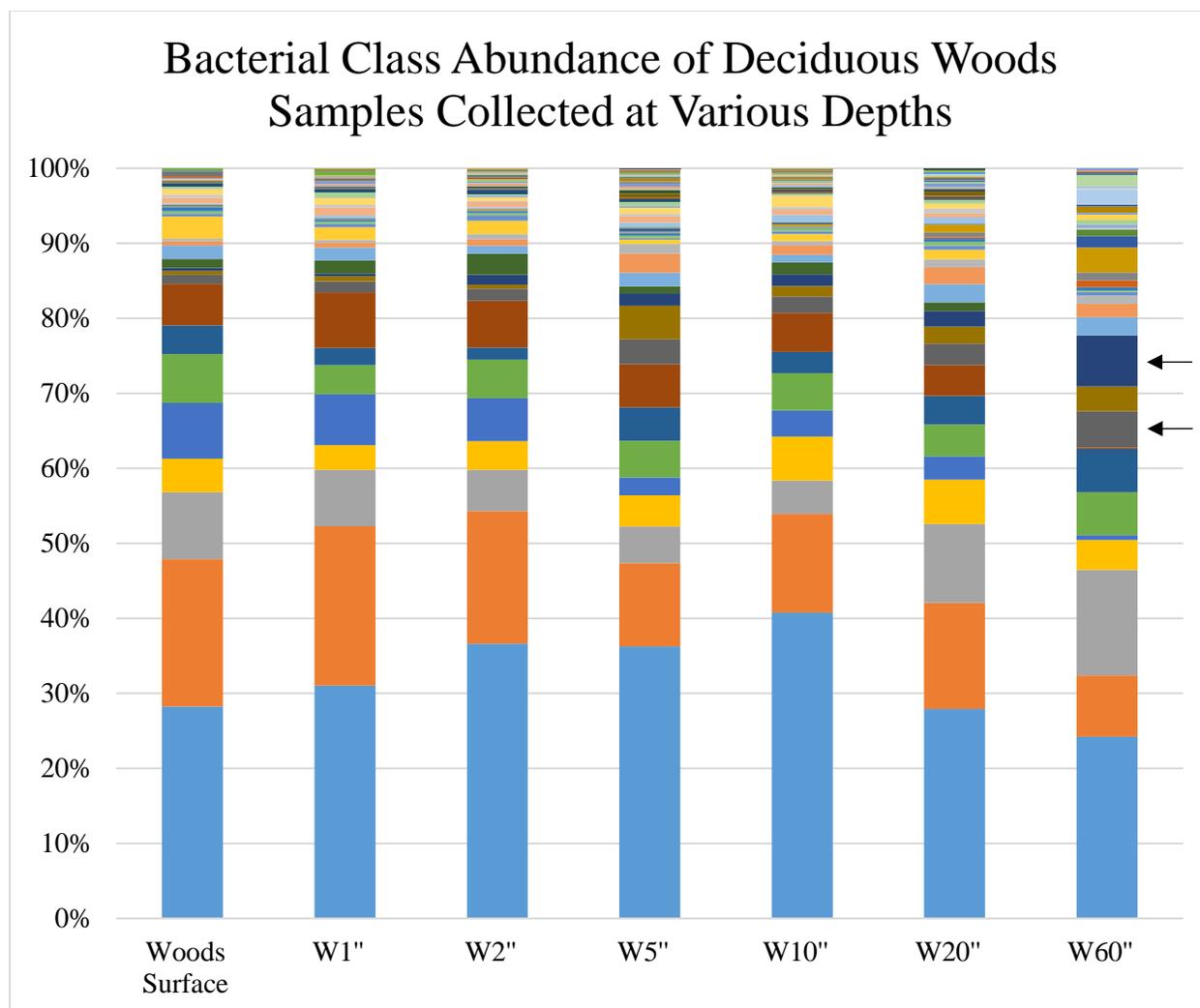


FIG. 22—Bacterial class abundance of woodlot depth samples. As depth increased, substantial differences in *Clostridia* and *Nitrospira* (denoted by arrows in ascending order) existed in all habitats.

### *Nonmetric Multidimensional Scaling*

Samples collected from the same habitat loosely clustered together in NMDS plots (FIG. 23), with profiles obtained from 50 and 100 feet from the center sampling site generally being the furthest from the middle of the clusters. The treated yard cluster was completely separated from the deciduous woods and yard clusters, but the latter two intermingled slightly.

The treated yard depth samples clustered separately in NMDS plots (FIG. 24), while the deciduous woods depth and yard depth samples intermingled. When deciduous woods and yard

soil samples were ordinated together (FIG. 25) clearer separation of habitats occurred. Within all habitats, a trend existed where the soil bacterial profiles moved farther away from the surface sample in multidimensional space as depth increased.

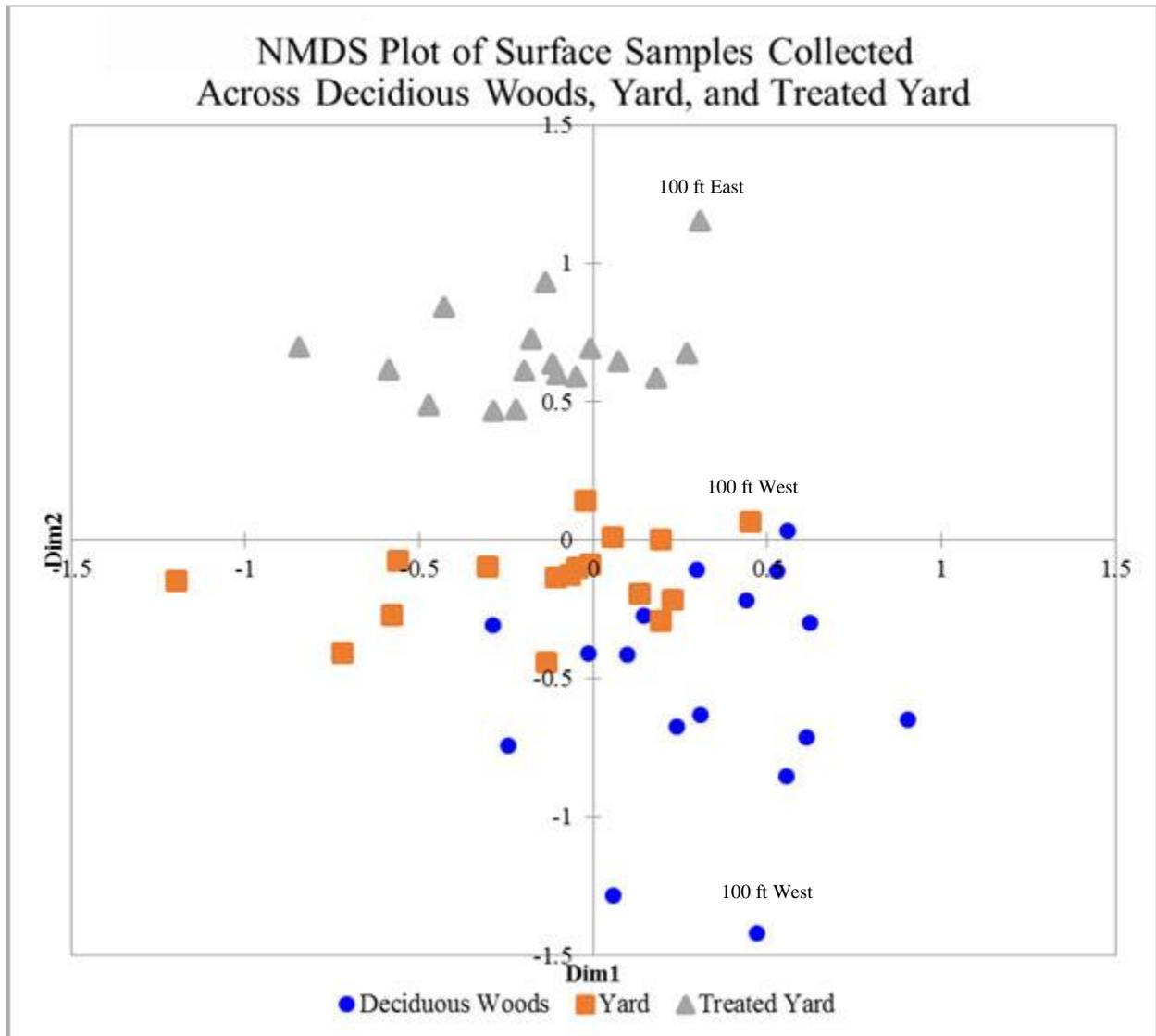


FIG. 23— NMDS plot of spatial samples collected from three habitats. Samples from each habitat formed clusters, but the deciduous woods and yard samples intermingled. Soils collected the furthest from the center sampling site, plotted further away in multidimensional space (one 100 foot sample for each habitat is labeled above the corresponding point).

