10-28-2016

Biogeographical Patterns of Soil Microbial Communities: Ecological, Structural, and Functional Diversity and their Application to Soil Provenance

Natalie Damaso
Florida International University, ndama001@fiu.edu

DOI: 10.25148/etd.FIDC001216

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

BIOGEOGRAPHICAL PATTERNS OF SOIL MICROBIAL COMMUNITIES: ECOLOGICAL, STRUCTURAL, AND FUNCTIONAL DIVERSITY AND THEIR APPLICATION TO SOIL PROVENANCE

A dissertation submitted in partial fulfillment of

the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BIOLOGY

by

Natalie Damaso

2016

To: Dean Michael R. Heithaus
College of Arts, Sciences, and Education

This dissertation, written by Natalie Damaso, and entitled Biogeographical Patterns of Soil Microbial Communities: Ecological, Structural, and Functional Diversity and their Application to Soil Provenance, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this dissertation and recommend that it be approved.

_____
Eric Von Wettberg

_____
Jeffrey Wells

_____
Jennifer Gebelein

_____
John Kominoski

_____
DeEtta Mills, Major Professor

Date of Defense: October 28, 2016

The dissertation of Natalie Damaso is approved.

_____
Dean Michael R. Heithaus
College of Arts, Sciences, and Education

_____
Andrés G. Gil
Vice President for Research and Economic Development
and Dean of the University Graduate School

Florida International University, 2016

DEDICATION

I dedicate this dissertation to my mother, Esther Martinez, and Scott Langevin. Thank you for all of your support and encouragement. I couldn't have done this without you.

ACKNOWLEDGMENTS

I first would like to acknowledge my support system throughout my dissertation, without which I could not have completed this project. My major advisor, Dr. DeEtta Mills for her support, patience, and great mentorship. Her guidance through both my academic and professional development has been invaluable. My committee members, Dr. Eric vonWettberg, Dr. John Kominoski, Dr. Jeffrey Wells, and Dr. Jennifer Gebelein for their time and guidance throughout the program. Your various expertise has made me think more broadly as well as helped me improve my scientific knowledge and skills. To my lab mates, your friendship and encouragement has made the lab feel like home; our coffee/croquetta breaks and gym classes will be greatly missed. I would also like to thank my family and friends, whom there are too many to list. Lastly, to Scott Langevin and my mother, Esther Martinez, for being my support structure and keeping me positive.

I would also like to acknowledge the Forensic DNA Profiling Facility within the International Forensic Research Institute at Florida International University for their instrumental support. I would also like to express my gratitude towards the Biology Department, McNair Graduate Fellowship and the FIU Graduate School Dissertation Year Fellowship for their support in funding me throughout my dissertation. This work was funded in part through a grant to DeEtta Mills, HM1582-09-1-0011 from the National Geospatial Intelligence Agency and a grant to Natalie Damaso, 800003135 from Sigma Xi Grants-in-Aids.

ABSTRACT OF THE DISSERTATION

BIOGEOGRAPHICAL PATTERNS OF SOIL MICROBIAL COMMUNITIES: ECOLOGICAL, STRUCTURAL, AND FUNCTIONAL DIVERSITY AND THEIR APPLICATION TO SOIL PROVENANCE

by

Natalie Damaso

Florida International University, 2016

Miami, Florida

Professor DeEtta Mills, Major Professor

The current ecological hypothesis states that the soil type (e.g., chemical and physical properties) determines which microbes occupy a particular soil and provides the foundation for soil provenance studies. As human profiles are used to determine a match between evidence from a crime scene and a suspect, a soil microbial profile can be used to determine a match between soil found on the suspect's shoes or clothing to the soil at a crime scene. However, for a robust tool to be applied in forensic application, an understanding of the uncertainty associated with any comparisons and the parameters that can significantly influence variability in profiles needs to be determined. This study attempted to address some of the most obvious uncertainties of soil provenance applications such as spatial variability, temporal variability, and marker selection (i.e., taxa discrimination). Pattern analysis was used to validate the ecological theories driving the soil microbial biogeography. Elucidating soil microbial communities' spatial and temporal variability is critical to improve our understanding of the factors regulating their structure and function. Microbial profiling and bioinformatics analyses of the soil

community provided a rapid method for soil provenance that can be informative, easier to perform, and more cost effective than approaches using traditional physico-chemical data. This study also showed that stable profiles may allow comparison between evidence and a possible crime scene despite the time lapse (4 years) between sample collections, however, this is dependent on the analysis method, site, vegetation, and level of disturbance. Marker selection was also an important consideration for profiling. Even though Fungi look promising for single taxon soil discrimination, the additional markers can help discriminate between a wide variety of soil types. As in human identification, the more DNA markers queried the greater the discrimination power. Lastly, this study illustrated a novel method to query the iron relating genes and ability to design a novel marker that can easily be used to profile the functional diversity of a soil community to enhance soil classification. Overall this research demonstrated the potential and effectiveness of using microbial DNA from soil, not just for comparison, but also for intelligence gathering to pinpoint the geographic origin of the soil.

TABLE OF CONTENT

CHAPTER PAGE

LIST OF TABLES

TABLE PAGE

LIST OF FIGURES

FIGURE PAGE

LIST OF ABREVIATIONS AND ACRONYMS

AFLP.....Amplified Fragment Length Polymorphisms

ANOSIM……Analysis of Similarity

ANOVA……...Analysis of Variance

Avg SE…….Average Standard Error

DGGE……...….Denatured Gradient Gel Electrophoresis

FAWN……Florida Automated Weather Network

GIS……Geographic Information Systems

ICP-OES……..Plasma Optical Emission Spectrometry

LH-PCR…….....Length Heterogeneity Polymerase Chain Reaction

MDS……...Multi-Dimensional Scaling

Mean Std……Mean Standardized Error

NGS……...……..Next Generation Sequencing

nMDS……Non-Metric Multidimensional Scaling

PCR……Polymerase Chain Reaction

RFUs……Relative Fluorescent Units

RMS……Root Mean Square Error

RMS std……...……...Root Mean Square Standardized

RMSE……..…….……..Root Mean Square Error

ROC AUC……...Receiver Operator Characteristics Area Under the Curve

SIMPER…….Similarity percentages

SOPs…….Standard Operating Procedures

SSCP……....Single Strand Conformation Polymorphism

SVM……...……Support Vector Machine

TOC……Total Organic Content

T-RFLP……...….Terminal Restriction Fragment Length Polymorphism

Introduction: Soil Microbial Profiling for Forensic Application-A Review

Soils are very important ecosystem components and contain a vast array of information, both abiotic and biotic, and are one of the most challenging natural environments to study, especially for microbiologists [1,2]. Microbial community interactions are complex with individuals within the functional guilds often relying on the presence and interactions of many other species [3]. Moreover, soil is known to contain more microbial species than can be detected by traditional culturing methods [4]. For example, a gram of soil can contain 10^6-10^{10} organisms [5]. However, only about one percent of soil microbes are culturable using standard laboratory cultivation methods. Therefore, the true intrinsic diversity of soil microbial communities is largely unexplored [1].

The soil communities are important drivers of ecosystem functions such as decomposition, nutrient cycling, and plant production [6-8]. Their spatial variability is often regarded as random noise [9]. However, elucidating their spatial and temporal variability is critical to improving our understanding of the factors regulating their structure and function. Studies have tried to verify the Beijerinck hypothesis that states "everything is everywhere but the environment selects" to specify which environmental factors exert the strongest influence on the microbial communities [10]. "Distance-decay" is a universal biogeographic pattern that is commonly observed with a wide variety of organisms and illustrates a decrease in community similarity with increasing geographic distance [11]. Understanding the cause of the "distance-decay" pattern is an area of intense research. Distance decay can be observed as a result of the environmental variables being spatially auto-correlated and organisms with specific niche preferences

being selected [11]. Modeling approaches can be used to study the soil processes and microbial patterns by observing the spatial and temporal distribution using abiotic and biotic information. Seasonal and temporal variability are important to understand as soil microbial communities are susceptible to seasonal fluctuations such as temperature, water content, and nutrient availability [12]. Pattern analysis can help to develop, test, and validate prominent ecological theories driving the biogeography [13]. Moreover, Beijerinck and distance decay hypotheses lay the foundation to use soil microbial communities in forensic field for intelligence purposes to predict where the soil evidence originated from geographically.

I. Forensic Applications of Soil

Soil can provide valuable corroborative evidence in forensic investigations due to its prevalence and its transferability (based on the Locard Exchange Principle: when two objects come into physical contact an exchange of material takes place [14]). Provenance and forensic investigations of soil are usually conducted by comparing questioned samples with samples of known origin to evaluate if they are significantly similar (inclusion) or different (exclusion) based on elemental [15] and physical characteristics [16]. Tests usually consist of looking at the physical properties of soil such as soil color, texture, consistency, density, porosity, and particle size [17]. A study in 1996 showed that approximately 79% of soils could be differentiated by comparing air-dried soil color to the Munsell color chart under the microscope [18]. These methods are cost-effective and non-destructive, but require some expertise in geology. However, when there are no distinguishable features observed using microscopy, more detailed methods are necessary such as elemental analysis [17]. Other analyses include rock and mineral identification

and chemical methodologies to identify unknown trace materials [19]. Moreover, no single physical analysis is sufficient to distinguish two soil samples are from the same source, as shown by a blind study that looked at independent and collective testing of three samples from four different experts. They concluded that independent interpretations were less accurate than those where multiple techniques were combined [20]. Pye & Blott (2009) supported this by also stating that elemental analysis should always be carried out in conjunction with other methods [16].

II. Microbial Community Profiling

A vast array of biotic information is associated with soils and, therefore, should provide valuable information for provenance of soil samples. The current ecological hypothesis states that the soil type (e.g., chemical and physical properties) determines which microbes occupy a particular soil [11,21,22] and provides the foundation for soil provenance studies. Therefore, soil microbial community profiling should produce a unique biotic profile at the community level and provide rapid and efficient methods to see a snapshot of the patterned diversity within the communities using expertise and instrumentation already employed in a DNA crime laboratory. Several studies to date have used microbial analysis for soil provenance using culture-independent, molecular biology techniques [17,23-26]. Before Horswell et al. in 2002, biological analysis of soils for forensic purposes was largely ignored [23]. Analysis consisted of looking at the morphology of plant material such as pollen grain, plant seeds, and fungal spores when mineral and chemical properties of soils were undistinguishable. However, methods required scanning electron microscope and transmission electron microscope, which are destructive and render the sample unusable for other analysis [17]. Recently with the

growth of molecular biology techniques, research has shown the potential of microorganisms for reliable forensic soil analysis [17].

Horswell et al. (2002) was the first to use microbial community analysis for soil forensics. Their results indicated that ecosystem profiles within the environments were significantly more similar to each other than to those from other habitats [24]. The pilot study provided preliminary data regarding the potential of the microbial community to identify which type of environment from which a soil sample may originate [27]. Further studies have shown that bacterial profiles within habitats are more similar than different environments, though there can be large spatial and temporal variability within habitats [28]. Since 2002, there have been an increasing number of studies looking at soil microbial dynamics for forensic applications. A majority of the studies explore the bacterial 16S rRNA genes using terminal restriction length fragment polymorphism (T-RFLP) [29-31] or length heterogeneity polymerase chain reaction (LH-PCR) [23]. In a study by Smalla et al. (2007), microbial metagenome profiling using T-RFLP was able to discriminate between soil types and showed higher resolution than Denatured Gradient Gel Electrophoresis (DGGE) and Single Strand Conformation Polymorphism (SSCP) methods [29]. In a study by Moreno et al. (2006), microbial community profiling using LH-PCR was concluded to be better able to discriminate between soil types with a high degree of reproducibility than elemental analysis with inductively coupled plasma optical emission spectrometry (ICP-OES) [23]. Overall, biotic results have been promising.

Microbial DNA fingerprinting has the potential to be a powerful method for forensic investigation as human DNA [30]. DNA fingerprinting techniques such as terminal restriction fragment length polymorphism (T-RFLP) and length heterogeneity

polymerase chain reaction (LH-PCR) have been widely used and validated for rapid microbial profiling for forensic applications [32-34]. The advantage of LH-PCR is that it is rapid, robust, reproducible, requires small sample size (<500mg), and can be done with equipment and expertise already found in most crime laboratories [23]. Their limitation is that the microbial diversity can be underestimated as more than one species can be represented by one peak [35]. These techniques look at the length heterogeneity instead of the sequence differences, therefore considers the community level rather than the individual organism level. Microbial communities can be complex, consisting of a wide variety of species and organisms therefore, it is impractical to use culture-based methods for species identification [25]. Analyzing the entire community enables fast and efficient ways to provide a glimpse of the diversity in the location, and develop links between community structure and the soil habitat [25,35]. As in human DNA profiling, the pattern of the amplicons (LH-PCR peaks) generated from the microbial profile can provide discrimination of samples. Soil microbial profiles have been shown to provide a unique soil fingerprint that could potentially be used as collaborative evidence to establish evidentiary relationship between suspect and crime scene as well as provide origin of the soil [24,32,36]. Furthermore, the amplicon lengths are phylogenetically relevant and can be sequenced to provide taxonomic identity if needed [33,37].