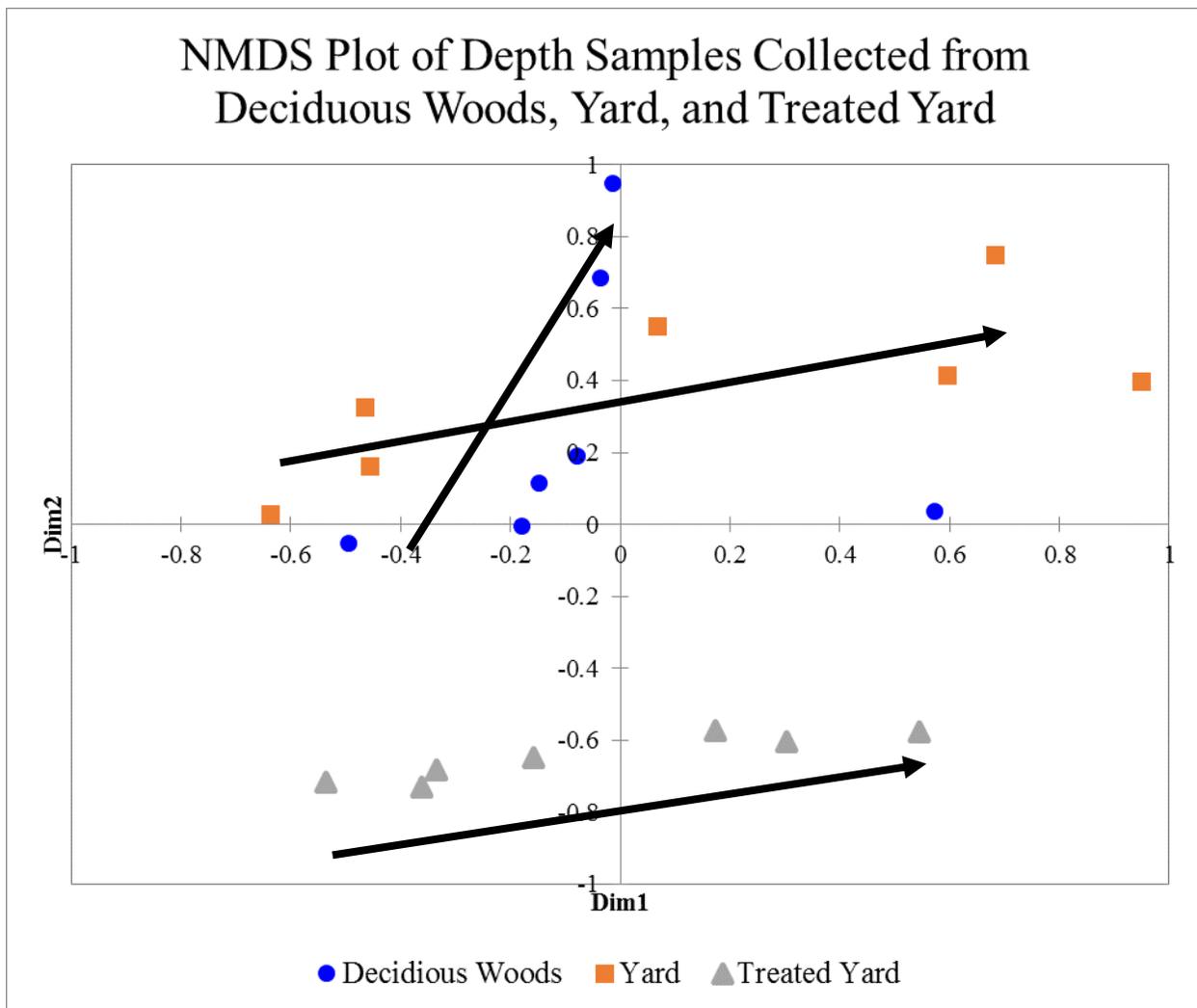


FIG. 24—Ordination of soil samples collected at various depths in three habitats. The treated yard samples clustered separately, while the deciduous woods and yard samples intermingled. A trend existed across all habitats where the soil bacterial profiles moved away from the surface sample in multidimensional space as depth increased (arrows point in the direction of increasing depth).

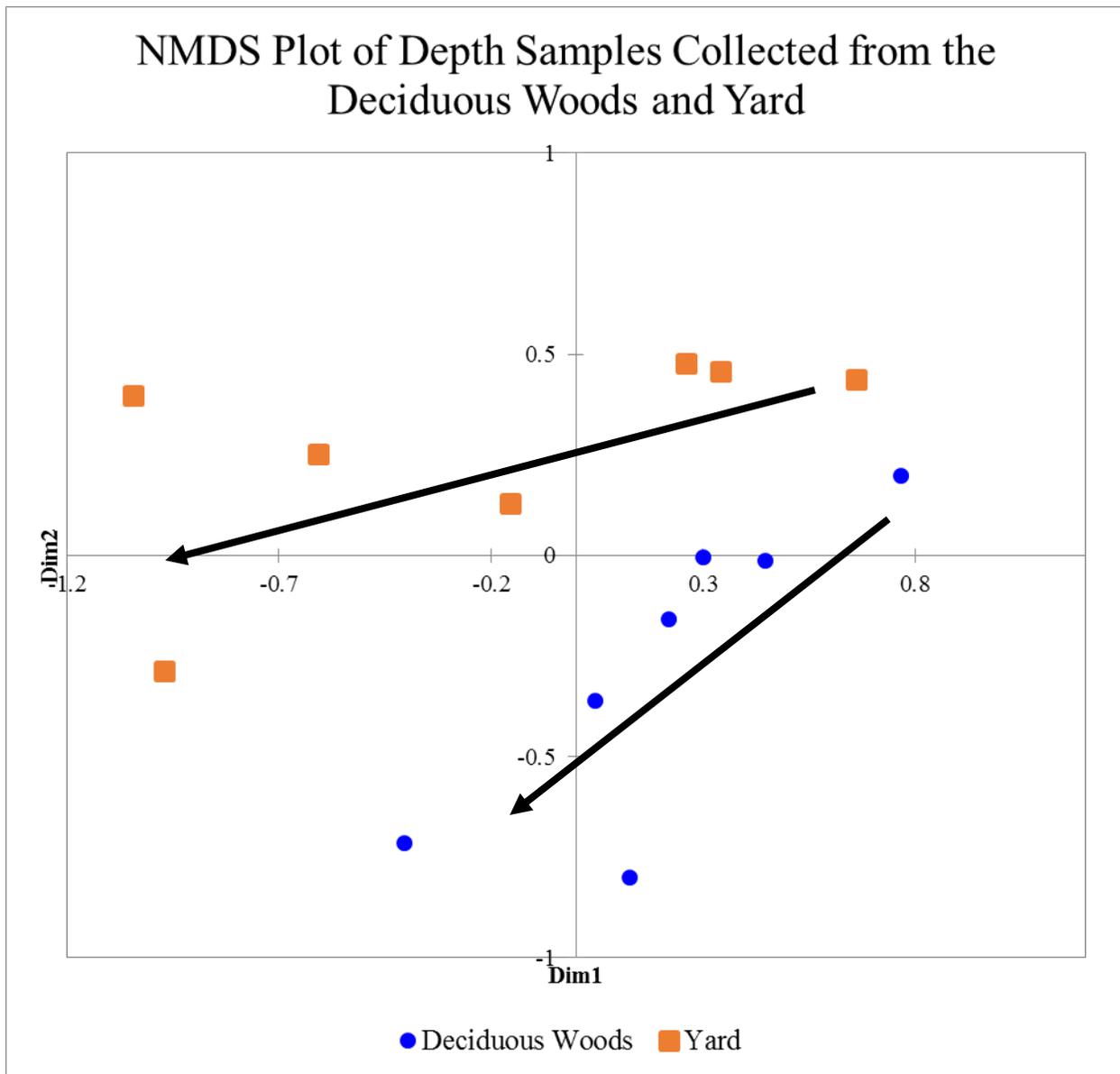


FIG. 25—Oridnation of deciduous woods and yard depth samples. Although intermingled when plotted with the treated yard samples, deciduous woods and yard samples separated when ordinated as a pair. Again, samples reflected the trend where soil bacterial profiles moved away from the surface sample in multidimensional space as depth increased (arrows point in the direction of increasing depth<sup>3</sup>).

<sup>3</sup> General orientation of samples in NMDS plots is random, thus, the different direction of arrows across plots is not analytically relevant.

### *k*-Nearest Neighbor

*k*-NN analysis of the soil samples collected across the habitat surface accurately assigned samples 94.4 – 97.2% (Table 4), depending on the samples used for the training set (see Materials and Methods). The most accurate categorization occurred when using the center plus one sample each from 5, 10, 50, and 100 feet distances as the training set.

Depth soil samples were accurately assigned 83% of the time when the shallowest five samples made up the training set (Table 4). The only misclassification was the 60 inch deciduous woods sample. Soil samples were accurately assigned 100% of the time with the surface, 2, 10, and 60 inch samples making up the training set for *k*-NN analysis (Table 4).

## **Analysis of Preliminary Evidentiary Samples**

### *Bacterial Abundance Charts*

Bacterial profiles collected from the various evidence types displayed abundance changes over time (e.g. FIG. 26). Of notable change was an increase in *Actinobacteria* and *Bacilli* and a decrease in *Acidobacteria*, *Sphingobacteria*, *Betaproteobacteria*, and *Spartobacteria* across all evidence types. The most extreme changes occurred within the first 6 months of storage, while less change was evident between samples collected after six months and one year.

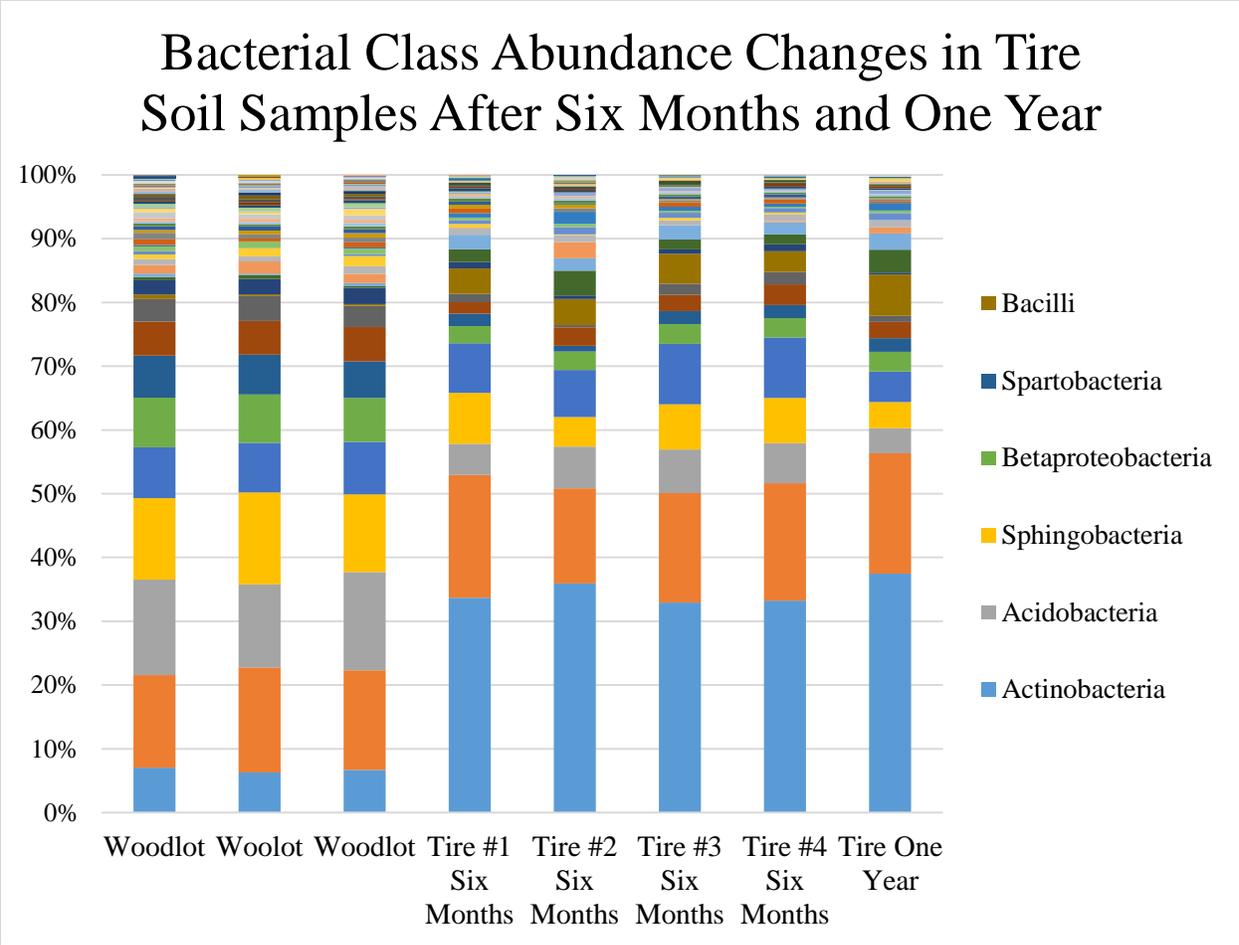


FIG. 26—Bacterial class abundance of soil samples collected from the woodlot of origin (left) and soil samples collected off of the tire after six months and one year in room temperature storage. Evidentiary profiles exhibited notable increase in *Actinobacteria* and *Bacilli* and a decrease in *Acidobacteria*, *Sphingobacteria*, *Betaproteobacteria*, and *Spartobacteria*.