III. Microbial Profiling Effectiveness

Microbial profiling effectiveness is dependent on the uniqueness among different habitat types, level of heterogeneity within a habitat, and stochastic processes in community over time [32]. The assumption states there should be limited temporal variability as soil should not change substantially over time to allow use of pattern

modeling of community analysis for forensic application [32]. Soil communities are not static and, therefore, can fluctuate with disturbance and seasons [38]. Spatial variability has been shown to be more significant than temporal variability [30,32]. Previous studies have assessed the temporal variability of the soil however, they were restricted within one year [24,27,28]. Lenz & Foran (2010) found that known soil samples can potentially be collected well after a crime occurred without detrimental outcomes as the time/season did not have a substantial negative influence on the ability to group soils from a habitat even though they were collected throughout a one year period [28]. Horswell et al. (2002) found that samples collected eight months apart were less similar to each other than those collected at the time of original sampling; however, they still showed a high degree of similarity (70% compared to 90%) [24]. However, if archived data and training sets are to be useful, long-term temporal variability (> 2yr) should be considered. Unlike human identification, the soil environment is dynamic and changes over time. Therefore, it is important to see if meaningful comparisons and links can still be made between soil evidence deposited at the crime and archived reference data previously collected (> 2yr) from a site can still be classified [5].

Most often soil forensic analyses have exclusively looked at bacteria. However, a study by MacDonald et al. (2008) illustrated a multiplex approach that analyzed bacteria, archaea, and fungi, which led to better discrimination [30]. Bacteria provide greater resolution between two sites, however, they appear to be more susceptible to air-drying, and sensitive to dehydration pressures that lead to population shifts. Many bacteria, however, have survival mechanisms that allow for rapid adaptation such as changing allocations of osmolytes or having thicker cell walls or sporulation capabilities as often

seen in gram-positive bacteria which help survive with spurts of dry-rewetting environmental conditions [39]. Fungi are less altered by air-drying, resilient to desiccation, tolerant to wider variation of pH (i.e., persist in acidic soils), and provide discrimination between sites [40]. Lastly, Archaea are most useful to identify saline or water logged soil environments. Therefore, a multi-taxon approach will lead to better discrimination than a single taxon approach [30] and microbial community profiling has the potential for use in forensics to link soil evidence to its origin.

IV. Functional Diversity

Over the past decade, a shift in research has been observed to study the functional diversity of an ecosystem versus the structural diversity. Biodiversity is usually defined as the species abundance and richness in an environment. However, the Millennium Assessment group (http://www.millenniumassessment.org/en/index.html) termed biodiversity as the genetic diversity, species richness and abundance, and functional traits present in an ecosystem [41]. Under global threats such as climate change (drought, flooded, etc.), major alterations of ecosystems are predicted, which can lead to substantial microbial community compositional changes affecting the ecosystem functioning and biogeochemical cycles. Biodiversity has been argued to influence ecosystem stability and resilience toward stress and disturbance. However, the relationship between the biotic diversity and microbial guild function in soil is understudied [12]. Currently, two overarching hypotheses regarding ecosystem function exist: ecological equivalence and functional dissimilarity (Figure 1). The ecological equivalence hypothesis states that the microbial communities in the same environment are functionally equivalent displaying functional redundancy [42]. The hypothesis assumes that the environment impacts

function, therefore soil type drives function [42]. In contrast, functional dissimilarity assumes that the community functions are dissimilar and not attributed to the environmental conditions but rather linked to the diversity of the microbes present in the system [42]. Therefore, a key question is whether all soil organisms are important for soil functioning or only a few species are more relevant while others are redundant. In other words, some organisms excel at a particular function and are critical to the system while others are capable but perhaps less efficient in getting the job done. Therefore, loss of those that excel could impact ecosystem services in different ways: processing rates, inferior metabolic by-products for community use, etc. Understanding these relationships will increase our understanding of the sensitivity of the composition-functioning relationship under accelerated or prolonged environmental changes.

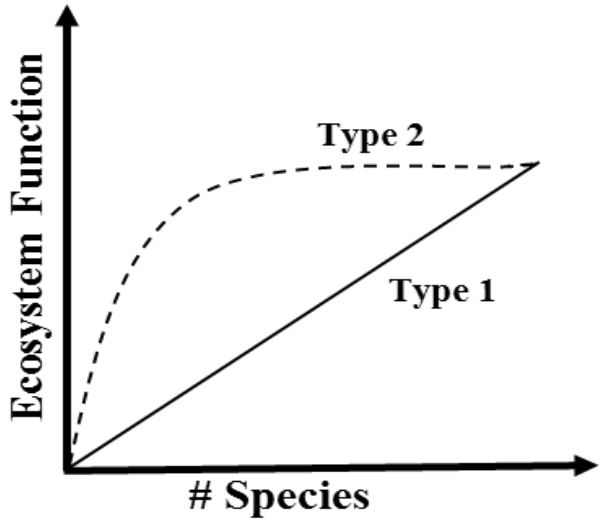

Figure 1. Species richness and ecosystem function graph adapted from Bengtsson 1998. Type 1=functional dissimilarity; Type 2=ecological equivalence. The figure depicts a hypothetical example of quantifying the relationship between ecosystem function and richness/diversity to determine the type of the relationship [43,44]. The straight line (Type 1) depicts that function only is maximized when the species diversity is maximized. Dotted line (Type 2) describes the maximium function is quickly reached

with low diversity but with critical species present and then all other species have a redundant function within the system.

The ecological equivalence hypothesis has been related to the biological insurance hypothesis, which states that redundancy within functional groups because of an increase in diversity will result in overall ecosystem performance and stability [43]. The ecological equivalence hypothesis assumes (a) that microbial communities under similar environments are more functionally similar across time; and (b) that highly diverse systems support a healthy ecosystem because many taxonomically unrelated organisms have intrinsic functional redundancy that buffer ecosystem services when environmental stress is experienced [45]. However, studies have shown that distance-decay which is commonly observed in structural genes (i.e., 16S rRNA) was not observed in a sulfate-reduction gene (*dsrA*) [46]. Therefore, more studies should be conducted to understand the regulating forces behind specific functional guilds to determine if soil type drives function or if other environmental factors (e.g., moisture) structure their biogeographical patterns. Using functional markers can be valuable to be used in forensics to discriminate soils. They can potentially reduce the complexity of assaying all bacteria that lead to high level of variability within and among habitats by profiling specific functional markers to discriminate the soils.

V. Statistical Approaches for Soil Microbial Analysis

In forensic science, the probability that a sample originated from one source rather than another selected at random must be evaluated with statistics such as the Random Match Probability or Likelihood Ratio commonly used for Human DNA profiling [47]. However, soil analysis differs from human identification as soil is not

discrete and the soil community is vulnerable to spatial and temporal variability. To date, there is no standard way to process T-RFLP or LH-PCR profiles--a standard method to quantify the calculated similarity in a forensic setting and develop a decision model to estimate evidential value of such similarities are needed [31]. Therefore, two soil samples cannot be said in the absolute sense to have originated from a single source [47,48] and it is only possible to establish a degree of probability regarding whether or not the sample derived from a given location [47]. Sorenson's similarity index has been commonly used to determine the variation within and between soil sites; however it is not optimal as the main differences between profiles may not be the presence or absence of peaks but the relative abundance (peak heights) [24]. The Bray-Curtis similarity measure is a non-parametric approach that takes into account the relative abundance to determine the similarities and differences between profiles and can be a more sophisticated approach for statistical interpretation for soil DNA profiles. Bray-Curtis similarity matrices are best suited for continuous datasets such as LH-PCR; however, negligible bias is introduced into the calculations as a result of shared absences of the amplicons [23,33]. For any statistical procedure, data transformation is often required since non-parametric and parametric tests can suffer when normality assumptions are violated. This is especially the case with microbial profiles, as true normality is rare in nature; therefore, data transformation is commonly used to improve the normality. Square-root transformation changes the values of the data points but not their rank and does not give special treatment to zeros. The square-root transformation moderates the imbalance of very abundant and rare peaks that are often observed in LH-PCR and T-RFLP profiles,

reducing the amount of noise and increasing the significant signals within the microbial community dataset [36].

Multidimensional scaling analysis (MDS) allows for a deeper inclusion of data than similarity indices as SIMPER only examines profile pairs by multiplying shared peaks by two and dividing by the total number of peaks present in both [28]. Similarity percentages can be a disadvantage in forensics, where precise origin of the sample is unknown. Multidimensional scaling generates a matrix of similarities that are weighted with peaks found in multiple habitats accentuated and background noise eliminated. The similarities are visualized in two-dimensional space resulting in an easy way to interpret profiles depicting similarities of the different soil samples [28]. In nMDS, each sample is represented as a point, and the relative interpoint distances reflect the relative dissimilarities between the sample pairs [23]. However, what do we mean by "sufficiently similar" or "sufficiently different" when we compare samples? Analysis of Similarity (ANOSIM) has been commonly used to provide a statistical significance to the dissimilarity. The best recommendation to date has been to project the unknown microbial DNA profile onto nMDS plot obtained from other localities and evaluate the similarities using ANOSIM statistics, but this is an investigative stage case [5].

Techniques for prediction and classification are developing rapidly [50]. Previous literature has discussed the usefulness of machine learning tools for classification [25,26,51]. These tools are statistical algorithms designed to study patterns in data that can then provide predictive models for the classification of unknown samples. Yang et al. (2006) illustrated the potential of supervised machine learning methods using Support Vector Machines (SVM) to classify samples using LH-PCR profiles for distinct soil use

and types. Support Vector Machines is a supervised learning method that has the ability to train on a known set of data and then be able to classify unknown soil samples to a high degree of certainty when tested against the trained set. Their results illustrated that there were a hidden pattern within the bacterial profiles that could be seen by mathematical tools [25]. Support Vector Machines, however, is not the only supervised machine learning tool that can be used for classification. Decision Trees, Random Forest, Neural Networks are other types of supervised machine learning tools. The forensic community could benefit enormously by the utilization of the classification tools and comprehensible reference database to distinguish soil samples and determine their geographic origin. Unlike nMDS that takes into account all LH-PCR peaks, of which are undoubtedly noisy, machine-learning tools throw out the noise and concentrate on those component that can define origin. Further bioinformatics trials are required to establish optimal data analysis pipeline and assess the signal to noise ratio and false positive/negative error rates [47].

The classification method, however, implies that the soil properties are discontinuous which is not correct as soil processes operate under different scales [52]. Therefore, meaningful data are usually lost during classification, which is a type of generalization that organizes the data into structural patterns to gain clarity. Soil properties have been known to vary spatially and can be related to several physical, chemical, and biological processes that act at different scales. Studies at a microscale have illustrated that soil structure and porosity as well as organic carbon content have been factors determining the soil microbes distribution while at field scale (10m-<200m), physicochemical characteristics such as texture, pH, and plant cover have been the main

factors structuring their distribution [22]. Geostatistics has been a popular method for soil science as it does not assume that the soil properties are discontinuous. Geostatistics uses the soil sample's spatial information to model spatial patterns, interpolate to unsampled locations, and assess uncertainty of the predictions [53]. Autocorrelation is at the heart of geostatistics, which is a term for spatial dependence, and queries the resemblance between "neighbors" as a function of spatial separation distance. When near neighbors are more similar than those farther away, the data are said to be autocorrelated, and therefore, violate the assumption that the data are independent [9]. The SoilFit project in the UK has used Geographic Information Systems (GIS) in soil forensics to integrate soil fingerprinting profiles with data held in spatially references soil databases to improve matching of evidentiary samples or predict provenance of soil [54].

Overall, environmental profiling has great potential to establish provenance of soil samples. Profiling involves comparing samples with those from a database to assess degree of similarity. Therefore, establishing an efficient database will aid in greater confidence of the conclusions reached [55]. There is currently no soil database to assist in interpretation of data and few attempts have been made for local forensic applications. Databases can provide useful information and assist in forensic investigation. However, degree of sample representation, the type and quality of information, discriminatory capacity, and the use of obsolete data or dynamic data need to be considered when constructing a searchable database [16]. It is also important to understand different sites spatial variability to see if different sampling designs are needed to accurately depict the soil site [56]. Previous studies have illustrated the differences of within site variability between homogeneous grassland over shrub land [57]. Local heterogeneity can be the

result of different soil properties and multiple environmental factors such as unique plant species, sunlight amount, and differing moisture content [28,58]. Geographic Information Systems' semivariograms can provide a useful tool for designing robust sampling strategies by estimating the variance (sill) that can be used to inform sample size in future studies as well as estimate the minimum distance required for samples to be considered spatially independent (range) that can be used to inform sample spacing to build a robust database for soil provenance. Further research is needed to understand the number of samples needed to represent the population and the discriminatory capacity to determine if one test fits all or if the model needs to be tuned to fit particular soil types or geographic situations.