*Nonmetric Multidimensional Scaling*

Soil samples collected off evidentiary items at both six months and one year after exposure clustered together in multidimensional space, away from all woodlots, but in closest proximity to the woodlot of origin (FIG. 27). Samples collected after one year of storage were slightly further away from the woodlot of origin than were the samples collected after six months.

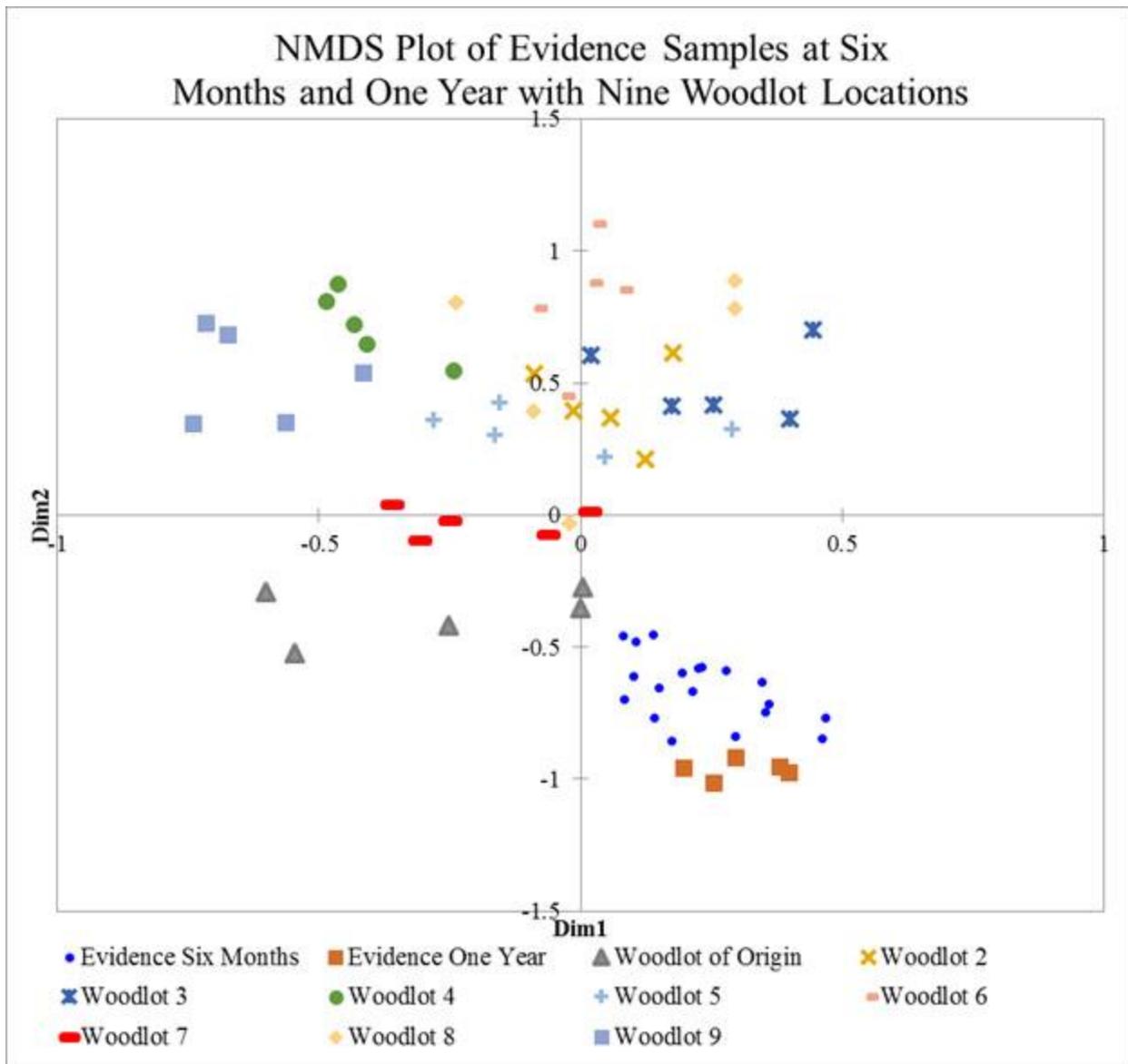


FIG 27—NMDS plot ordinating evidentiary and nine woodlot soil samples. Evidence samples collected after both six months and one year in storage clustered together, nearest the woodlot of origin, with the year samples plotting slightly further away.

*k-Nearest Neighbor*

*k*-NN accurately assigned soil collected off of evidentiary items back to their location of origin 100% of the time after both six months and one year (Table 4).

## Analysis of Secondary Evidentiary Samples

### Bacterial Abundance Charts

Bacterial profile abundance changes were evident in samples collected from t-shirts over the four month period (e.g. FIG. 28, FIG. 29) regardless of storage temperature. The same changes that were seen on the other evidentiary items occurred; an increase in *Actinobacteria* (FIG. 30) and *Bacilli* and a decrease in *Acidobacteria*, *Sphingobacteria* (FIG. 31), *Betaproteobacteria*, and *Spartobacteria*. These abundance changes began more slowly in the samples collected from t-shirts stored at 4°C.

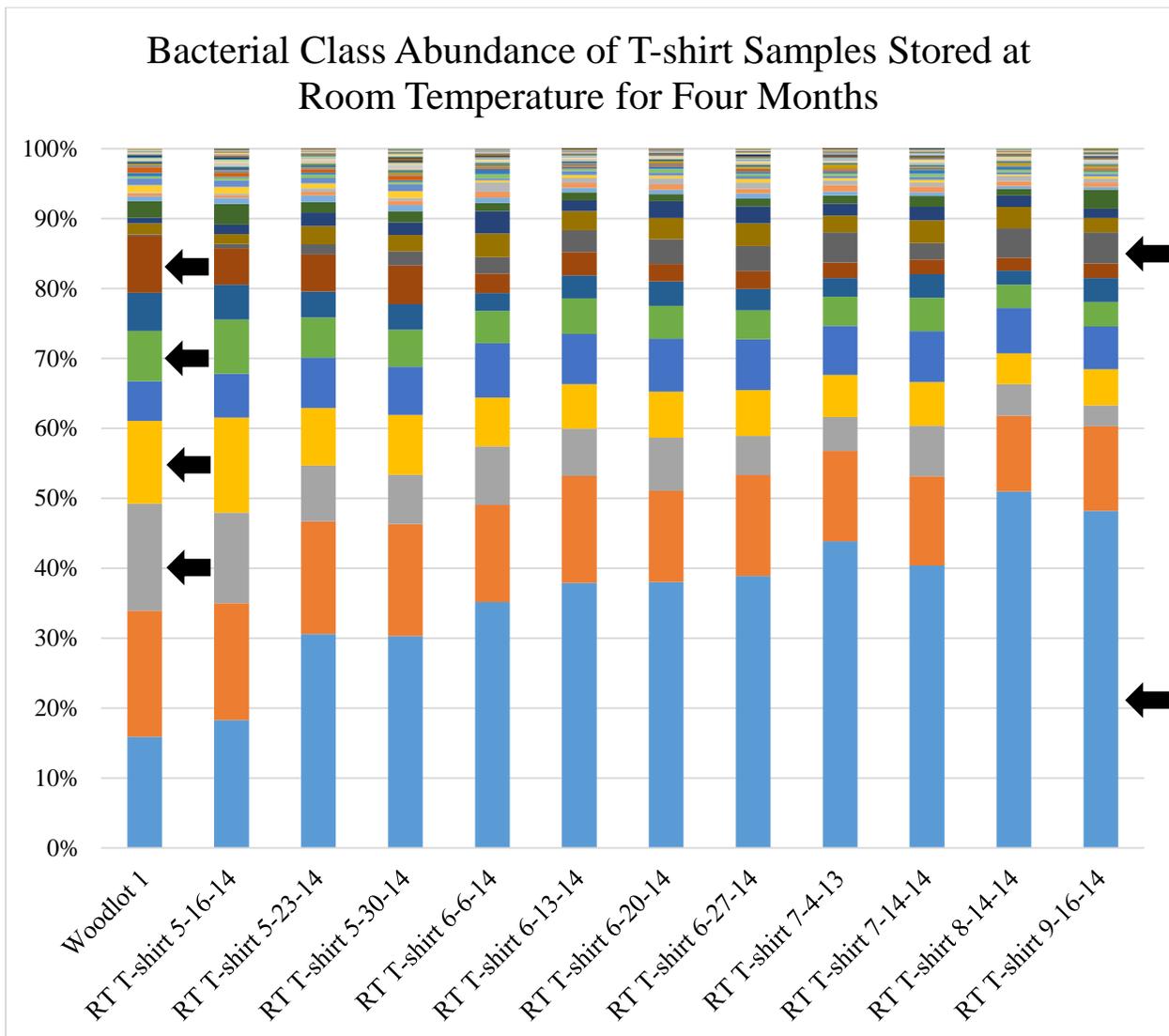


FIG. 28—Bacterial class abundance chart of room temperature t-shirt soil samples collected over the four month sampling period compared to soil collected from the woodlot of origin. Evidentiary soil samples exhibited notable increases in *Actinobacteria* and *Bacilli* (designated by arrows in ascending order on the right of the figure) and decreases in *Acidobacteria*, *Sphingobacteria*, *Betaproteobacteria*, and *Spartobacteria* (designated by arrows in ascending order on the left of the figure). RT=Room Temperature.