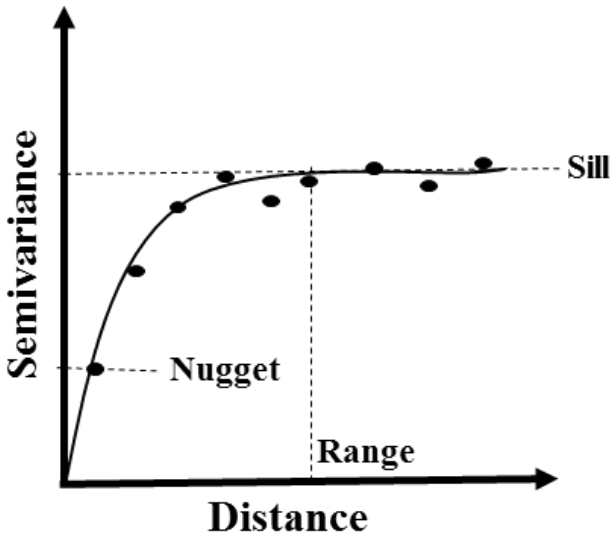

Figure 2. The semivariograms show the hypothetically observed distance class (filled circles) and the fitted model (solid line). The theoretical semivariogram model fitting is usually expressed by three parameters: nugget, sill, and range. The nugget represents the measurement errors or spatial dependence at scales not explicitly sampled. The sill represents the variance of the correlated measurements. The range shows the extent of heterogeneity (i.e., zone of influence or distance of dependence) [9,53].

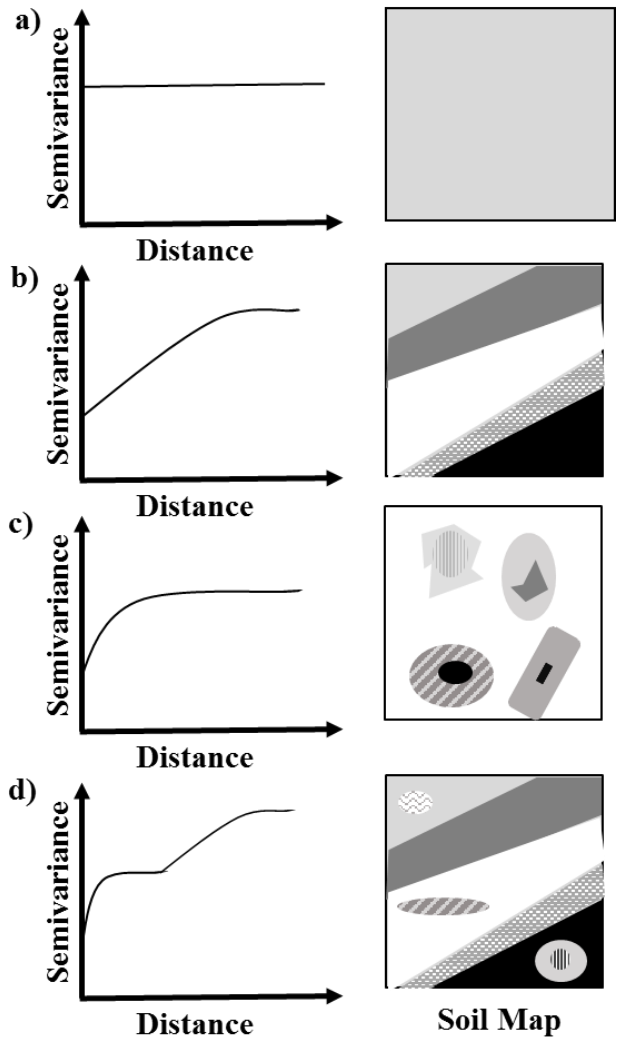

Figure 3. Plots illustrate different hypothetical semivariograms with their associated surface maps adapted from Ettema & Wardle (2002). A) pure nugget effect: no spatial structure was observed at the spatial extent studied. This can occur as a result of random sampling variance or variability that is occurring at other spatial scales not examined in the spatial extent. B) Large-scale heterogeneity: few, large and smoothly continuous gradients. C) Small-scale heterogeneity: many, small, sharply discontinuous patches. D) Nested heterogeneity: multiple scales of patchiness where more than one factor is influencing the pattern at different scales [9].

VI. Validation of Soil Analysis

Soil DNA profiling has great potential as a forensic tool and research to date has been promising. However, for microbial soil fingerprinting to be forensically useful, optimization and standardization needs to be conducted [36]. Research examining the

ability to discriminate soil samples and their limitations are needed [59]. Impact of abiotic conditions (moisture and organic content) as well as seasonal and temporal affects are critical to establish the robustness of this method in practice, furthermore determining the most reliable and robust target taxa or multiple taxa is important. Analytical approaches to microbial community profiling needs to combine discriminatory power, robustness and, reliability as well as statistical methods must be identified to provide objective measures for assessing the similarities/differences between samples [5]. Like human DNA fingerprinting, a validated statistical method to provide definite proof linking suspects or victim to crime scene based on the soil microbial fingerprint is needed [36]. For a robust tool to be applied in forensic application, an understanding of the uncertainty associated with any comparisons and the parameters that can significantly influence variability in profiles needs to be determined. These issues include selecting suitable microbial markers and the influence of temporal variability on the DNA profile.

Soil analysis can be time consuming and complicated as the techniques vary [18]. Also there is no single application or set of techniques that are suitable for all circumstances and to date there is no standard forensic soil examination method [14]. Therefore, standardization and validation for forensic soil analyses are required. Validation is a common process in forensic science to generate reliable, robust, confident, and discriminatory power analyses. The validation process determines the conditions required to obtain results, limitations of the methods, areas that need to be monitored and controlled, and interpretation guidelines to express significance. Validation of methods of collection, preservation, extraction, analysis, and interpretation are required to document their specificity, sensitivity, reproducibility, bias, precision, false

positives/negatives, appropriate controls, and interpretative thresholds [60]. Therefore, there is a need for standard operating procedures from the collection to the interpretation of microbial forensic analysis to be accepted in the court of law [38].

VII. Objectives of study

In this study, modeling approaches will be used to study the microbial patterns and drivers of the variability by observing the spatial and temporal distribution of microbes using abiotic and biotic information. Soil biotic content at both the structural and functional level will be assessed. Pattern analysis will be used to validate the ecological theories driving the soil microbial biogeography. Moreover, the spatio-temporal variability of the soil will be observed to determine the usefulness for soil provenance studies. Bioinformatic tools and Geographic Information Systems will be used to determine if soil biotic profiles can be used to classify soil on the basis of soil type or location and their ability interpolate to un-sampled locations. Four taxa will be observed together and separately to determine their discrimination power for soil classification. Lastly, functional diversity profiles using iron cycling genes will be assessed to determine if soil type drives function and if the addition of these data can enhance soil classification.

A. AIM 1: Comparison of machine learning algorithms for the classification and provenance of soil samples using biotic content

Hypothesis: Soil microbial communities exhibit biogeographical patterns based on the soil type and therefore, they can be used for soil provenance applications as these patterns are predictable.

The first aim was to first determine if the soil samples collected were spatially correlated

to their respective geographic locations. Analysis of both the geographic location and genetic profiles of the soils permit the evaluation of the hypothesis that geographically closer samples will display similar microbial profiles than those farther apart. Secondly, five supervised machine learning algorithms were evaluated based on their accuracy in classification of samples at different spatial scales (soil types, transects, and subplot) to determine the potential to use soil microbial profiles and bioinformatics tools for determination of soil origin.

B. AIM 2: Geographic Information Systems approach to characterize the spatial variability of the soil microbial community and the application to forensics

Hypothesis: Geographic Information Systems' semivariograms can provide a useful tool for designing robust sampling strategies to build a microbial community database for forensic provenance applications.

The second aim was to examine the organization of the microbial community structure at multiple spatial scales across Miami-Dade, Florida using multivariate statistics and geostatistics to observe patterns as a function of distance. Geographic Information Systems provides a useful tool that can be used to inform sample strategies to build a robust database for soil provenance applications. Geographic Information Systems' semivariograms can provide a useful tool for designing robust sampling strategies by estimating the variance (sill) between sampling points as well as estimate the minimum distance required for samples to be considered spatially independent (range).

C. AIM 3: Assessing temporal variability and DNA marker selection for forensic soil provenance applications

Marker Hypothesis: The more molecular markers queried the greater the discrimination

power therefore, the four taxa approach will provide the highest degree of discrimination between and within sites.

Temporal Hypothesis: Soil microbial communities will display limited temporal variability over a four-year time span.

The third aim was to determine the microbial community effectiveness by determining the marker discrimination as well as the temporal variability. Bacteria, fungi, archaea, and plant community profiles will be assessed independently and combined to determine the best marker or markers for forensic comparison of soil evidence. Secondly, the temporal variability of the soil microbial community will be assessed after a four year time span (2010 and 2014). This is vital as microbial communities should be stable enough to be able to use for forensic purposes over a reasonable time span.

D. AIM 4: Analysis of the microbial functional diversity using iron genes across different soil types in Miami-Dade County, FL

Hypothesis: Soil iron and moisture content are responsible for the microbial community's functional guild (biogeographical) distribution and therefore, the functional diversity profiles can supplement other biotic information for soil provenance applications.

The fourth aim was to use GeoChip microarray to query the iron cycling genes across different soil types in Miami-Dade, FL to determine the distribution of their structural and functional diversity. One of the discriminatory iron genes (*feoB*) detected in GeoChip was used to design a novel degenerate primer that can be used to make functional diversity profiles to determine if it adds to the discrimination for soil provenance.

VIII. References

[1] R. Daniel, The metagenomics of soil, Nature Reviews Microbiology. 3 (2005) 470-478.

[2] S. Mocali, A. Benedetti, Exploring research frontiers in microbiology: the challenge of metagenomics in soil microbiology, Research in Microbiology. 161 (2010) 497-505.

[3] G.T. Hill, N.A. Mitkowski, L. Aldrich-Wolfe, L.R. Emele, D.D. Jurkonie, A. Ficke, S. Maldonado-Ramirez, S.T. Lynch, E.B. Nelson, Methods for assessing the composition and diversity of soil microbial communities, Applied Soil Ecology. 15 (2000) 25-36.

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

[5] G.F. Sensabaugh, Microbial community profiling for the characterisation of soil evidence: forensic considerations, Criminal and Environmental Soil Forensics, Springer, 2009, pp. 49-60.

[6] R.D. Bardgett, W.D. Bowman, R. Kaufmann, S.K. Schmidt, A temporal approach to linking aboveground and belowground ecology, Trends in Ecology & Evolution. 20 (2005) 634-641.

[7] C. Emmerling, M. Schloter, A. Hartmann, E. Kandeler, Functional diversity of soil organisms-a review of recent research activities in Germany, Journal of Plant Nutrition and Soil Science. 165 (2002) 408.

[8] P. Garbeva, J.A. Van Veen, J.D. Van Elsas, Microbial diversity in soil: selection microbial populations by plant and soil type and implications for disease suppressiveness. Annual Review Phytopathology. 42 (2004) 243-270.

[9] C.H. Ettema, D.A. Wardle, Spatial soil ecology, Trends in Ecology & Evolution. 17 (2002) 177-183.

[10] N. Fierer, R.B. Jackson, The diversity and biogeography of soil bacterial communities. Proceedings of the National Academy of Sciences. 103 (2006) 626-631.

[11] J.B. Martiny, J.A. Eisen, K. Penn, S.D. Allison, M.C. Horner-Devine, Drivers of bacterial beta-diversity depend on spatial scale, Proceedings of the National Academy of Sciences 108 (2011) 7850-7854.

[12] V. Torsvik, L. Øvreås, Microbial diversity and function in soil: from genes to ecosystems, Current Opinion in Microbiology. 5 (2002) 240-245.

[13] C. Violle, P.B. Reich, S.W. Pacala, B.J. Enquist, J. Kattge, The emergence and promise of functional biogeography, Proceedings of the National Academy of Sciences. 111 (2014) 13690-13696.

[14] R.W. Fitzpatrick, M.D. Raven, S.T. Forrester, A systematic approach to soil forensics: criminal case studies involving transference from crime scene to forensic evidence, Criminal and Environmental Soil Forensics, Springer, 2009, pp. 105-127.

[15] S.C. Jantzi, J.R. Almirall, Characterization and forensic analysis of soil samples using laser-induced breakdown spectroscopy (LIBS), Analytical and Bioanalytical Chemistry. 400 (2011) 3341-3351.

[16] K. Pye, S.J. Blott, Development of a searchable major and trace element database for use in forensic soil comparisons, Science & Justice. 49 (2009) 170-181.

[17] S.A. Larson, Developing a high throughput protocol for using soil molecular biology as trace evidence, (2012). Theses and Dissertations in Biochemistry.

[18] R. Sugita, Validity of color examination for forensic soil identification, Forensic Science International. 83 (1996) 201; 201-210; 210.

[19] R.C. Murray, Forensic Examination of Soils, Forensic Chemistry Handbook. (2012) 109-130.

[20] B.G. Rawlins, S.J. Kemp, E.H. Hodgkinson, J.B. Riding, C.H. Vane, C. Poulton, K. Freeborough, Potential and pitfalls in establishing the provenance of Earth-related samples in forensic investigations, Journal of Forensic Science. 51 (2006) 832-845.

[21] L. Moreno, B. McCord, Separation of DNA for forensic applications using capillary electrophoresis, in: Landers J.P. (Ed.), Handbook of Capillary and Microchip Electrophoresis and Associated Microtechniques, CRC Press, Boca Raton, FL, 2007, pp. 761-784.

[22] P. Maron, C. Mougel, L. Ranjard, Soil microbial diversity: methodological strategy, spatial overview and functional interest, Comptes Rendus Biologies. 334 (2011) 403-411.

[23] L.I. Moreno, D.K. Mills, J. Entry, R.T. Sautter, K. Mathee, Microbial metagenome profiling using amplicon length heterogeneity-polymerase chain reaction proves more effective than elemental analysis in discriminating soil specimens, Journal of Forensic Science. 51 (2006) 1315-1322.

[24] J. Horswell, S.J. Cordiner, E.W. Maas, T.M. Martin, K.B.W. Sutherland, T.W. Speir, B. Nogales, M. Osborn, Forensic comparison of soils by bacterial community DNA profiling, Journal of Forensic Science. 47 (2002) 350-353.

[25] C. Yang, D. Mills, K. Mathee, Y. Wang, K. Jayachandran, M. Sikaroodi, P. Gillevet, J. Entry, G. Narasimhan, An ecoinformatics tool for microbial community studies: Supervised classification of Amplicon Length Heterogeneity (ALH) profiles of 16S rRNA, Journal of Microbiological Methods. 65 (2006) 49-62.

[26] J.A. Entry, D. Mills, K. Mathee, K. Jayachandran, R. Sojka, G. Narasimhan, Influence of irrigated agriculture on soil microbial diversity, Applied Soil Ecology. 40 (2008) 146-154.

[27] L.E. Heath, V.A. Saunders, Assessing the potential of bacterial DNA profiling for forensic soil comparisons, Journal of Forensic Science. 51 (2006) 1062-1068.

[28] E.J. Lenz, D.R. Foran, Bacterial profiling of soil using genus-specific markers and multidimensional scaling, Journal of Forensic Science. 55 (2010) 1437-1442.

[29] K. Smalla, M. Oros-Sichler, A. Milling, H. Heuer, S. Baumgarte, R. Becker, G. Neuber, S. Kropf, A. Ulrich, C.C. Tebbe, Bacterial diversity of soils assessed by DGGE, T-RFLP and SSCP fingerprints of PCR-amplified 16S rRNA gene fragments: Do the different methods provide similar results? Journal of Microbiological Methods. 69 (2007) 470-479.

[30] L.M. Macdonald, B.K. Singh, N. Thomas, M.J. Brewer, C.D. Campbell, L.A. Dawson, Microbial DNA profiling by multiplex terminal restriction fragment length polymorphism for forensic comparison of soil and the influence of sample condition, Journal of Applied.Microbiology. 105 (2008) 813-821.

[31] F.C.A. Quaak, I. Kuiper, Statistical data analysis of bacterial t-RFLP profiles in forensic soil comparisons, Forensic Science International. 210 (2011) 96-101.

[32] M.S. Meyers, D.R. Foran, Spatial and temporal influences on bacterial profiling of forensic soil samples, Forensic Science International. 53 (2008) 652-660.

[33] D.K. Mills, J.A. Entry, J.D. Voss, P.M. Gillevet, K. Mathee, An assessment of the hypervariable domains of the 16S rRNA genes for their value in determining microbial community diversity: the paradox of traditional ecological indices, FEMS Microbiology Ecology. 57 (2006) 496-503.

[34] D.K. Mills, J.A. Entry, P.M. Gillevet, K. Mathee, Assessing microbial community diversity using amplicon length heterogeneity polymerase chain reaction, Soil Science Society of America Journal. 71 (2007) 572-578.

[35] M.J. Johnson, K.Y. Lee, K.M. Scow, DNA fingerprinting reveals links among agricultural crops, soil properties, and the composition of soil microbial communities, Geoderma. 114 (2003) 279-303.

[36] Z. Pasternak, A. Al-Ashhab, J. Gatica, R. Gafny, S. Avraham, S. Frenk, D. Minz, O. Gillor, E. Jurkevitch, Optimization of molecular methods and statistical procedures for forensic fingerprinting of microbial soil communities, International Research Journal of Microbiology. 3 (2012) 363-372.

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

[38] A. Gunn, S.J. Pitt, Review Paper Microbes as forensic indicators, Tropical Biomedicine. 29 (2012) 311-330.

[39] J. Schimel, T.C. Balser, M. Wallenstein, Microbial stress-response physiology and its implications for ecosystem function, Ecology. 88 (2007) 1386-1394.

[40] G.A. Evdokimova, N.P. Mozgova, The effect of drying of soil samples on the number of bacteria and fungi, Eurasian Soil Science. 36 (2003) 546-549.

[41] S. Naeem, Biodiversity, ecosystem functioning, and human wellbeing: an ecological and economic perspective, Oxford University Press, Oxford; New York, 2009.

[42] M.S. Strickland, C. Lauber, N. Fierer, M.A. Bradford, Testing the functional significance of microbial community composition, Ecology. 90 (2009) 441-451.

[43] D.C. Coleman, W.B. Whitman, Linking species richness, biodiversity and ecosystem function in soil systems, Pedobiologia. 49 (2005) 479-497.

[44] J. Bengtsson, Which species? What kind of diversity? Which ecosystem function? Some problems in studies of relations between biodiversity and ecosystem function, Applied Soil Ecology. 10 (1998) 191-199.

[45] S.D. Allison, J.B. Martiny, Colloquium paper: resistance, resilience, and redundancy in microbial communities, Proceedings of the National Academy of Sciences. 105 Suppl 1 (2008) 11512-11519.

[46] A. Angermeyer, S.C. Crosby, J.A. Huber, Decoupled distance–decay patterns between *dsrA* and 16S rRNA genes among salt marsh sulfate-reducing bacteria, Environmental Microbiology. (2015).

[47] J.M. Young, L.S. Weyrich, J. Breen, L.M. Macdonald, A. Cooper, Predicting the origin of soil evidence: High throughput eukaryote sequencing and MIR spectroscopy applied to a crime scene scenario, Forensic Science International. 251 (2015) 22-31.

[48] R.C. Murray, J.C. Tedrow, Forensic geology, Prentice-Hall 1991.

[49] F. Wickelmaier, An introduction to MDS, Sound Quality Research Unit, Aalborg University, Denmark. 46 (2003).

[50] M. Kovacevic, B. Bajat, B. Gajic, Soil type classification and estimation of soil properties using support vector machines, Geoderma. 154 (2009) 340.

[51] C. Kampichler, R. Wieland, S. Calmé, H. Weissenberger, S. Arriaga-Weiss, Classification in conservation biology: A comparison of five machine-learning methods, Ecological Informatics. 5 (2010) 441-450.

[52] J. McKinley, How useful are databases in environmental and criminal forensics? Geological Society, London, Special Publications. 384 (2013) 109-119.

[53] P. Goovaerts, Geostatistical tools for characterizing the spatial variability of microbiological and physico-chemical soil properties, Biology and Fertility of Soils. 27 (1998) 315-334.

[54] E. Nissan, The Forensic Disciplines: Some Areas of Actual or Potential Application, Computer Applications for Handling Legal Evidence, Police Investigation and Case Argumentation, Springer, 2012, pp. 841-989.

[55] K. Pye, D.J. Croft, Forensic geoscience: introduction and overview, Geological Society, London, Special Publications. 232 (2004) 1-5.

[56] H. Loescher, E. Ayres, P. Duffy, H. Luo, M. Brunke, Spatial variation in soil properties among North American ecosystems and guidelines for sampling designs, PloS One. 9 (2014) e83216.

[57] D.L. Mummey, P.D. Stahl, Spatial and temporal variability of bacterial 16S rDNA-based T-RFLP patterns derived from soil of two Wyoming grassland ecosystems, FEMS Microbiology Ecology.46 (2003) 113-120.

[58] R.B. Franklin, A.L. Mills, Multi-scale variation in spatial heterogeneity for microbial community structure in an eastern Virginia agricultural field, FEMS Microbiology Ecology. 44 (2003) 335-346.

[59] J.M. Young, L.S. Weyrich, A. Cooper, Forensic soil DNA analysis using high-throughput sequencing: a comparison of four molecular markers, Forensic Science International: Genetics. 13 (2014) 176-184.

[60] B. Budowle, S.E. Schutzer, S.A. Morse, K.F. Martinez, R. Chakraborty, B.L. Marrone, S.L. Messenger, R.S. Murch, P.J. Jackson, P. Williamson, R. Harmon, S.P. Velsko, Criteria for validation of methods in microbial forensics, Applied and Environmental Microbiology. 74 (2008) 5599.

Chapter 1: A Comparison of Machine Learning Algorithms for the Classification and Provenance of Soil Samples Using Biotic Content

This chapter is currently under review in Forensic Science International.

Natalie Damaso[1,2], Julian Mendel[1,2], Maria Mendoza[1,2], Yanie Oliva[1], Ashley Diaz[2], Eric J von Wettberg[1,3], Giri Narasimhan[4], and DeEtta Mills[1,2*]

[1]Department of Biological Sciences, Florida International University, Miami, Florida, United States of America

[2]International Forensic Research Institute, Florida International University, Miami, Florida, United States of America

[3]International Center for Tropical Botany, Florida International University, Miami, Florida, United States of America

[4]Bioinformatics Research Group (BioRG), School of Computing and Information Sciences, and Biomolecular Sciences Institute, Florida International University, Miami, Florida, United States of America

I. Abstract

Soil type (i.e., elemental composition, chemical/physical properties) is often correlated with the microbial community that inhabits it and the food web it supports. Therefore, soil metagenomic profiling should produce a distinguishable biotic profile from a specific soil type and location. Additional bioinformatic analyses of the soil community can provide a rapid method for soil provenance that can be informative, easier to perform and more cost effective than approaches using physico-chemical data. However, the intrinsic spatio-temporal heterogeneity of soil needs to be considered in community analyses for forensic applications. The objective of this study was to compare five machine learning tools for their predictive ability to recognize biotic patterns for rapid classification of soils at different spatial scales. Metagenomic DNA was extracted from 1268 soil samples that represent the six soil types in Miami-Dade County, FL. Bacteria, archaea, fungi, and plant universal DNA markers were amplified, separated by capillary electrophoresis and profiled. Autocorrelations were conducted using Mantel tests which linked metagenomic content to soil type as well as to specific transects within a soil type with strong accuracy. Seasonal changes (wet and dry) did not reduce the correlation; however, soil disturbance did. Five machine learning tools were employed for soil classification at different spatial scales: K-Nearest Neighbor, Decision Trees, Random Forests, Neural Networks, and Support Vector Machines. Of all of these tools, Random Forests had higher accuracy than the others, and were able to accurately classify soils at the level of soil type, transect, and subplot. These methods illustrated the potential of using soil metagenomic profiles and bioinformatic tools for soil provenance testing.

II. Introduction

Soil can provide valuable information as evidence in forensic investigations; its value is associated with its prevalence and transferability. Several forensic soil studies have used microbial analyses for soil provenance using culture-independent, molecular biology techniques [1-6]. Previous studies have shown the enormous potential of methods that use microbial community profiling to determine soil origin of an unknown sample. However, spatial distribution and sensitivity of the analysis method to detect differences in microbial communities from similar soil types (i.e., similar physical and chemical properties) and local scales (i.e., similar location) have to be investigated before being applied in the forensic context [7]. Currently, knowledge of the spatial and temporal distribution of the microbial communities at multiple scales is lacking. The effectiveness of using microbial community profiling to differentiate forensic soil samples depends on the existence of a quantitative measure or method (a) that helps distinguish soils from different types of habitats, (b) that soils exhibit spatial autocorrelation, and (c) that remains relatively stable within limited temporal scales [8]. In the present study, soil using four-taxa profiles (bacteria, archaea, fungi, plant) at multiple spatial scales—soil type, transect (>1.6 km apart), subplot level (within 1 m), and over one year period (seasons-dry and wet)—were conducted to determine the effectiveness of this method for forensic applications.

Machine learning tools have been used for pattern discovery, classification, and prediction and many studies have indicated the usefulness of these different algorithms for classification [4,5,9]. These tools are statistical algorithms designed to study patterns in the data that can provide predictive models for the classification of unknown samples.

Machine learning tools are separated into two categories: supervised and unsupervised. Supervised learning involves using a training set to build a model of causation for the desired classification, whereas unsupervised learning does not make such assumptions and attempts to discover patterns and structures in the data without a training set [10]. In the present study, community DNA profiles using universal primers for four taxa were used to generate data to evaluate five different machine-learning algorithms for their ability to determine soil provenance: K-Nearest Neighbor, Decision Trees, Random Forests, Neural Networks, and Support Vector Machines.

These machine-learning tools have been used previously to discriminate bacterial communities in different microbiomes. For example, Yang et al. (2006) used Support Vector Machines and K-Nearest Neighbor to classify samples using length heterogeneity PCR 16S bacterial profiles across distinct soil types under different agricultural use [4]**.** Beck and Foster (2014) used Logistic Regression, Genetic Programming, and Random Forest to classify bacterial vaginosis characteristics from female microbiomes [11]. However, studies have shown that no single classification method is superior in every case [9,12,13]. For example, Kampichler et al. (2010) showed that modelling methods for one dataset might not be optimal for another [9]. A similar conclusion was arrived at by Tan & Gilbert (2003) [13]. Each classification tool has its own learning and prediction procedure and differs in complexity and computation time. For example, Neural Networks and Support Vector Machines are more complex as compared to Decision Trees, Random Forests, and K-Nearest Neighbors. Recent studies have used bacterial metagenomics data and machine learning algorithms but none have compared five different algorithms for their ability to accurately classify soils using four-taxa profiles at

different spatial scales. Therefore, the current study compared the five machine learning tools to determine which tool was the best to predict provenance of soil using four-taxa (bacteria, archaea, fungi, plant) profiles at different spatial scales. The profiles can be quickly and easily generated and interpreted without the need for data analysis pipelines for complex data analyses such as that needed for metagenomic analyses. For forensic provenance applications, the technique should provide rapid analyses, be reproducible, have a high degree of classification accuracy, and be easily interpretable for implementation in a court of law—all the attributes satisfied by the approaches used in this study.