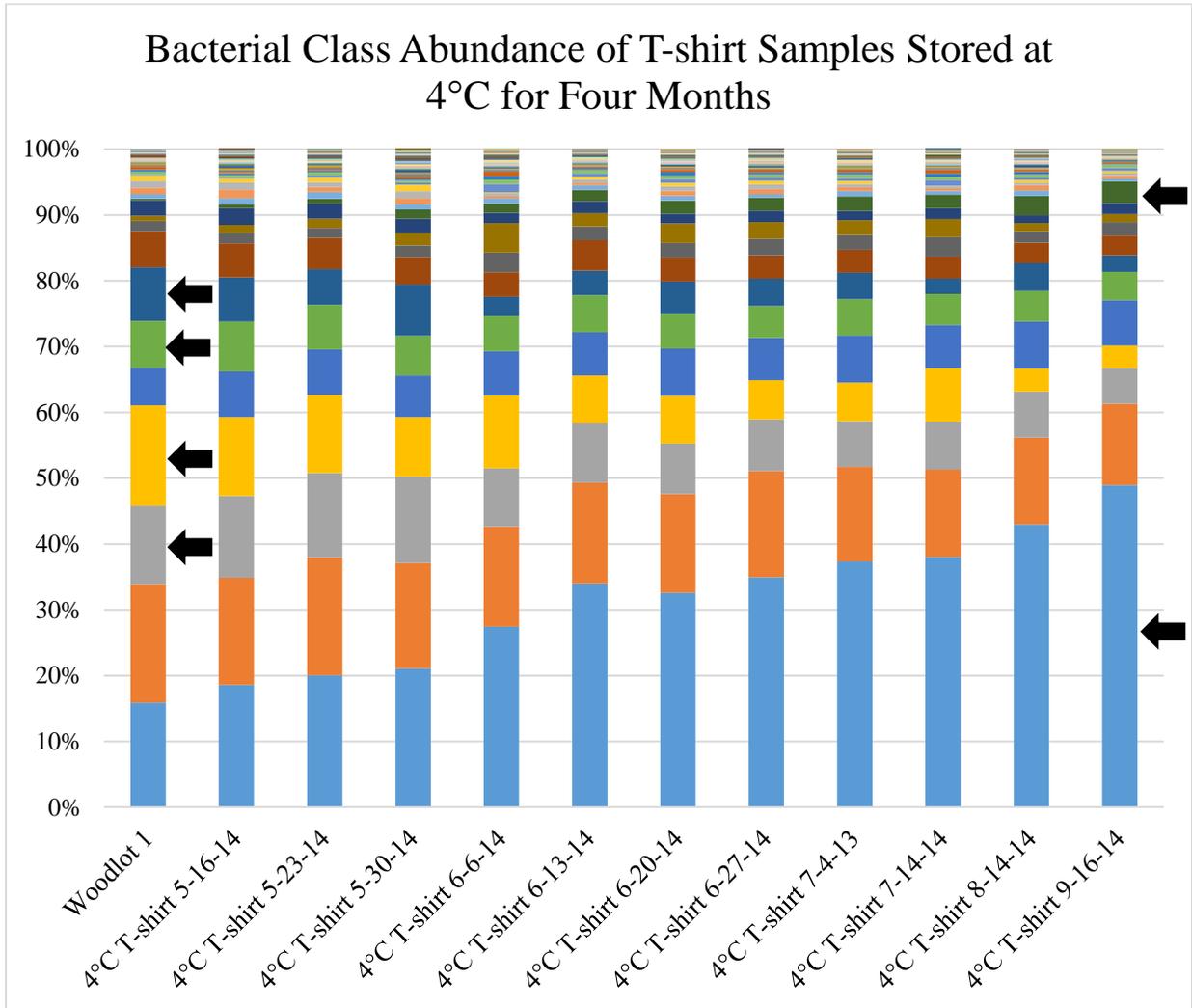


FIG. 29—Bacterial class abundance chart of 4°C t-shirt samples collected over the four month sampling period compared to one collected from the woodlot of origin. Evidentiary soil samples exhibited notable increases in *Actinobacteria* and *Bacilli* (designated by arrows in ascending order on the right of the figure) and decreases in *Sphingobacteria*, *Acidobacteria*, *Betaproteobacteria*, and *Spartobacteria* (designated by arrows in ascending order on the left of the figure).

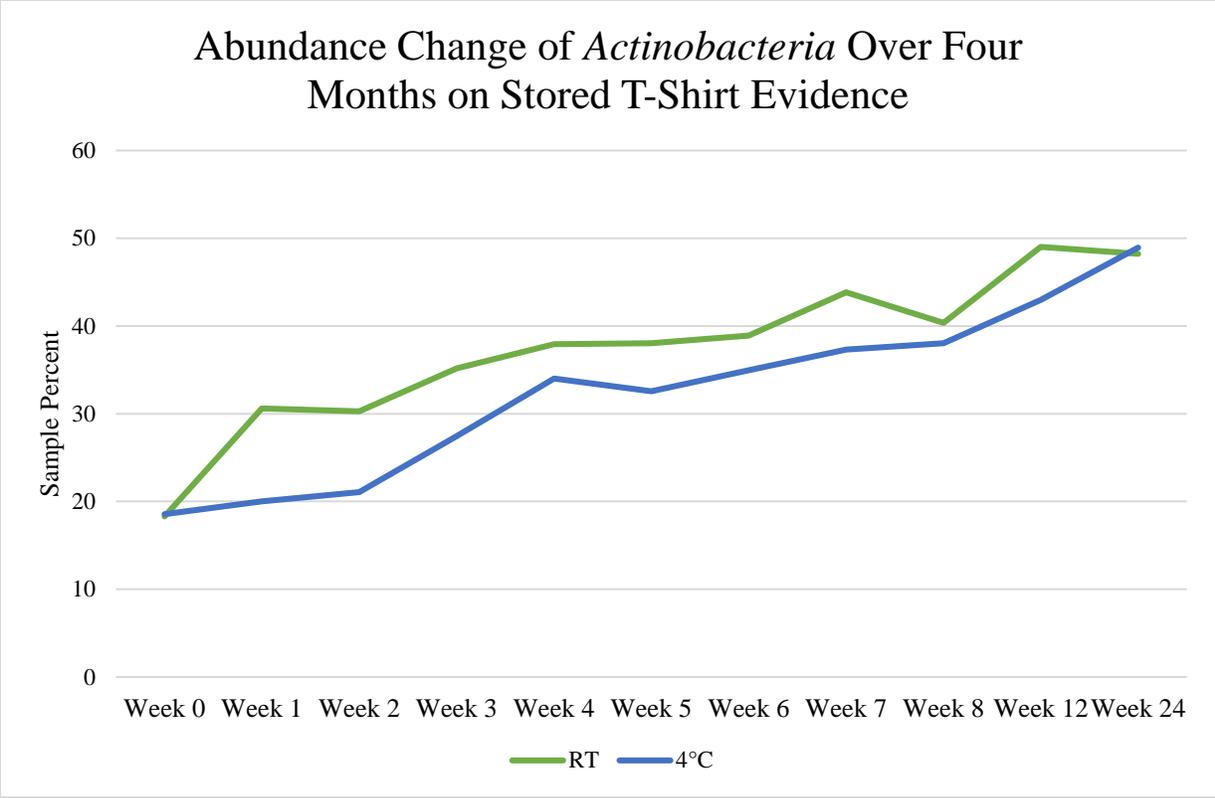


FIG. 30—Average (n=4) *Actinobacteria* abundance changes for bacterial profiles collected from t-shirts stored at room temperature (RT) and 4°C over a four month storage period. Members of this class increased in abundance over time in storage.

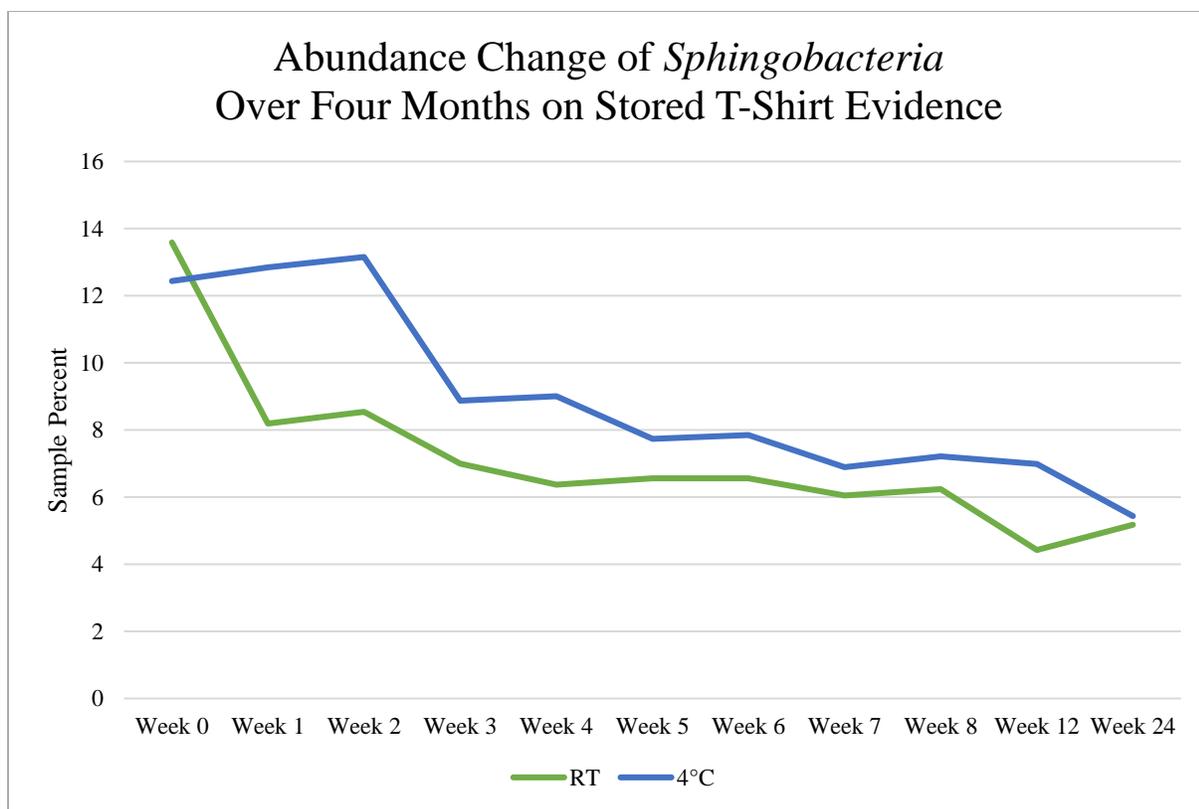


FIG. 31—Average (n=4) *Sphingobacteria* abundance changes for bacterial profiles collected from t-shirts stored at room temperature (RT) and 4°C over four month storage period. Members of this class decreased in abundance over time in storage.

### *Nonmetric Multidimensional Scaling*

T-shirt soil samples clustered together near their woodlot of origin initially (FIG. 32) and began to drift away from all woodlots in multidimensional space over the four month period (e.g. FIG. 33 and FIG. 34).

### *k-Nearest Neighbor*

*k*-NN accurately assigned 100% of the soil samples collected from t-shirt evidence to their woodlot of origin over the four month period (Table 4).