III. Materials and Methods

A. Soil Collection

Soil samples (N = 1268) were collected across Miami-Dade County, Florida. Given that the collections were made from public access sites and did not involve endangered or protected species, no special permits were required. Six soil types with 2-4 transects within each were surveyed. Each transect was at least 1.6 km away from each other, transects were 100 m in length and within each transect, six subplots were randomly selected. Within each subplot, six cored samples were taken within a 1.0 m^2 quadrat. A five-centimeter diameter soil corer was used to collect the top 5-10 cm of the soil (Figure 4). Samples were collected during one year, with one transect (FIU) collected over a 1.5 year period. In South Florida with its monsoonal subtropical climate, sampling was repeated at the same sites during both the dry and wet seasons (and dry-wet-dry in the 1.5 year sequence). Most transects were established in undisturbed sites that had limited public access. The soils were labeled as one of six different soil types

according to USDA soil surveys [14]: 1-Urban Land-Udorthents, 2- Lauderhill Dania-Pahokee, 3- Rock Outcrop-Biscayne-Chekika, 4- Perrine-Biscayne-Pennsuco, 5- Krome Association, 6- Perrine-Terra Ceia-Pennsuco. Global Positioning System (GPS) coordinates for each subplot for each transect were recorded. Wet and dry seasons were defined by Florida Automated Weather Network (FAWN, http://fawn.ifas.ufl.edu) where seasons in Florida are classified based on the average rainfall. The wet season is defined as the period during which the average rainfall is four times more than that in the corresponding dry season. The wet season occurs from May-October, while the dry season is from November-April [15].

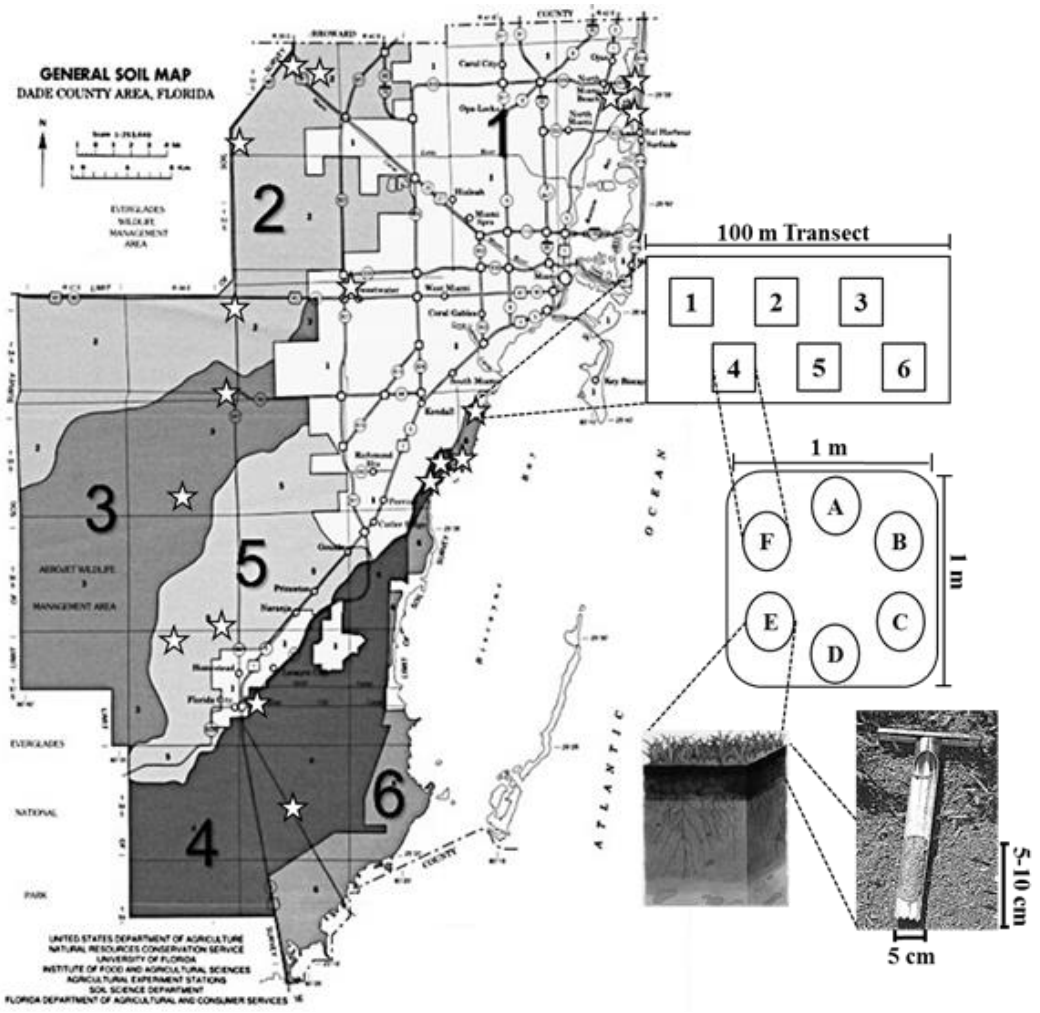

Figure 4. Map of Miami-Dade County, FL. Shaded areas represent the six soil types of Miami-Dade according to USDA: 1-Urban Land-Udorthents, 2-Lauderhill Dania-Pahokee, 3-Rock Outcrop-Biscayne-Chekika, 4-Perrine-Biscayne-Pennsuco, 5-Krome Association, 6-Perrine-Terra Ceia-Pennsuco [14]. Stars indicate transect sites. Within each 100 m transect, six subplots were sampled and six cored samples were taken within a 1.0 m^2 quadrat from each subplot. A five-centimeter diameter soil corer was used to collect the top 5-10 centimeters of the soil.

B. DNA Extraction

The soil samples were transported back to the laboratory on ice, manually homogenized, and sieved to remove large objects and debris. The DNA was extracted using the BIO 101 Fast DNA Spin Kit for Soil® and FastPrep®-24 System homogenizer (MP Bio, Solon, OH). The Fluorescent DNA Quantitation Kit (Bio-Rad, Berkeley, CA) and Modulus™ Microplate Multimode Reader (Turner Biosystems, Sunnyvale, CA) were used to quantitate the extracted metagenomic DNA. Samples were diluted to a working stock of 20 ng/µl. Lastly, a 1% agarose yield gel was run to assess the integrity and quality of the extracted DNA.

C. Length Heterogeneity Polymerase Chain Reaction

The DNA was amplified using Length Heterogeneity Polymerase Chain Reaction (LH-PCR) using two PCR duplexes: (1) bacteria and fungi, and (2) plant and Archaea. Universal primers for the following genomic regions for each taxa were used: 16S rRNA for bacteria (27-F, 355-R) [16] and Archaea (21-F, 518-R) [17,18], ribosomal internal transcribed spacer region (ITS) for fungi (ITS5-F, ITS2-R) [19], and chloroplast *trn*l intergenic region for plant (trnL-F, trnL-R) [20]. Forward primers were labeled with 6-FAM fluorescent dye. PCR reaction mixtures were: 1X reaction buffer, 2.5 mM $MgCl_2$, 250 µM dNTPs (Promega, Madison, WI), 1% BSA (fraction V, Fisher Scientific, Pittsburgh, PA), 1% DMSO (Promega, Madison, WI), various concentrations of primers

(bacteria, 0.5 µM; archaea and fungi, 0.4 µM; plant, 0.6 µM), 40 ng DNA, 0.5 U AmpliTaq Gold® DNA Polymerase (Applied Biosystems, Foster City, CA), and diethylpyrocarbonate-treated (DEPC) water to a final volume of 20 µl. Each duplex was amplified with the same program using the ABI 9700™ thermocycler (Applied Biosystems, Foster City, CA) with the following parameters: initial 10 min denaturing step at 95°C, 25 cycles of denaturation at 95°C annealing at 54°C and extension at 74°C each for 30 sec with a final extension at 74°C for 10 min.

D. Capillary Electrophoresis

Samples from the two duplexes were co-loaded where 0.5 µl of each duplex PCR product was added to a mixture of 11.5 µl Hi-Di™ Formamide (Applied Biosystems, Foster City, CA) and 0.65 µl internal size standard, GeneScan LIZ® 600 (Applied Biosystems, Foster City, CA), denatured by heating for 2 min at 95°C and then snap-cooled on ice for 2 min. The amplicon data were analyzed using the DS-33 matrix and filter set G (Applied Biosystems, Foster City, CA). The samples were electrokinetically injected at 15 kV for five sec and separated at 60°C on an ABI Prism™ 310 (Applied Biosystems, Foster City, CA) using Performance Optimized Polymer 4 (POP4) (Applied Biosystems, Foster City, CA) with laser power at 9.9 mW and capillary length of 36 cm well to read (WTR) distance to the detection window.

E. Analyses

Raw data were analyzed using the GeneMapper™ research software, version 4.0 (Applied Biosystems, Foster City, CA). Local Southern size calling was used for the analysis parameters with a minimum threshold of 50 relative fluorescent units (RFUs). Bins were created to separate amplicons that differed from each other in length by a

single base pair. The relative ratios were calculated by normalizing the heights of each peak in the genotype to the total peak intensities resulting in the ratio for each peak height as a decimal value from zero to one. Data from all taxa were concatenated for subsequent analyses. Relative GPS coordinates were taken for each sample by making the center of the subplot as the true GPS coordinate.

F. Mantel Test

Mantel tests were performed using R programming language with the ade4 library [21]. Two distance matrices were tested: geographic distance and genetic distance with data imported as binary data (presence/absence). The Mantel tests were performed and plotted using the function mantel.randtest in the ade4 package and calculated based on the random permutation using the Monte Carlo method. The method relies on repeated random sampling (using 999 permutations) to compute the results so that no assumptions regarding the statistical distributions of samples in the matrix were needed. The rows and columns of one matrix were randomly permutated followed by recalculation of the correlation after each permutation, thereby testing the significance. Detailed script can be found in supporting information (S1 File).

G. Machine Learning Tools

The R software package was used for all the classification methods; the specific R packages used for these methods include: class for K-Nearest Neighbor [22], rpart for Decision Trees [23], randomForest for Random Forests [24], neuralnet for Neural Networks [25], and e1071 for Support Vector Machines [26]. Detailed scripts can be found in supporting information (S2 File). Two thirds of the dataset was used for training and one third was used for testing for each replicate run and across each algorithm. For

comparison and reproducibility, the datasets were re-tested by randomly selecting a different training and testing set, three different times. Comparisons of the methods were conducted by calculating the percent of samples correctly classified into soil type, transect, or subplot for each test set. Second performance criterion evaluated was the area under an ROC curve (AUC), which is a widely used measure of performance for supervised classification methods using their ranking quality of sensitivity (true positive rate) as a function of the specificity (false positive rate). An AUC value of 1 illustrates a perfect test that has zero false positives and zero false negatives. Multi-class AUC was conducted using the pROC package in R [27]. Random Forest and Support Vector Machines were re-evaluated using different minimum ratio thresholds for the electrophoretic data (1%, 5%, 10%, and 20%) to check if the classification accuracy changed. Student two-sample T-tests were conducted to determine significant differences between different classification scales and machine learning tools. Random Forest analysis was conducted to provide the most important variables for classification for each spatial scale-soil type, transect, and subplot.

H. Similarity Percentages

A SIMPER analysis using Primer-E v.7 was conducted to determine the percent dissimilarity within and between samples at multiple spatial scales-soil type, transect, subplot- and seasonal differences (wet, dry). The analysis was also conducted to identify unique LH-PCR peaks contributing to dissimilarity between sites and was compared to the Random Forest important variable plot that illustrated the significant LH-PCR peaks to discriminate between samples.

IV. Results

A. Spatial Autocorrelation Analysis: Mantel Test

The genetic profiles from most sites in Miami-Dade displayed a significant positive spatial autocorrelation between its geographic location and biotic composition illustrating that samples that were geographically closer together were statistically similar in their biotic composition. Out of eighteen transects, six transects had non-significant correlations (Table 1). These six sites were found to have been previously disturbed (e.g., fire, illegal dumping, agricultural disturbance). These sites included: OSP1 (ob = -0.04, p = 0.58) from soil type 1, CH (ob = 0.03, p = 0.29) from soil type 3, PE (ob = -0.17, p = 0.99) from soil type 4, and USDA 1 (ob = -0.05, p = 0.73) from soil type 5 during the wet season as well as OSP1 (ob = 0.09, p = 0.15) from soil type 1, OSP2 (ob = -0.08, p = 0.81) from soil type 1, NW 137 (ob = -0.01, p = 0.42) from soil type 2, and USDA 1 (ob = -0.17, p = 0.98) from soil type 5 for the dry season.