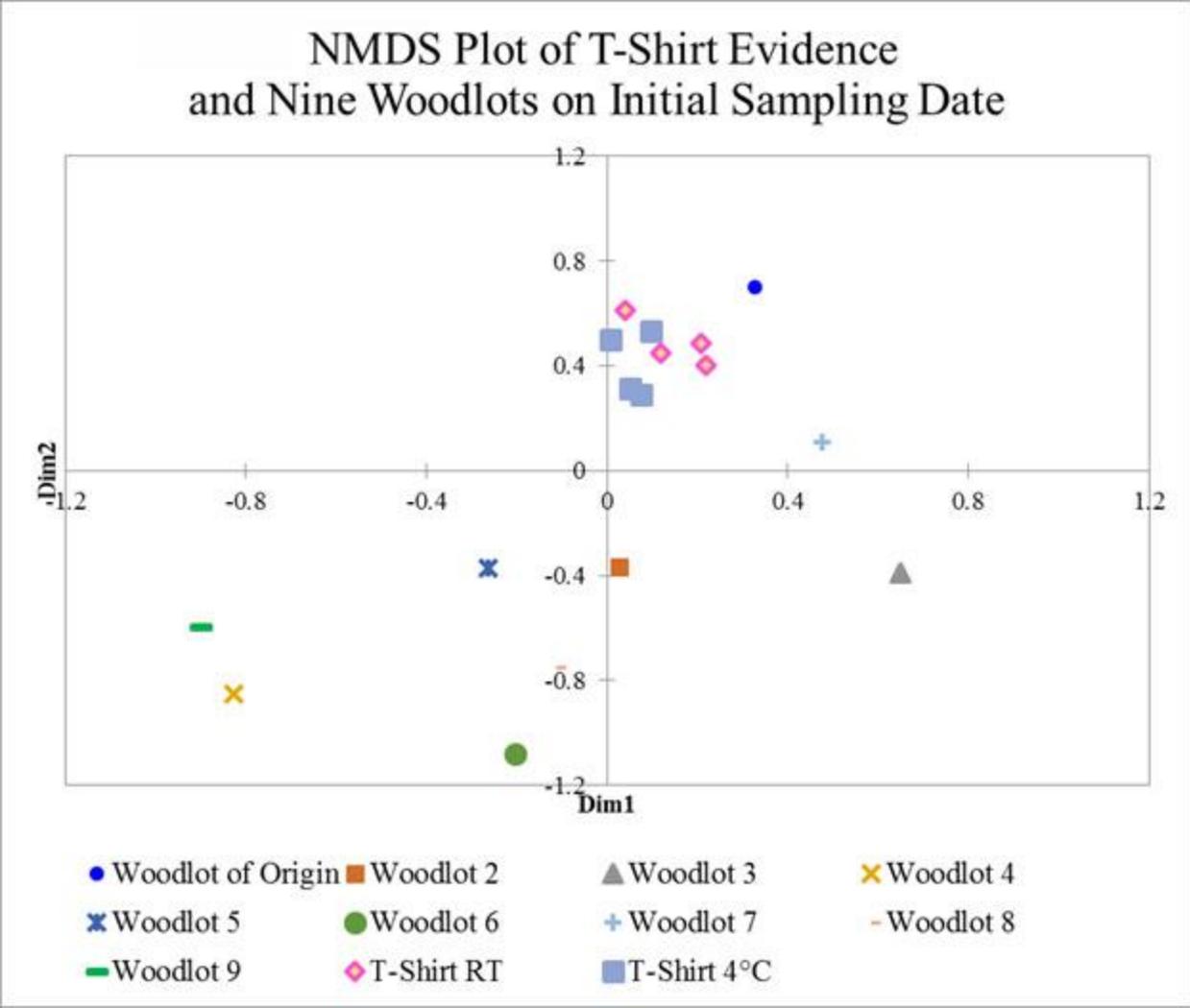


FIG. 32—NMDS plots of initial woodlot and t-shirt soil samples. Evidentiary soil samples clustered together nearest the woodlot of origin sample. RT=Room Temperature.

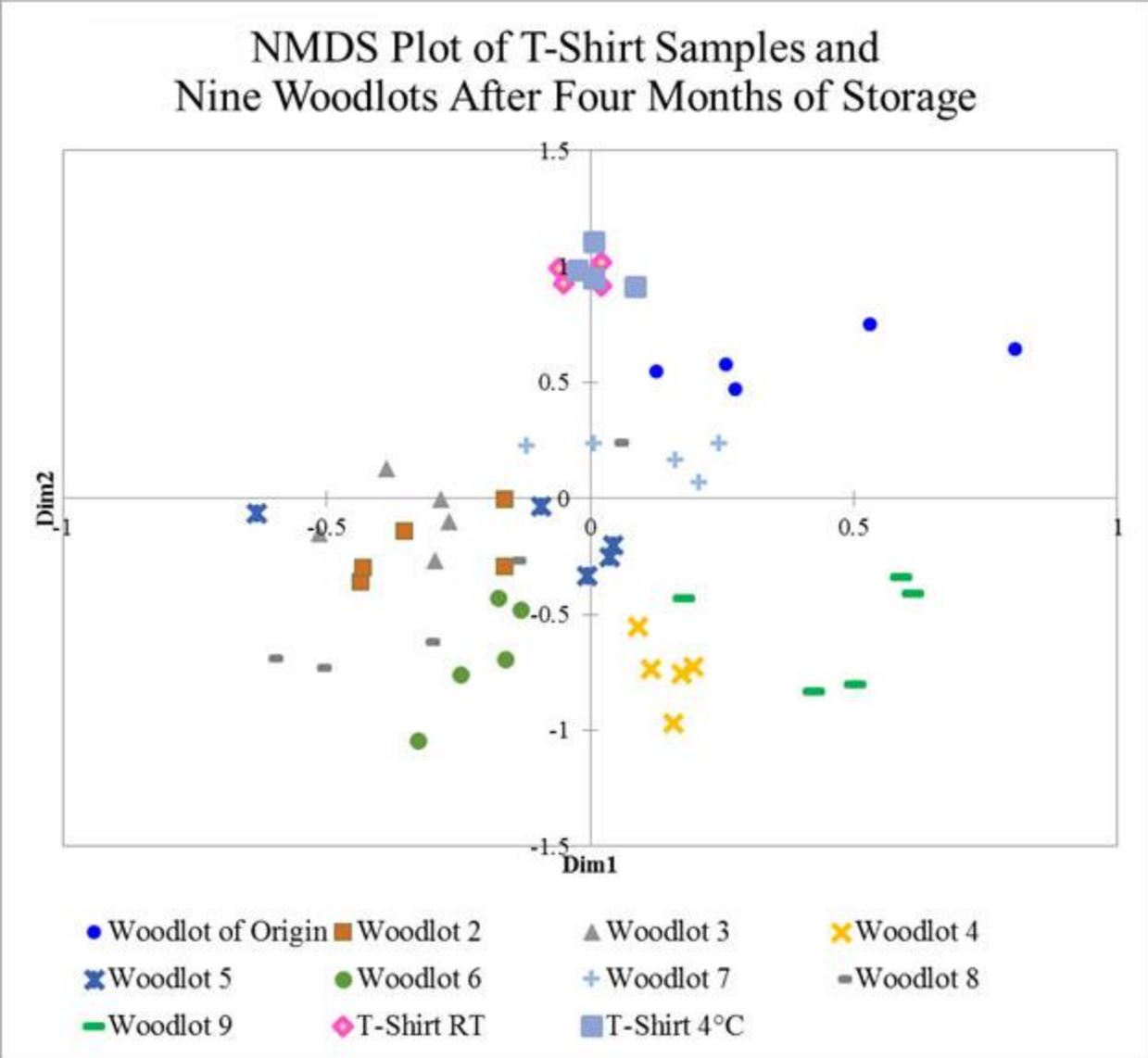


FIG. 33—NMDS plot of nine woodlots and t-shirt evidentiary samples after four months of storage. T-shirt samples from both storage temperatures drifted away from all woodlots, remaining closest to the woodlot of origin cluster. RT=Room Temperature.

### NMDS Plot of T-Shirts at Different Storage Temperatures and Nine Woodlots Over Four Months

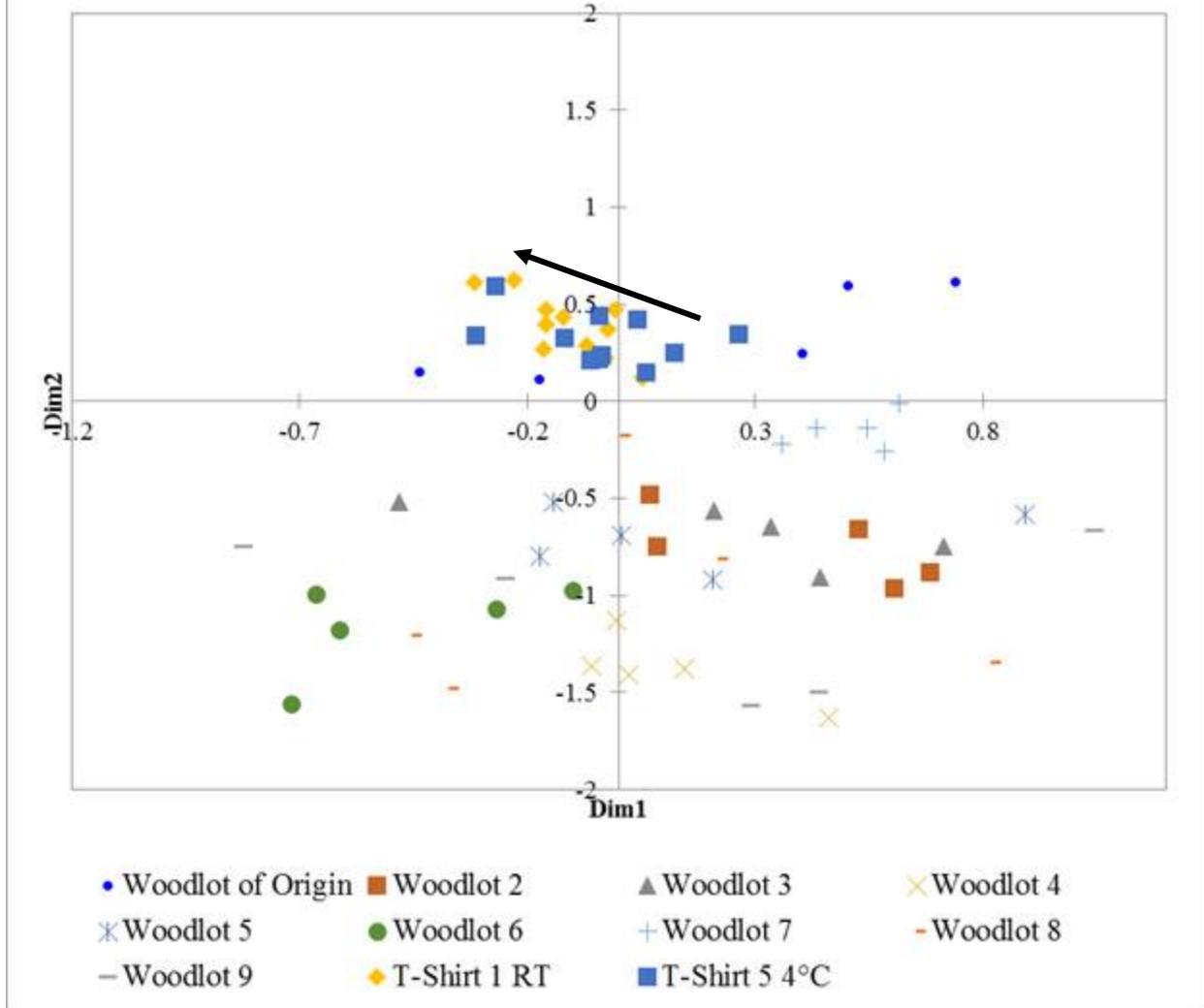


FIG. 34—NMDS plot of nine woodlots and soil samples collected from one t-shirt at each storage temperature (1 and 5) over four months. T-shirt soil samples cluster together near the woodlot of origin cluster. Samples drifted away from all woodlots over time (in the direction of the arrow).