Table 1. The Mantel test results for all of Miami-Dade County's six soil types, transects within each soil type, for each season (wet and dry). Numbers in parentheses are p values.

Soils	Wet	Dry
1 - FIU	0.43 (0.001)	0.18 (0.001)
1 – OSP1	**- 0.04 (0.583)**	**0.09 (0.148)**
1 – OSP2	0.24 (0.003)	**- 0.08 (0.812)**
1 - OSP3	0.19 (0.011)	0.22 (0.006)
2 - CC6	0.29 (0.001)	0.55 (0.001)
2 - KNT	0.43 (0.001)	0.16 (0.023)
2 - KS8	0.41 (0.001)	0.53 (0.001)
2 - NW137	0.25 (0.001)	**- 0.01 (0.424)**
3 - CH	**0.03 (0.285)**	0.42 (0.001)
3 - KK	0.35 (0.001)	0.46 (0.001)
4 - CS	0.17 (0.008)	0.35 (0.001)
4 - PE	**- 0.17 (0.993)**	0.22 (0.002)
5 - HA	0.46 (0.001)	0.13 (0.001)
5 - TREC	0.21 (0.006)	0.16 (0.011)
5 - USDA1	**- 0.05 (0.725)**	**- 0.17 (0.984)**
5 - USDA2	0.42 (0.001)	0.31 (0.001)

6 - USDA3	0.15 (0.006)	0.37 (0.001)
6 - FC	0.51 (0.001)	0.43 (0.001)

B. Soil Classification: Comparison of Five Machine Learning Algorithms

Using only the soil type as the classifier, the biotic information provided 98%, 95%, 99%, 91% and 91% accuracy (AUC= 0.93-1) with K-Nearest Neighbors, Decision Tree, Random Forest, Neural Networks, Support Vector Machines, respectively (Figure 5). At the transect level, accuracies were 92%, 85%, 98%, 64% and 89% (AUC= 0.95-1) and at the subplot level, classification accuracies dropped to 51%, 6%, 67%, 13%, 45% (AUC= 0.97-0.99) with K-Nearest Neighbors, Decision Tree, Random Forest, Neural Network, Support Vector Machines, respectively (Figure 5). Irrespective of which machine learning was used, soil type classification resulted in significantly higher accuracy when compared to transect, and subplot ($p< 0.007$). With all three classifiers, Random Forest had the highest classification accuracy compared to the other algorithms. Student t-test results show that Random Forest significantly outperformed all other algorithms regardless of the level selected (e.g., soil type ($p< 0.044$), transect ($p< 0.001$), subplot ($p< 0.001$) except K-Nearest Neighbors ($p= 0.065$) for Soil type.

Table 2. Prediction accuracy and AUC values (±SD of the mean) for each of the five machine learning tools (KNN, DT, RF, NN, SVM) based on three repeats.

		Soil Type	**Transect**	**Subplot**
KNN	**Accuracy**	98.5 ± 0.44	92.57 ± 0.42	51.67 ± 2.65
	AUC	0.99 ± 0.01	0.98 ± 0.01	0.98 ± 0.01
DT	**Accuracy**	95.26 ± 1.54	85.45 ± 1.37	6.16 ± 2.06
	AUC	0.97 ± 0.01	0.96 ± 0.00	0.94 ± 0.01
RF	**Accuracy**	99.76 ± 0.24	98.1 ± 0.36	67.98 ± 1.52
	AUC	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.00
NN	**Accuracy**	91.61 ± 1	64.71 ± 0.67	13.91 ± 0.94
	AUC	0.93 ± 0.01	0.95 ± 0.01	0.97 ± 0.00
SVM	**Accuracy**	91.86 ± 0.84	89.47 ± 0.68	45.95 ± 1.86

	AUC	0.95 ± 0.01	0.97 ± 0.00	0.98 ± 0.02

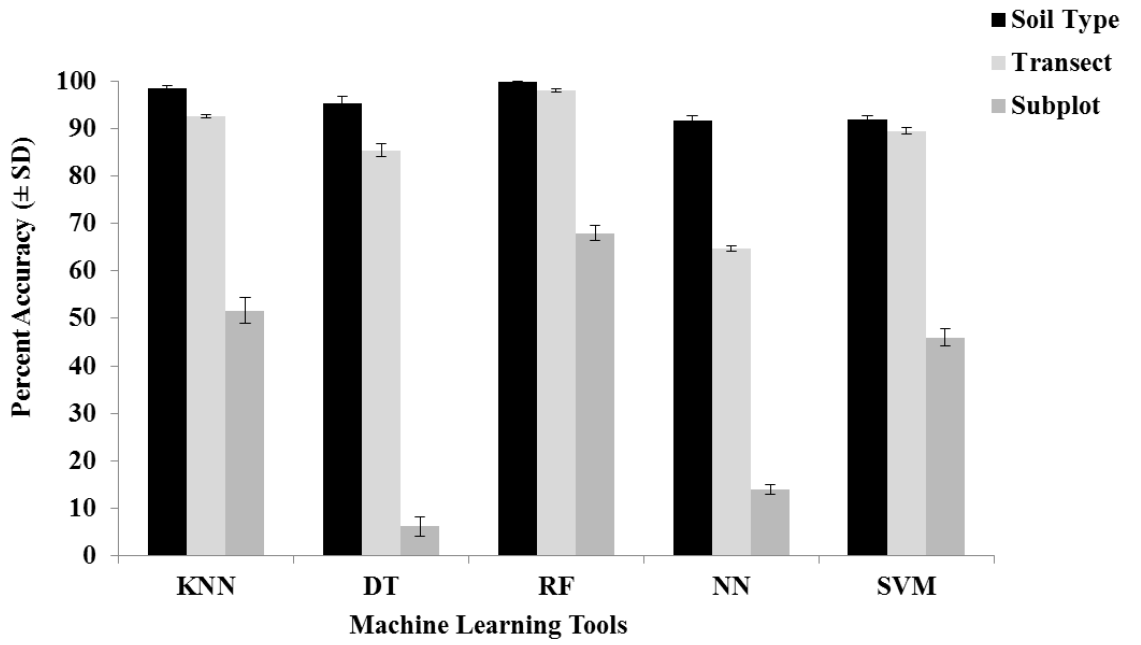

Figure 5. Prediction accuracy values and AUC values (±SD of the mean) for each of the five machine learning tools (KNN, DT, RF, NN, SVM) using training and test sets randomly chosen three different times from the complete database. Black bars = soil type, light grey bars = transect, dark grey bars = subplot.

Two machine learning tools, Random Forest and Support Vector Machines, were re-evaluated using different minimum relative ratios of the electrophoretic data (1%, 5%, 10%, 20%) to determine if reducing the number of variables would increase or decrease the classification accuracy. For example, 5% threshold indicated that relative peak ratios under 0.05 were marked as zero. Testing only Random Forest and Support Vector Machine algorithms, the classification accuracy did not significantly alter by increasing the electrophoretic threshold to 5% ($p= 0.528$). However, increasing the threshold to above 10% did significantly reduce the classification accuracy ($p< 0.001$). Using 1% and 5% threshold, Random Forest significantly outperformed Support Vector Machines ($p< 0.004$); however, under higher thresholds (10%, 20%) the two machine learning tools

were not significantly different (p= 0.570, 0.848, respectively). Prediction accuracy and AUC values are given in Table 3.

Table 3. Prediction accuracy and AUC values (±SD of the mean) for Random Forest and Support Vector Machines using different minimum relative ratios of electrophoretic data (1%, 5%, 10%, 20%).

		Soil Type		Transect	
	Threshold	**Accuracy**	**AUC**	**Accuracy**	**AUC**
RF	**1%**	99.76 ± 0.24	1.00 ± 0.00	98.10 ± 0.36	1.00 ± 0.00
	5%	99.19 ± 0.32	0.99 ± 0.01	94.27 ± 0.89	0.99 ± 0.00
	10%	93.23 ± 0.12	0.97 ± 0.00	73.97 ± 0.78	0.95 ± 0.00
	20%	63.30 ± 1.59	0.86 ± 0.01	35.72 ± 0.84	0.86 ± 0.02
SVM	**1%**	91.86 ± 0.84	0.95 ± 0.01	89.47 ± 0.68	0.97 ± 0.00
	5%	93.21 ± 0.55	0.96 ± 0.01	87.94 ± 0.23	0.96 ± 0.00
	10%	89.63 ± 0.88	0.95 ± 0.00	70.38 ± 0.94	0.92 ± 0.01
	20%	64.52 ± 0.21	0.87 ± 0.01	37.9 ± 0.65	0.85 ± 0.02

C. Similarity Percentages

A SIMPER analysis conducted at different scales—soil type, transect, subplot and season illustrated the dissimilarities between and within each scale. (See Tables 4 and 5) For example, at the level of soil types (labeled "Between"), Table 4 shows the average dissimilarity of one soil type (i.e., soil type 1) when compared to the other five soil types, while the "Within" comparisons consider the average dissimilarity of the 2-4 transects within a soil type. In contrast, as shown in Table 5, the "Between" column compares the average dissimilarity of one transect (i.e., FIU) when compared to the other seventeen transects, while the "Within" column compares the average dissimilarities of the six subplots within the transect. Overall, the average dissimilarity between site comparisons were greater than within sites. For example, between soil type dissimilarities ranged from 80-88% and between transects dissimilarities ranged from 74-92%, while their within site dissimilarities ranged from 50-80% and 28-65% for soil type and transect, respectively.

Seasonal dissimilarity was different based on soil type and transect. For example, for transect level, KS8 had the lowest season dissimilarity of 28% while, PE had the largest seasonal dissimilarity of 76%. This can be attributed to the soil physical characteristics.

Table 4. SIMPER analysis illustrating the average dissimilarity between and within each soil type. (± is the SD of the mean % dissimilarity).

	Dissimilarity (%)		
Soil Type	**Between**	**Within**	**Season**
1	80.27 ± 8.17	67.52 ± 6.42	64.57
2	81.82 ± 6.56	54.40 ± 2.59	51.08
3	81.78 ± 7.19	50.82 ± 0.00	54.90
4	88.39 ± 5.21	80.84 ± 0.00	79.07
5	82.67 ± 3.87	66.01 ± 6.70	61.71
6	81.12 ± 1.82	65.73 ± 0.00	59.89

Table 5. SIMPER analysis illustrated the average dissimilarity between and within each transect (± is the SD of the mean % dissimilarity).

	Dissimilarity (%)		
Transect	**Between**	**Within**	**Season**
FIU	78.07 ± 9.04	62.33 ± 4.44	63.24
OSP1	77.79 ± 9.46	52.92 ± 8.77	51.09
OSP2	74.14 ± 9.06	42.41 ± 3.40	46.54
OSP3	80.01 ± 7.92	52.95 ± 10.84	54.90
NW	75.03 ± 12.07	42.72 ± 7.12	44.78
KNT	77.27 ± 12.14	53.93 ± 9.28	59.45
KS8	76.62 ± 12.64	28.54 ± 3.34	28.57
CC6	74.64 ± 11.58	31.73 ± 3.54	35.01
KK	79.16 ± 9.95	46.86 ± 3.54	57.81
CH	78.95 ± 10.21	47.74 ± 8.74	52.45
CS	85.44 ± 4.64	64.10 ± 6.97	70.00
PE	92.18 ± 6.34	60.27 ± 2.25	76.06
HA	75.95 ± 8.63	31.16 ± 4.12	31.70
TREC	80.05 ± 8.71	52.50 ± 4.21	54.78
USDA1	81.88 ± 7.53	53.94 ± 4.60	63.94
USDA2	78.74 ± 7.29	45.16 ± 4.36	46.43
USDA3	79.92 ± 5.95	46.08 ± 2.75	64.16
FC	80.64 ± 5.54	41.80 ± 4.90	42.65

D. Discriminatory LH-PCR Peaks

Random Forest analysis provided a "Mean Decrease Accuracy" for the different LH-PCR peaks to determine the most important variable to discriminate between the soils being classified. The higher the accuracy decrease due to the exclusion of a single variable, the more important that variable is deemed. Three scales were analyzed-soil type, transect, and subplot. The Mean Decrease Accuracy is calculated based on an out of bag error calculation phase to determine if the accuracy of the random forest prediction decreases when the single variable is excluded. Overall, the results illustrated that with finer resolution scale (i.e., subplot vs soil type) more peaks were important to accurately classify the soil's origin (Figure 6). Moreover, these data support the threshold data (Table 3) and illustrated that all four taxa were important to discriminate between soils. Lastly, the Random Forest analyses were supported by SIMPER analysis results of their unique LH-PCR peaks that contributed to the dissimilarity between sites.