## DISCUSSION

Soil evidence offers the potential to link parties or objects involved in a crime back to a scene. Bacterial populations within the soil can be used to evaluate the similarity between two samples and determine whether they originated from the same location. Past techniques for generating bacterial profiles have lacked in resolving power and required subjective analysis to form conclusions. The goal of the research detailed here was to determine the utility of a much more powerful technique, next-generation sequencing, for the production of bacterial profiles, and to test their applicability for forensic soil comparisons. Employing various analysis methods already utilized in microbiological research, the ability to differentiate soils from diverse and similar habitats, while also assessing differences within single locations spatially and temporally, was determined. Finally, the knowledge gleaned from these studies was applied to mock evidence to evaluate the utility of next-generation sequencing and the selected analysis techniques for tracing forensic soil samples back to a location of origin.

Earlier studies at Michigan State University using next-generation sequencing of soil bacteria (Hopkins 2014), acted as preliminary data for the NIJ funded research presented here. Employing 454 pyrosequencing, the utility of that technique for differentiating soil samples collected from three habitats was established. The data produced via pyrosequencing and the MiSeq platform used in the current research differed slightly. Although both sequencing methods resulted in similar averages of bacterial taxonomic classes across all samples (~50), the MiSeq platform produced more consistent results, with class numbers ranging from 42 – 58 (excluding the dirt road) while the 454 platform had a much wider range of 28 – 78 (excluding dirt road). 454 data variability was most prevalent among different runs, suggesting pyrosequencing data may not be as reproducible as those generated by the MiSeq. Of course, reproducibility is vital in

a forensic context as samples are often tested multiple times, potentially by several different individuals, and concordant results are requisite. Additionally, the data production capabilities and per sample cost of Illumina MiSeq sequencing likely make it more useful for forensic application. The 454 platform produced approximately 10,000 sequence reads at a cost of about \$100 per sample while the MiSeq produced fifteen fold more sequence data per sample at a cost of approximately \$15, and these costs are only expected to decrease.

Next-generation sequencing is a highly technical process that produces massive amounts of data, and myriad analysis techniques have been employed for data simplification and statistical comparisons. Several analysis methods already used by microbiologists to analyze next-generation sequencing data of microbiomes from both soil and other sources (e.g., Fierer and Jackson, 2006; Yang et al., 2006; Barnard et al., 2013; Li et al., 2014) were examined in our preliminary research. However, the demands of forensic soil analysis can be quite different from those in microbiology. In general, microbiologists utilize techniques that examine community structure to determine basic properties of different soils, while forensic scientists need more objective statistical measures that can lead to definitive assignments of soil samples to a location. Further, forensic scientists may be asked to present their highly technical results to lay audiences (attorneys, judges, juries) in a way that can be readily comprehended. This task can be aided by having visual interpretations of the data, acting as demonstrative evidence, which is often presented in court. Five analysis techniques were examined in the current research to determine which, if any, could satisfy these forensic needs. Pairwise comparisons and supervised classification acted as the objective measures, while bacterial abundance charts, HCA, and NMDS acted as visualization tools.

Two of the analysis methods examined in our preliminary studies were eliminated in the current research. The first, pairwise comparison, offers an objective statistical evaluation of sequence data, but the massive number of sequences and slight abundance fluctuations within a single habitat make pairwise statistics overly sensitive and thus untenable for forensic comparisons. When replicate soil samples from a single collection were compared using Jaccard, LIBSHUFF and UniFrac the majority differed significantly, showing how even the most minor bacterial variability can affect pairwise analysis. Forensically, pairwise comparisons could result in statistical differentiation of two samples that came from the exact same location, leading to a false exclusion.

The second procedure eliminated, HCA, provides a useful visualization of dissimilarity measures among bacterial profiles, however it does not provide any information not offered by other techniques. More importantly, the output from HCA varied based on the linkage method used, raising subjectivity concerns. If two experts used HCA while employing different linkage methods, the result they obtained could be substantially different, which is clearly undesirable forensically. In the end, the elimination of pairwise comparisons and HCA methodologies for forensic analysis left three techniques that were useful for characterizing bacterial profiling data while also meeting forensic science criteria: bacterial abundance charts, NMDS, and supervised classification techniques (e.g. *k*-NN).

Abundance charts offer a clear visualization of taxonomic classes present within a bacterial profile, while also exhibiting differences in quantity and diversity. In this research, virtually all of the soil samples contained the major bacterial classes in similar numbers, however, no two charts looked exactly alike. Bacterial abundance charts among diverse habitats appeared less alike than did those among similar habitats, which looked less alike than those

from within a habitat. The only real exception to this was the depth samples, which had clear differences as depth of collection increased (e.g., higher levels of *Clostridia* and *Nitrospira*). Environmental factors in deep soils (lower oxygen and nitrogen levels [Hinchee and Leeson, 1997; Schramm et al., 1999]) offer a possible explanation for why these bacterial classes were abundant in deeper samples from all habitats, as their members thrive under such conditions (O'Brien and Morris, 1971). Another interesting example of bacterial profile differences apparent through abundance charts was the dirt road samples. Soil collected from the dirt road exhibited surprising differences in specific bacterial classes when compared to other habitats, as well as a lower level of diversity. Upon further investigation, this likely resulted from treatment of the road with calcium chloride to reduce dust levels (Shiawassee County Road Commission, Personal Communication). Treatment can increase the salinity levels in the soil, which has been shown to lessen bacterial diversity (Hollister et al., 2010), while also favoring halophilic bacteria (Amoozegar et al., 2005; Quesada et al., 1983), many of which are found in the bacterial classes that were unusually high (*Clostridia*, *Bacilli*, *Flavobacteria*, and *Gammaproteobacteria* [Oren, 1983; Albuquerque et al., 2008; Ventosa et al., 1998; Sorokin et al., 2010]). As is evident from these examples, abundance charts can help explain why other analysis techniques are producing odd results. For example, if a known sample is forcing others together in an NMDS plot or consistently misclassifying when building a training set, an abundance chart can show why these odd phenomena are occurring and removal of that sample might be justified. However, while helpful for such irregular results, abundance charts of very similar soil samples will appear almost identical, therefore, other analysis techniques are necessary that can tease out subtle differences among bacterial profiles.

NMDS can help differentiate very similar soil samples that may not be detected through abundance charts. Furthermore, it too generates a visualization of the data, providing the advantages noted above. In the studies presented here, NMDS accurately reflected differences apparent in bacterial abundance charts through distant clustering (e.g. the dirt road samples plotted the furthest from other habitats and the depth samples plot progressively further from the surface sample) as well as providing new information among samples that had similar bacterial abundances. For example, abundance charts of the nine woodlots appeared similar, but distinct clusters of soil samples from each woodlot were formed when ordinated in multidimensional space. In both the diverse and similar habitat studies, NMDS plots exhibited intermingling of clusters when many samples were ordinated together, making interpretation difficult. However, separation of these clusters occurred when ordinating intermingling locations as pairs or triads. In a forensic context, it is more likely that a small number of locations will be considered as the origin of a questioned soil, as opposed to the nine or ten originally tested here, so the ordination of smaller numbers of sites reflects a more realistic scenario. In contrast, NMDS plots did not separate all samples from a location in an undesirable fashion; rather, it reflected similarities seen in abundance charts by plotting soils collected from the same location in close groupings. For example, abundance charts of soils within each habitat in the temporal study appeared very similar, and NMDS reflected this through tight clustering. Unfortunately, like abundance charts, NMDS does not produce any numerical measures of association within and among clusters, meaning a subjective interpretation is necessary to form conclusions.

Supervised classification techniques have the potential to fill this gap by providing an objective assignment of soil samples to specific groups based on training sets (forensic knowns), resulting in a definitive classification regarding a soil's origin. In the current research, *k*-NN

reliably assigned samples back to a location of origin with a small number of exceptions (see below). Another advantage of  $k$ -NN is that it offers classification information on both the training and test sets throughout the modeling process. If a training sample is classifying poorly, it can be identified and possibly removed. Furthermore, if a test soil sample is not classifying like its replicates,  $k$ -NN can be rerun with both the training set its replicates classify to and the training set it is misclassifying to. This was done in the diverse habitat study (e.g., between marsh and fallow agricultural field samples), which resulted in 100% accurate assignment. A disadvantage of  $k$ -NN for forensic purposes is that it is a hard classifier, meaning if knowns from two locations (for instance) are being used, an unknown sample will be classified to a group even if it does not belong to either, producing a misleading result, and possibly a false inclusion. Additionally,  $k$ -NN does not produce a p-value or other test statistic to measure the confidence of a classification. Although the added interpretation of abundance charts and NMDS can provide some confidence in location of origin assignments, other supervised classification techniques that can produce statistical values may perform better than  $k$ -NN for forensic analysis.  $k$ -NN is, however, recognized as a baseline supervised classification technique reflecting how datasets will perform using more advanced classification algorithms (Lavine and Davidson, 2006), so its success in the current research establishes the strength of supervised classification techniques for next-generation sequencing data interpretation.

After assessing the analytical utility of the next-generation sequence data analysis techniques, the first step of validating the process for forensic use was to determine if soils from diverse habitats could be differentiated. In previous soil studies, researchers were able to distinguish a small number of habitats in multidimensional space or through the presence or absence of specific bacteria (Fierer and Jackson, 2006; Lenz and Foran, 2010; Lauber et al.,

2009; Lauber et al., 2013; Kim et al., 2013), but overlap among habitats often occurred. In the current research, similar distinguishability and overlap existed when examining many habitats simultaneously, but this was resolved when fewer habitats were compared in both multidimensional space and via *k*-NN analysis. This increased differentiability may have been achieved due to the greater resolution of next-generation sequencing technology, where data production is much more extensive and bacterial profile generation is much more objective than in previous methods. Improved resolution may also be a result of more robust analysis techniques. Supervised classification was not attempted in past forensic soil research, and association of soil samples may have improved if it had been employed. Regardless of the reason for the improved results, it is clear that higher resolution next-generation sequencing and supervised classification greatly enhanced our ability to differentiate soils from diverse habitats.