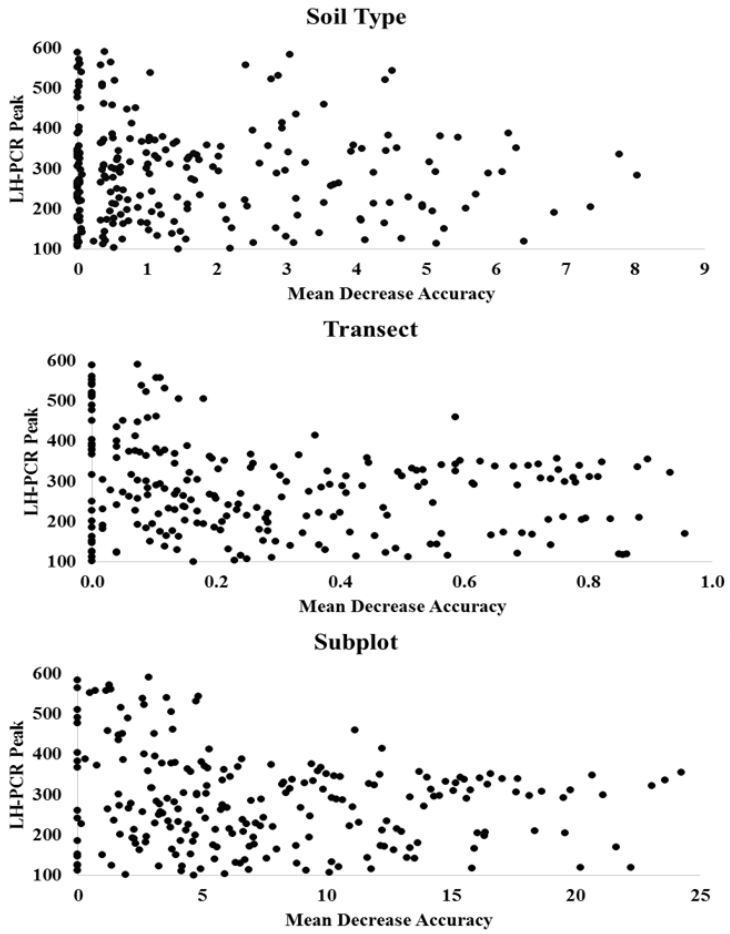

Figure 6. Most Important Variables for discriminating between soils at multiple spatial scales (soil type, transect, and subplot) based on Random Forest analysis. The greater the Mean Decrease Accuracy, the more important the LH-PCR peak for classification.

V. Discussion

Several studies have recently shown that bacterial community profiling of metagenomic samples using Next Generation Sequencing (NGS) technologies is strongly correlated with soil location and soil disturbance [28,29]. In the present study, Length Heterogeneity PCR (LH-PCR) data instead of NGS metagenomic sequencing were used for several reasons: First, the method has been proven to be fast, robust and reproducible in studying microbial community dynamics [30,31]. Second, many forensic laboratories have not implemented NGS technologies in their laboratories and LH-PCR is one that

can be used with the standard forensic DNA instrumentation. Third, generating LH-PCR data is less expensive than generating NGS data. Finally, the purpose was to compare algorithms for their ability to discriminate patterns within a large data set, regardless of the method in which the data were generated. Previous study by MacDonald et al. utilized a multiplex T-RFLP approach that analyzed bacteria, archaea, and fungi, which led to better discrimination of soil samples as different taxa responded differently to spatio-temporal ecological drivers [6]. They showed prokaryotic composition provided greater resolution between two sites, but were more susceptible to air-drying and sensitive to dehydration pressures when processing the soil samples that led to population shifts. Many bacteria do, however, have survival mechanisms that allow for rapid adaptation such as changing allocations of osmolytes or having thicker cell walls or sporulation capabilities as often seen in Gram-positive bacteria which help survive with spurts of dry-rewetting environmental conditions [32]. Fungi provided discrimination between sites and archaea were most useful in identifying saline or water logged soil environments [6]. Plants also have a potential to be used to discriminate soils, as they are dependent on the soil's microbes, water, and nutrients. In this study, a four-taxa approach was employed to include plants as well as bacteria, archaea, and fungi, for better discrimination of soils for provenance applications. Moreover, five supervised machine-learning tools were evaluated to determine the best tool for classification and determine at what spatial scale they can predict origin of the soil. To our knowledge, the work presented here is one of the first studies to use bioinformatic tools for soil forensic application using four-taxa and is unique in its consideration of multiple spatial scales.

The present study builds on our growing knowledge of spatial relationships in microbial communities by applying the Mantel statistic to this dataset to illustrate that the biotic patterns and their geographic location are indeed spatially auto-correlated in Miami-Dade soils (Table 1). The non-significant spatial autocorrelation proved to be an indicator of disturbed or constructed sites when compared to the undisturbed transects within the same soil type—value added for discrimination of sites for provenance or forensic applications. Sites that showed non-significant spatial autocorrelation were found to have been disturbed by humans (i.e., CH had been burned six months prior to soil collection, PE was an old abandoned nursery, NW137 was an illegal mixed trash dump site, and the OSP transects spanned mixed forest vegetation to abandoned construction sites) [8]. Previous study conducted by Meyers & Foran (2008) showed that extensive human activity appeared to homogenize bacterial content [8]. On the basis of the four-taxa microbial profiles and Mantel test, correlation between biotic content and geographic location was observed, thus justifying the use of machine learning tools to predict biotic patterns that can be applied for determination of soil provenance.

Five supervised machine-learning algorithms were evaluated for their predictive value when using four-taxa biotic profiles for soil classification. Three spatial scales (soil type, transect (>1.6 km), subplot (within 100 m)) were evaluated to determine the scale at which the algorithms are able to accurately classify the microbial profiles. The main performance criterion that was evaluated was the classification accuracy, which is the measurement of the correctly classified instances (accuracy = total number of samples correctly classified/total number of samples) as well as a measure of the overall error rate. The second performance criterion evaluated was the area under an ROC curve

(AUC). Area under an ROC curve is widely used to measure the performance of supervised classification methods based on their ranking quality of sensitivity (true positive rate) as a function of the specificity (false positive rate). For forensics, a balance of sensitivity is required in which the method should be sensitive to detect differences but avoid false positive results [29]. Each classification tool has its own learning and prediction procedure; therefore, to be able to compare between the five different supervised machine learning tools, all classifiers were generated using the same training and test sets.

Studies have shown that no single classification method is superior in every case [9,12,13]. In the present study, we compared the five machine learning tools to determine which is the best tool to predict provenance of soil using four-taxa biotic LH-PCR profiles. For forensic provenance applications, the model needs to have a high degree of classification accuracy and be easily interpretable for implementation in a court of law. In the current study, all algorithms were able to classify the soil samples with high accuracy and high AUC values. Irrespective of the spatial scale (soil type, transect, or subplot), Random Forest had the highest classification accuracy and AUC value compared to the other algorithms (Figure 5). Moreover, Random Forest was able to accurately predict the origin of the soil using the four-taxa profiles at subplot level (samples within 100 m apart) with 67% accuracy (Figure 5). Those samples that misclassified were still classified within the transect of origin. The SIMPER analysis illustrated that the within transect variability was less than the variability between transects (Table 5). Multiple studies support these results and have also found that bacteria profiles within a habitat are more similar to each other than those from other ecosystems [33,34]. For example, in this

study, site KK had a 46% within site dissimilarity compared to 79% dissimilarity between transect (Table 5). This can be attributed to the similar and almost identical microbial flora and fauna within some transects. Previous studies have illustrated the differences of within site variability between homogeneous grassland over shrubland. Mummey & Stahl (2003) showed that homogeneous grasslands had a highly similar bacterial community and lower within site variability [35]. Local heterogeneity can be to the result of different soil properties and multiple environmental factors such as unique plant species, sunlight amount, and differing moisture content [34,36]. Seasonal dissimilarities also varied between transects (Table 5); however, it did not alter the classification of the microbial profiles at the different spatial scales. Lenz & Foran (2010) also found that there was a large level of variability within habitats spatially and temporally, but the variability did not have a substantial negative influence on the ability to group soils from a habitat from samples collected throughout a year [34].

When choosing the most appropriate algorithm, it is important to take into account the dataset. Neural Networks and Support Vector Machines are more complex algorithms as compared to Decision Trees, Random Forests, and K-Nearest Neighbors. For example, Decision Trees and Random Forest are simpler classifiers that perform better with discrete and categorical data as they approach the variables with the purpose of finding the most discriminative variable that classifies and repeats this process until all of the data are classified [13]. Support Vector Machines and Neural Networks essentially find the maximal margin that can distinguish different classes that result in a highly comprehensible model but at times can also have the potential to over-fit the data [4,13]. Therefore, Support Vector Machines and Neural Networks are capable of working with

high-dimensional and continuous data, but require variable selection and do not perform well with large number of irrelevant variables [13].

As variable selection can significantly influence the performance of machine learning tools, Random Forest and Support Vector Machines were re-evaluated using different minimum thresholds based on the relative ratios electrophoretic data (1%, 5%, 10%, and 20%). Increasing the minimum threshold resulted in a continuous reduction of peaks, essentially removing the less intense peaks. It was expected that Support Vector Machines classification accuracy would increase with more stringent electrophoretic thresholds as we expected "irrelevant" variables (i.e., low intensities) would be reduced with higher RFU thresholds. The results indicated that for both Random Forest and Support Vector Machines, the classification accuracy did not significantly alter by increasing the electrophoretic threshold to 5% ($p= 0.528$). However, increasing the threshold above 10% did significantly reduce the classification accuracy ($p< 0.001$) of both algorithms. With the higher thresholds (10%, 20%) the two machine learning tools were not significantly different ($p= 0.570$, 0.848, respectively) in their prediction accuracy; however, Random Forest outperformed Support Vector Machines at thresholds of 1% and 5% ($p< 0.004$). A previous study (Meyers and Foran, 2008) found that using the "top (highest peak intensity) 40 peaks" of a bacterial profile generated with universal primers was as effective in discriminating soil samples versus using all of the electrophoretic peaks and similarity indices. They determined that observing the top 40 peaks, reduced the inclusion of small non-reproducible peaks that can occur by slight differences in amount of DNA injected into the capillary [8]. In our study, using four-taxa profiles showed that as the minimum relative ratio threshold was raised, the majority of

the peaks that were removed were archaea, fungi, and plant peaks (100-200 and >350bp) and resulted in a decrease in classification accuracy. Increasing the threshold resulted in lowering the number of data points being analyzed and resulted in losing peaks that represented distinguishing taxa. Therefore, these 'rare' peaks representing various members of the community are important to the specific habitat and provided "uniqueness" to the sample, which is important in forensics and provenance studies. Our results show that as you increase the spatial resolution from soil type to subplot level all the LH-PCR amplicons are important to discriminate locations (Figure 6). A peak threshold between 1-5% was needed that included all taxa in order to provide the identity of an unknown sample to its approximate origin. Therefore, this demonstrated the significance of using four taxa to provide higher accuracy and discrimination between sites.

The current study demonstrated that using four-taxa biotic profiles combined with user-friendly classification algorithms can provide a significant tool to the forensic and intelligence community. The biotic analyses can be conducted with the DNA expertise and instrumentation already employed in many crime laboratories making it easy to implement and can be used with ≤ 500 mg of soil [31,37]. The implementation of these methods could provide a routine use of soil microbial community profiles for soil provenance and assist in intelligence gathering or forensic investigations. This study recommends the use of Random Forest Supervised Machine Learning Tool with a threshold below 5% as a data analysis pipeline for best classification of soil provenance. Previous study by Beck and Foster (2014) also concluded that Random Forest was computationally efficient and easy to extract important model features [11]**.** Kampichler

et al. (2010) also recommended the utilization of Random Forest for biologists and decision makers due to their ease of interpretability of classifiers and clarity of the method [9]. Further studies should be conducted to determine the sampling design- number of samples collected and distance between samples across different habitats- to utilize soil microbial profiling for intelligence based forensic investigations and ultimately establish a usable database for soil provenance.

VI. Conclusion

In conclusion, this study showed that there was a correlation based on the four-taxa biotic community profiles and the geographic locations from which they were collected. The sites that displayed non-significance correlation between its geographic location and microbial profile information still provided important information illustrating site disturbance (i.e., recent fire, constructed site). These 'red flags' could allow further corroboration for soil evidence to a particular site. Moreover, this study has demonstrated the power of bioinformatic tools such as machine learning algorithms for identifying patterns in data. While each of the tools utilized in this study performed with high accuracy and would prove useful, Random Forest analysis demonstrated consistently high accuracy at all spatial scales and would be recommended for use in provenance and soil forensics.

VII. References

[1] L.I. Moreno, D.K. Mills, J. Entry, R.T. Sautter, K. Mathee, Microbial metagenome profiling using amplicon length heterogeneity-polymerase chain reaction proves more effective than elemental analysis in discriminating soil specimens, J. Forensic Sci. 51 (2006) 1315-1322.

[2] S.A. Larson, Developing a high throughput protocol for using soil molecular biology as trace evidence, Theses and Dissertation in Biochemistry. (2012).

[3] J. Horswell, S.J. Cordiner, E.W. Maas, T.M. Martin, K.B.W. Sutherland, T.W. Speir, et al., Forensic comparison of soils by bacterial community DNA profiling, J. Forensic Sci. 47 (2002) 350-353.

[4] C. Yang, D. Mills, K. Mathee, Y. Wang, K. Jayachandran, M. Sikaroodi, P. Gillevet, J. Entry, G. Narasimhan, An ecoinformatics tool for microbial community studies: Supervised classification of Amplicon Length Heterogeneity (ALH) profiles of 16S rRNA, J. Microbiol Methods. 65 (2006) 49-62.

[5] J.A. Entry, D. Mills, K. Mathee, K. Jayachandran, R. Sojka, G. Narasimhan, Influence of irrigated agriculture on soil microbial diversity, Applied Soil Ecol. 40 (2008) 146-154.