Once it was evident that soils from environmentally diverse locations could be differentiated based on bacterial profiles, the next step was to determine if the same was true for similar habitats at locations in close proximity (a maximum of 6 miles apart). This presents a much greater challenge, as similar habitats are likely to share many of the same physical and chemical characteristics that affect bacterial communities. Researchers who have collected soil from similar habitats (including some of the studies cited above) generally pooled ecologically similar soils together for diverse habitat comparison rather than attempting to differentiate locations. Although this procedure is useful for microbiological comparisons because it establishes basic properties of soil types, forensic scientists may need to differentiate ecologically similar locations in criminal cases. This challenging task was largely achieved in the current research through differentiation of nine woodlot sites, with the exception of woodlots 8 and 9, which exhibited substantial intra-location variation. This was resolved by ordinating fewer

locations in NMDS plots, but could not be totally overcome in *k*-NN analysis. In this regard, recommendations for collection of additional soil samples in actual forensic settings, where clustering and classification within a location is vital, can be considered. Multiple collections on the same day may have produced a more accurate training set, while also being more likely in a forensic context. Despite the results for two of the woodlots, the very similar habitats were largely differentiated and correctly classified in this study, demonstrating the high resolution of the bacterial profile data produced via MiSeq sequencing.

With the strength of next-generation sequencing of soil bacteria for differentiating diverse and similar habitats established, factors that may influence bacterial profiles within a location were considered. It is fundamentally impossible to collect known soil samples precisely when a crime occurs; consequently, temporal changes in bacterial makeup must be examined. Past studies assessing change over time (through T-RFLP [Meyers and Foran, 2008] and pyrosequencing [Lauber et al., 2013; Hopkins, 2014] of the 16S locus) have shown substantial differences in bacterial profiles collected across seasons, resulting in intermingling of habitats and different levels of bacterial diversity within a habitat. Some temporal changes were evident in the current study, but bacterial profiles remained relatively stable across seasons, most likely due to the superior resolving power of the MiSeq platform. Again, the results from this study provide important sample collection information as the few soil samples that did not cluster well in NMDS plots and misclassified in *k*-NN analysis were collected in February and March. The bacterial profiles generated from these samples may have been affected by water, because the bags contained ice and snow during transport and storage. If this was the cause of profile deviation, crime scene investigators will need to use caution when collecting soils in icy conditions, possibly by drying soils after collection. Aside from these ‘slushy’ soil samples,

seasonal changes did not affect analysis accuracy in this research, and soils were correctly assigned to their origin across all seasons.

It is also unlikely that known soil samples will be collected from the precise spot to which the evidence item was exposed, but instead could be collected feet, yards, or greater distances away, highlighting the importance of considering spatial variability of soil within a location. Differences in bacterial profiles over small distances has been attributed to microenvironmental factors such as foliage, pH, nutrient supply, etc. (Ettema and Wardle, 2002; Eichorst et al., 2007), although in reality, any number of factors could come into play. In a forensic study, Meyers and Foran (2008) described variability, sometimes substantial, among soil samples collected even 10 feet away from one another. Similar variation or patchiness was reflected in soil samples collected in the current research, where surface samples collected across a habitat or soils at different depths did not always cluster well in multidimensional space. Despite this variation, *k*-NN analysis resulted in the accurate assignment of spatial samples, highlighting the importance of using an effective training sets to capture within habitat variation. Soil samples collected at different depths showed the most bacterial variability of all studies (excluding the dirt road), and using the surface soils alone as a training set did not reflect this variation, as deeper samples misclassified. When a range of samples (surface, 2", 10", and 60") from each habitat were used as a training set, the other soil samples were all correctly assigned to their site of origin. This presents a possible collection strategy for situations like burials, where known soil samples from throughout the burial depth, either individually or mixed, should likely be collected to build an accurate training set. Analysis of soil samples collected across the surface of a habitat also reflected the importance of proper training sets, as different knowns resulted in different classification success. It is noteworthy that samples that misclassified were

always 90 feet or more away from the nearest training soil sample. Given this, developing multiple training sets for each location in question could possibly be applied to more confidently determine an unknown soil sample's origin. Overall, although bacterial variation was apparent in the spatial studies, effective analysis through incorporation of various training sets allowed for accurate assignment of soil samples.

The final portion of this research combined similar habitats, temporal changes, and mock evidence to assess traceability of soils in more realistic forensic scenarios. Most importantly, soil evidence traced back to its location of origin 100% of the time even after a full year of storage, highlighting the success of the sequencing technique and supervised classification for forensic analysis. Further, similar abundance changes occurred across all stored evidentiary items, where specific classes of bacteria increased or decreased over time. Rather than being a hindrance for soil evidence investigation, predictable changes in bacterial profiles offer the potential to act as a biological clock of how long soil has been removed from a habitat, providing a valuable tool in criminal investigations. Although evidence was only exposed to soil from a single woodlot, it is possible these same bacterial class abundance changes will occur in soil from other habitats, as these taxonomic classes that changed are common in most or all soils. While we do not yet fully understand why these changes are occurring, in retrospect, some sense can be made of them. For example, *Actinobacteria*, which increased over time, has members that thrive in dry environments (Ghorbani-Nasrabadi et al., 2013), and *Bacilli*, which also increased, has members that react to changing environments very efficiently due to spores that are highly resilient compared to other bacterial strains (Claus and Berkeley, 2009). Such characteristics could aid the endurance of these classes in storage. If these same abundance changes occur in soils from other habitats, there is strong potential for the generation of a biological clock to predict how long soil

has been removed from a location. It should be noted that such a clock will be influenced by storage temperature, given that cooler temperatures slowed the start of bacterial abundance change across evidentiary items. This temperature dependent transformation will affect how biological clocks are calibrated (e.g. if a piece of evidence is stored in an outdoor shed during winter versus an indoor closet, the extent of abundance changes over that period of time would differ). Therefore, when soil evidence is collected it will be important to note storage temperature. Most importantly, the evidence studies show that soil bacterial profiles can be effectively used to trace evidence back to a location of origin regardless of the length of time it was stored.

## **CONCLUSIONS**

Based on the research presented, next-generation sequencing of soil bacteria has met the requirements for a new microbial technique to be forensically applicable. The amount of data produced and the resolution of next-generation technology (e.g. Illumina MiSeq sequencing) far surpass that of previous bacterial profiling methods, allowing for more robust, reproducible data generation and the potential for soil individualization. Evidentiary soil samples may be contested as being from one location or another, and the research presented here demonstrates that next-generation sequencing data can be used to differentiate the locations and place a questioned soil back to one of them. Within habitat factors (i.e., time and space) had little or no effect on soil analysis, further reinforcing the utility of bacterial profile generation via next-generation sequencing. The results of mock evidentiary studies increased confidence in next-generation sequencing for forensic soil profiling, while also introducing the possibility of estimating the time soil has been removed from a source. Finally, the type of data analysis performed played a critical role in this research. The results show that it may be highly worthwhile to utilize more

than one analysis method so as to meet several goals. Clear visual representations will aid the jury's understanding of highly technical soil evidence, which are effectively generated through abundance charts and NMDS. The two techniques act in a complementary manner wherein the former provides a categorization and quantification of the copious sequences, and the latter produces relative similarity information. Together they can then be used by the expert witness to explain categorizations generated by supervised classification techniques, which deliver the objective assignment of samples to their location of origin. In combination, these three analysis strategies can effectively be used for next-generation sequencing data, providing an avenue for forensic soil analysis to enter the courtroom.

The process of implementing next-generation sequencing of soil bacteria into crime laboratories will involve multiple steps. First, sequencing technology must be currently available, or in most cases added to the laboratory. However, it is unlikely every crime laboratory will have the means or desire to purchase expensive equipment like a next-generation sequencer, especially if only a handful of samples are run per year. In this regard, the technology could be made available through a central or regional laboratories, much like the mitochondrial DNA laboratories equipped by the FBI. Next, the methods must be accepted by the courts, satisfying the Frye and Daubert standards for any new scientific technique. Both of these will take time, yet despite such hurdles, this research has shown the tremendous potential for next-generation sequencing of forensic soil samples, which far surpasses the class characteristics generated through traditional soil analysis methods. The combination of powerful next-generation sequencing technology and reliable, robust data analysis means it is feasible to link an evidentiary item, victim, or suspect to a crime scene.

## IMPLICATIONS FOR FURTHER RESEARCH

Scientific research always leads to further questions, as is the case following the studies detailed herein. Next-generation sequencing has been established as a powerful technique for soil differentiation and traceability, but additional studies will help to ensure its feasibility as a forensic procedure. The minimum amount of soil required for testing was not examined in this research. The evidentiary studies described here were only conducted using soil from deciduous woods habitats, and other habitat types need to be examined. Similarly, the evidentiary biological clock was only detected using soil from a deciduous woods, which should be expanded to other habitat type to ensure similar changes occur. Crime scene evidence is rarely in pristine condition when exposed to soil (as were the t-shirts in this study), and it is unknown how bodily fluids on evidentiary items (e.g., perspiration or blood) might affect bacterial profile analysis, given they can have their own microbiome or act as nutrients. The three data analysis methods in combination proved useful in the current research, but further consideration of other supervised classification techniques should be undertaken to find a more statistical measure of confidence than is provided by  $k$ -NN, and one that is not a hard classifier. There should also be research into strategies for collecting known samples that act as training sets for the supervised classification methods. Finally, the experiments detailed in this report were never conducted in a blind fashion, which is how they would exist in criminal investigation. Once the factors above are considered, blind testing of the methods should take place.

## DISSEMINATION OF RESEARCH FINDINGS

Hopkins and Foran. Forensic Exploration of Temporal and Spatial Differences of Soils Using Next-Gen Sequencing. The American Academy of Forensic Sciences Annual Meeting, Feb. 2014, Seattle, WA

Foran and Jesmok. Collaborative Efforts in Developing Reliable Genetic and Statistical Methods for Microbial Profiling of Soils. NIJ Grantees Meeting Feb. 2015, Orlando, FL

Jesmok, Hopkins, and Foran. The Influence of Distance, Depth, and Time on Forensic DNA Profiling of Soil Bacteria. The American Academy of Forensic Sciences Annual Meeting, Feb. 2015, Orlando, FL

Jesmok and Foran. Bacterial Profiling of Soil Evidentiary Items Using Next-Generation Sequencing. The American Academy of Forensic Sciences Annual Meeting, Feb. 2015, Orlando, FL

Jesmok, Hopkins, and Foran. Next-Generation Sequencing of the Bacterial 16S rRNA Gene for Forensic Soil Comparison: A Feasibility Study. Under review: Journal of Forensic Sciences<sup>4</sup>

Hopkins, Jesmok, and Foran. Statistical Methods for the Analysis of Forensic Soils based on Next-Generation Sequencing. Manuscript in prep<sup>4</sup>

---

<sup>4</sup> Portions of this report form the basis of this manuscript

## **ACKNOWLEDGEMENTS**

The authors would like to thank past students involved in soil research at Michigan State University, including: Melissa Meyers, Erin Lenz, and Ethan Smith. Thank you to Dr. Tom Schmidt, Dr. Jeff Landgraf, and Dr. Ruth Smith for helping to develop techniques for bacterial profile generation and analysis.

## REFERENCES

- Albuquerque L, Tiago I, Taborda M, Nobre MF, Verissimo A, da Costa MS. *Bacillus isabelliae* sp. nov., a halophilic bacterium isolated from a sea salt evaporation pond. *Int J Syst Evol Microbiol* 2008; 58(Pt 1): 226 – 30.
- Alden A. Sherlock Holmes, Forensic Geologist.  
<http://geology.about.com/od/bookreviews/a/sherlock.htm>
- Amoozegar MA, Hamed J, Dadashpour M, Shariatpanahi S. Effects of salinity on the tolerance to toxic metals and oxyanions in native moderately halophilic spore-forming bacilli. *World Journal of Microbiology and Biotechnology* 2005; 21(6-7): 1237 – 43.
- Barnard RL, Osborne CA, Firestone MK. Responses of soil bacterial and fungal communities to extreme desiccation and rewetting. *ISME J* 2013; 7: 2229 – 41.
- Beebe KR, Pell RJ, Seasholtz MB. *Chemometrics a practical guide*. New York: John Wiley & Sons, Inc., 1998.
- Bergslien E. *An Introduction to Forensic Geoscience*. New Jersey: John Wiley & Sons, Ltd., 2012.
- Borg I, Groenen PJF, Mair P. *Applied multidimensional scaling*. Berlin: Springer, 2013.
- Bray JR and Curtis JT. An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecol Monogr* 1957; 27(4): 325 – 49.
- Bressan D. C.S.I Geology: Forensic geosciences and its application to criminal investigations. <<http://historyofgeology.fieldofscience.com/2010/07/csi-geology-forensic-geosciences-and.html>>. Accessed 2015 March 30.
- Caporaso GJ, Lauber CL, Walters WA., Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M, Gormley N, Gilbert JA, Smith G, Knight R. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J* 2012; 6: 1621 – 4.
- Claus D, Berkeley RCW. *Bergey's Manual of Systematic Bacteriology, Second Edition, Volume 2*. Springer Dordrecht Heidenberg London, New York, 2009: 1105 – 39.
- Cox T. *Multidimensional Scaling*. Boca Raton: Chapman and Hall, 2001.
- Coyle HM, editor. *Nonhuman DNA Typing*. Boca Raton: Taylor & Francis Group, 2008.
- Daniel R. The Metagenomics of Soil. *Nat Rev Microbiol* 2005; 3: 470 – 8.
- Daubert v. Merrell Dow Pharmaceuticals (92-102), 509 U.S. 579 (1993).

- Dice LR. Measures of the Amount of Ecologic Association Between Species. *Ecology* 1945; 26(3): 297 – 302.
- Eichorst SA, Breznak JA, Schmidt TM. Isolation and Characterization of Soil Bacteria That Define *Terriglobus* gen. nov., in the Phylum Acidobacteria. *Appl Environ Microbiol* 2007; 73(8): 2708 – 17.
- Ettema CH and Wardle DA. Spatial soil ecology. *Trends Ecol Evol* 2002; 17(4): 177 – 83.
- Fierer N, Jackson JA. The diversity and biogeography of soil bacterial communities. *Proc Natl Acad Sci USA* 2006; 103(3): 626 – 31.
- Ghorbani-Nasrabadi R, Greiner R, Alikhani HA, Hamed J, Yakhchali B. Distribution of actinomycetes in different soil ecosystems and effect of media composition on extracellular phosphatase activity. *J Soil Sci Plant Nutr* 2013; 13(1): 223 – 36.
- Heath LE, Saunders VA. Assessing the Potential of Bacterial DNA Profiling for Forensic Soil Comparisons. *J Forensic Sci* 2006; 51(5): 1062 – 8.
- Hinchee RE, Leeson A. Soil Bioinventing: Principles and Practices.
- Hollister EB, Engledow AS, Hammett AJM, Provin TL, Wilkinson HH, Gentry TJ. Shifts in microbial community structure along an ecological gradient of hypersaline soils and sediments. *ISME J* 2010; 4: 829 – 38.
- Hopkins JM. Forensic soil bacterial profiling using 16S rRNA gene sequencing and diverse statistics. Thesis. Michigan State U, 2014. Print.
- Kaltenbach HM. A concise guide to statistics. New York: Springer, 2012.
- Kim M, Boldgiv B, Singh D, Chun J, Lkhagva A, Adams JM. Structure of soil bacterial communities in relation to environmental variables in a semi-arid region of Mongolia. *J Arid Environ* 2013; 89: 38 – 44.
- Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the MiSeq Illumina Sequencing Platform. *Appl and Environ Microbiol* 2013; 79(17): 5112 – 20.
- Lauber CL, Hamady M, Knight R, Fierer N. Pyrosequencing-Based Assessment of Soil pH as a Predictor of Soil Bacterial Community Structure at the Continental Scale. *Appl Environ Microbiol* 2009; 75(15): 5111 – 20.
- Lauber CL, Ramirez KS, Aanderud Z, Lennon J, Fierer N. Temporal variability in soil microbial communities across land-use types. *ISME J* 2013; 1 – 10.

- Lavine BK, Davidson CE. Gemperline, P ed. *Practical Guide to Chemometrics 2<sup>nd</sup> Edition*. CRC Press 2006; 9: 339 – 78.
- Lenz EJ, Foran DR. Bacterial Profiling of Soil Using Genus-Specific Markers and Multidimensional Scaling. *J Forensic Sci* 2010; 55(6): 1437 – 42.
- Li J, Chen Q, Long LJ, Dong JD, Yang J, Zhang S. Bacterial dynamics within the mucus, tissue and skeleton of the coral *Porites lutea* during different seasons. *Sci Rep* 2014; 4: 7320.
- Liu WT, Marsh TL, Cheng H, Forney LJ. Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. *Appl Environ Microbiol* 1997; 63(11): 4516 – 22.
- Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ. Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnol* 2012; 30(5): 434 – 9.
- Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* 2005; 71(12): 8228 – 35.
- MacDonald L, Singh B, Thomas N, Brewer M, Campbell C, Dawson L. Microbial DNA profiling by multiplex terminal restriction fragment length polymorphism for forensic comparison of soil and the Influence of Sample Condition. *J Appl Microbiol* 2008; 105(3): 813 – 21.
- MacLean D, Jones JDG, Studholme DJ. Application of ‘next-generation’ sequencing technologies to microbial genetics. *Microbiol* 2009; 7: 287 – 96.
- Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA *et al*. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 2005; 437(15): 376 – 80.
- Metzker ML. Sequencing technologies—the next generation. *Nature Rev* 2010; 11(1): 31 – 46.
- Meyers MS, Foran DR. Spatial and temporal influences on bacterial profiling of forensic soil samples. *J Forensic Sci* 2008; 53(3): 652 – 60.
- Mohri M, Rostamizadeh A, Talwalkar A. *Foundations of Machine Learning*. The MIT Press 2012. ISBN 9780262018258.
- Moreno LI, Mills DK, Entry J, Sautter RT, Mathee K. Microbial metagenome profiling using amplicon length heterogeneity-polymerase chain reaction proves more effective than elemental analysis in discriminating soil specimens. *J Forensic Sci* 2006; 51(6): 1 – 8.

- Muyzer G, Smalla K. Application of denaturing gradient gel electrophoresis (DGGE) and temperature gradient gel electrophoresis (TGGE) in microbial ecology. *Antonie van Leeuwenhoek* 1998; 73(1): 127 – 41.
- National Research Council. (2009). *Strengthening Forensic Science in the United States: A Path Forward*. Washington, D.C.: National Academies Press.
- Oren A. *Clostridium lortetii* sp. nov., a halophilic obligatory anaerobic bacterium producing endospores with attached gas vacuoles. *Arch Microbiol* 1983; 136: 42 – 8.
- Pye K, Blott SJ, Croft DJ, Carter JF. Forensic comparison of soil samples: Assessment of small-scale spatial variability in elemental composition, carbon and nitrogen isotope ratios, colour, and particle size distribution. *Forensic Sci Int* 2006; 163(1 – 2): 59 – 80.
- Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 2013; 41 (D1): D590 – 6.
- Quesada E, Ventosa A, Rodriguez-Valera F, Megias L, Ramos-Cormenzana A. Numerical taxonomy of moderately halophilic gram-negative bacteria from hypersaline soils. *J Gen Microbiol* 1983; 129: 2649 – 57.
- Ritz K, Dawson L, Miller D. *Criminal and Environmental Soil Forensics*. Springer, 2008.
- Saferstein R, editor. *Forensic Science Handbook*. Volume 1, 2<sup>nd</sup> edition. Upper Saddle River: Prentice Hall, 2002.
- Schloss PD. Evaluating different approaches that test whether microbial communities have the same structure. *ISME J* 2008; 2: 265 – 75.
- Schloss PD, Westcott SI, Ryabin T, Hall JR, Hartmann M, Hollister EB. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 2009; 75(23): 7537 – 41.
- Schramm A, Beer D, Heuvel JC, Ottengraf S, Amann R. Microscale distribution of populations and activities of *Nitrosospira* and *Nitrospira* spp. along a macroscale gradient in a nitrifying bioreactor: Quantification by in situ hybridization and the use of microsensors. *Appl Environ Microbiol* 1999; 65(8): 3690 – 6.
- Sensabaugh GF. Microbial community profiling for the characterization of soil evidence: Forensic considerations. In *Criminal and Environmental Soil Forensics* Springer 2009; 49 – 60.
- Shokralla S, Spall JL, Gibson JF, Hajibabaei M. Next-generation sequencing technologies for environmental DNA research. *Mol Ecol* 2012; 21(8):1794 – 805

- Singleton DR, Furlong MA, Rathbun SL, Whitman WB. Quantitative comparisons of 16S rRNA gene sequence libraries from environmental samples. *Appl Environ Microbiol* 2001; 67(9): 4372 – 6.
- Sørensen T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Kongelige Danske Videnskabernes Selskab* 1948; 5(4): 1 – 34.
- Sorokin DY, Kovaleva OL, Tourova TP, Muyzer G. *Thiohalobacter thiocyanaticus* gen. nov., sp. nov., a moderately halophilic, sulfur-oxidizing gammaproteobacterium from hypersaline lakes, that utilizes thiocyanate. *Int J Syst Evol Microbiol* 2010; 60: 444 – 50.
- Sutton NB, Maphosa F, Morillo JA, Al-Soud WA, Langenhoff AAM, Grotenhuis T, Rijnaarts HHM, Smidt H. Impact of Long-Term Diesel Contamination on Soil Microbial Community Structure. *Appl Environ Microbiol* 2013. 79(2): 619 – 30.
- Tukey JW. Bias and confidence in not-quite large samples. *Ann Math Stat* 1958; 29: 614 – 5.
- Ventosa A, Nieto JJ, Oren A. Biology of Moderately Halophilic Aerobic Bacteria. *Microbiol Mol Biol Rev* 1998; 62(2): 504 – 44.
- Whittaker R. Vegetation of the Siskiyou mountains, Oregon and California. *Ecol Monographs* 1960; 30(3): 279 – 338.
- Yang C, Mills D, Mathee K, Wang Y, Jayachandran K, Sikaroodi M *et al.* An ecoinformatics tool for microbial community studies: Supervised classification of amplicon length heterogeneity (ALH) profiles of 16S rRNA. *J Microbiol Meth* 2006; 65(1): 49 – 62.