[6] L.M. Macdonald, B.K. Singh, N. Thomas, M.J. Brewer, C.D. Campbell, L.A. Dawson, Microbial DNA profiling by multiplex terminal restriction fragment length polymorphism for forensic comparison of soil and the influence of sample condition, J. Appl Microbiol. 105 (2008) 813-821.

[7] C.A. Macdonald, R. Ang, S.J. Cordiner, J. Horswell, Discrimination of soils at regional and local levels using bacterial and fungal T-RFLP profiling, J. Forensic Sci. 56 (2011) 61-69.

[8] M.S. Meyers, D.R. Foran, Spatial and temporal influences on bacterial profiling of forensic soil samples, J. Forensic Sci. 53 (2008) 652-660.

[9] C. Kampichler, R. Wieland, S. Calmé, H. Weissenberger, S. Arriaga-Weiss, Classification in conservation biology: A comparison of five machine-learning methods, Ecol Inform. 5 (2010) 441-450.

[10] A.L. Tarca, V.J. Carey, X. Chen, R. Romero, S. Drăghici, Machine learning and its applications to biology, PLoS Comput Biology. 3 (2007) e116.

[11] D. Beck, J.A. Foster, Machine learning techniques accurately classify microbial communities by bacterial vaginosis characteristics, PLoS One. 9 (2014) e87830.

[12] R. Caruana, A. Niculescu-Mizil, An empirical comparison of supervised learning algorithms, Proc 23rd Intl Conf Mach Learning. ACM (2006) 161-168.

[13] A.C. Tan, D. Gilbert, An empirical comparison of supervised machine learning techniques in bioinformatics, Proc First Asia-Pacific Bioinfo Conf Bioinfo. (2003) 219-222.

[14] C.V. Noble, R.W. Drew, J.D. Slabaugh, Soil survey of Dade County area, Florida, USDA NRCS, Gainesville, FL. (1996).

[15] M. Duever, J. Meeder, L. Meeder, J. McCollom, The climate of South Florida and its role in shaping the Everglades ecosystem, Everglades: The ecosystem and its restoration, St. Lucie Press, Boca Raton, FL, 1994, pp. 225-248.

[16] M. Suzuki, M.S. Rappe, S.J. Giovannoni, Kinetic bias in estimates of coastal picoplankton community structure obtained by measurements of small-subunit rRNA gene PCR amplicon length heterogeneity, Appl Environ Microbiol. 64 (1998) 4522-4529.

[17] E.F. DeLong, Archaea in coastal marine environments, Proc Natl Acad Sci. 89 (1992) 5685-5689.

[18] L. Cocolin, M. Manzano, C. Cantoni, G. Comi, Denaturing gradient gel electrophoresis analysis of the 16S rRNA gene V1 region to monitor dynamic changes in the bacterial population during fermentation of Italian sausages, Appl Environ Microbiol. 67 (2001) 5113-5121.

[19] T.J. White, T. Bruns, S. Lee, J. Taylor, Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics, PCR protocols: a guide to methods and applications. 18 (1990) 315-322.

[20] P. Taberlet, L. Gielly, G. Pautou, J. Bouvet, Universal primers for amplification of three non-coding regions of chloroplast DNA, Plant Mol Biol. 17 (1991) 1105-1109.

[21] S. Dray, A.B. Dufour, The ade4 package: implementing the duality diagram for ecologists. J. Stat Soft. 22 (2007) 1-20.

[22] B.D. Ripley, Pattern recognition and neural networks, Cambridge University Press 2007.

[23] L. Breiman, J. Friedman, C.J. Stone, R.A. Olshen, Classification and regression trees, CRC press 1984.

[24] A. Liaw, M. Wiener, Classification and Regression by randomForest, R news. 2 (2002) 18-22.

[25] F. Günther, S. Fritsch, neuralnet: Training of neural networks, The R Journal. 2 (2010) 30-38.

[26] E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, A. Weingessel, Misc functions of the Department of Statistics (e1071), TU Wien, R package. (2008) 1.5-24.

[27] D.J. Hand, R.J. Till, A simple generalisation of the area under the ROC curve for multiple class classification problems, Mach Learning. 45 (2001) 171-186.

[28] D. Buckley, T. Schmidt, The structure of microbial communities in soil and the lasting impact of cultivation, Microb Ecol. 42 (2001) 11-21.

[29] J.M. Young, L.S. Weyrich, J. Breen, L.M. Macdonald, A. Cooper, Predicting the origin of soil evidence: High throughput eukaryote sequencing and MIR spectroscopy applied to a crime scene scenario, Forensic Sci Int. 251 (2015) 22-31.

[30] L.I. Moreno, D. Mills, J. Fetscher, K. John-Williams, L. Meadows-Jantz, B. McCord, The application of amplicon length heterogeneity PCR (LH-PCR) for monitoring the dynamics of soil microbial communities associated with cadaver decomposition, J. Microbiol Methods. 84 (2011) 388-393.

[31] M. Doud, E. Zeng, L. Schneper, G. Narasimhan, K. Mathee, Approaches to analyze dynamic microbial communities such as those seen in cystic fibrosis lung, Hum Genomics. 3 (2009) 246-256.

[32] J. Schimel, T.C. Balser, M. Wallenstein, Microbial stress-response physiology and its implications for ecosystem function, Ecology. 88 (2007) 1386-1394.

[33] L.E. Heath, V.A. Saunders, Assessing the potential of bacterial DNA profiling for forensic soil comparisons, J. Forensic Sci. 51 (2006) 1062-1068.

[34] E.J. Lenz, D.R. Foran, Bacterial profiling of soil using genus-specific markers and multidimensional scaling, J. Forensic Sci. 55 (2010) 1437-1442.

[35] D.L. Mummey, P.D. Stahl, Spatial and temporal variability of bacterial 16S rDNA-based T-RFLP patterns derived from soil of two Wyoming grassland ecosystems, FEMS Microbiol Ecol. 46 (2003) 113-120.

[36] R.B. Franklin, A.L. Mills, Multi-scale variation in spatial heterogeneity for microbial community structure in an eastern Virginia agricultural field, FEMS Microbiol Ecol. 44 (2003) 335-346.

[37] G.F. Sensabaugh, Microbial community profiling for the characterisation of soil evidence: forensic considerations, Criminal and Environmental Soil Forensics, Springer, 2009, pp. 49-60.

VIII. Supplemental Information

A. S1 File. Script example for the Mantel tests performed in this study.

Below demonstrates how to perform the Mantel test for soil type level (ex. soil type 1). This was performed using R software with the ade4 library. Each soil type/ transect was made up of subsets out of the dataset by using rows that displayed the samples of interest. Mantel test was performed based on the genetic distance (E.dists)

and gps distance (gps.dists) based on 999 permutations. The Mantel tests were plotted using the function mantel.randtest.

```
Require (ade4)
# Input data
> All<-read.csv(file.choose(), header=TRUE)
# Make subsets to get row of interest
> Soil1<-subset(All[c(1:303),])
# GPS matrix
Soil1.gps.dists<-dist(cbind(Soil1$West,Soil1$North))
# Genetic matrix
> Soil1.E.dists<-dist(cbind(Soil1[,4:235]))
# Mantel test based on 999 permutations
> Mantel.Soil1<-mantel.randtest(Soil1.gps.dists,Soil1.E.dists,nrepet=999)
# Plot Mantel test
> plot(Mantel.Soil1<-mantel.randtest(Soil1.gps.dists,Soil1.E.dists),main="Soil 1
Mantel Test")
```

B. S2 File. Script examples for Machine Learning tests performed in this study.

a=K-Nearest Neighbor b=Decision Tree, c=Random Forest, d=Neural Networks, e=Support Vector Machine). Below demonstrates how to perform the different Machine Learning tests for soil type level (ex. soil type 1).

a) K-Nearest Neighbor:

```
Require (class)
# Input data: Testing and Training data separately
> Train<-read.csv(file.choose(), header=TRUE)
> Test<-read.csv(file.choose(), header=TRUE)
# Column Classification
> cl<-traindata[,1]
# Predict classification
> pred<-knn(traindata[,2:11],testdata[,2:11, cl,k=3])
# Print prediction results
> print(pred)
```

b) Decision Tree:

```
Require (rpart)
# Input data: Testing and Training data separately
> Train<-read.csv(file.choose(), header=TRUE)
> Test<-read.csv(file.choose(), header=TRUE)
# Model the training data
> model.train_Type<-rpart(Train_Type$Soil.Type~.,method="class",data=Train)
# Prune the training model
> model.prunetrain_Type<-prune(model.train_Type,newdata=model.train_Type,
cp=model.train_Type$Soil.Type[which.min(model.train_Type$Soil.Type[,"xerror"]),
"CP"])
```

```
# Print the training model
> model.prunetrain_Type
# Plot the training model
> plot(model.prunetrain_Type, uniform=TRUE, main="Pruned Tree for Biotic
Training Data for
Soil Type")
> text(model.prunetrain_Type, use.n=TRUE, cex=.5, pretty=0)
# Predict classification of testing data based on the training model
> pred.prunetest_Type<- predict(model.prunetrain_Type,newdata=Test,type="class")
# Print prediction results
> pred.prunetest_Type
```

c) Random Forest:

```
Require (randomForest)
# Input data: Testing and Training data separately
> Train<-read.csv(file.choose(), header=TRUE)
> Test<-read.csv(file.choose(), header=TRUE)
# Model the training data
> model.rf_Type <-
randomForest(Train_Type$Soil.Type~.,data=Train,importance=TRUE,
mtry=3)
# Print the training model
> model.rf_Type
# Plot the important variables for classification
> varImpPlot(model.rf_Type,type=1,sort=FALSE,n.var=10,main="Variable
Importance for
Biotic Training Data for Soil Type", cex=.5)
# Round and print the important variables
> round(importance(model.rf_Type),2)
# Predict classification of testing data based on the training model
> pred.rf_Type <- predict(model.rf_Type,newdata=Test,type="class")
# Print prediction results
> pred.rf_Type
```

d) Neural Networks:

```
Require (neuralnet)
# Input data: Testing and Training data separately
> Train<-read.csv(file.choose(), header=TRUE)
> Test<-read.csv(file.choose(), header=TRUE)
# Model the training data
> net.train_Type<-neuralnet(Train_Type$Soil.Type~.,data=Train,hidden=155,
threshold=0.01,
rep=5)
# Print the training model
> print(net.train_Type)
```

```
# Predict classification of testing data based on the training model
> net.results_Type <- compute(net.train_Type, Test_Type)
# Print prediction results
> Type<-print(net.results_Type$net.result)
```

e) Support Vector Machines:

```
Require (e1071)
# Input data: Testing and Training data separately
> Train<-read.csv(file.choose(), header=TRUE)
> Test<-read.csv(file.choose(), header=TRUE)
# Model the training data
> model<-svm(Soil~Type.,data=traindata,type="C-classification")
# Print the training model
> model
# Predict classification of testing data based on the training model
> pred<-predict(model,testdata)
# Print prediction results
> table(pred)
```

Chapter 2: Geographic Information Systems approach to characterize the spatial variability of the soil microbial community and the application to forensics

This chapter is currently under review in Applied Geography.

Natalie Damaso[1,2], Jennifer Gebelein[3], and DeEtta Mills[1,2*]

[1]Department of Biological Sciences, Florida International University, Miami, Florida, United States of America

[2]International Forensic Research Institute, Florida International University, Miami, Florida, United States of America

[3]Department of Earth and Environment, Florida International University, Miami, Florida, United States of America

I. Abstract

Soil DNA profiling has great potential as a forensic tool and research to date have been promising. As human profiles are used to determine a match between evidence from a crime scene and a suspect, a soil microbial profile can be used to determine a match between soil found on the suspect's shoes or clothing to the soil at a crime scene. Soil properties are known to vary spatially and can be related to several physical, chemical, and biological processes that act at different scales and are important in shaping the composition of the microbial community. Therefore, spatial scale is an important consideration for forensic application. Understanding the spatial variability of the microbial community and the extent to which other soil variables might shape the community structure are important factors needed to develop sampling strategies. This variability is important for understanding the spatial range to determine the sampling scheme required to represent an ecosystem. In this study, a survey was conducted to examine the organization of the microbial community structure at multiple spatial scales across Miami-Dade, Florida. Multivariate statistics and geostatistics were used to observe the patterns as a function of distance. The results illustrated that semivariograms can provide a useful tool for designing robust sampling strategies by estimating the variance (sill) between sampling points as well as estimate the minimum distance required for samples to be considered spatially independent (range). Therefore, GIS provides a useful tool that can be used to inform sample strategies to build a robust database for soil provenance.

II. Introduction

"Distance-decay" is a universal biogeographic pattern that is commonly observed

with a wide variety of organisms and illustrates a decrease in community similarity with increasing geographic distance (Martiny, Eisen, Penn, Allison, & Horner-Devine, 2011). Studies have shown that soil organisms are not randomly distributed and exhibit spatially predictable, aggregated patterns with scale-dependent controls (Ettema & Wardle, 2002). Soil variability can be influenced by a combination of different factors such as spatial location, resolution or map scale, and specific soil properties (Lin, Wheeler, Bell, & Wilding, 2005a). Studies have attempted to verify the Beijerinck hypothesis "everything is everywhere but the environment selects" (a.k.a., soil type determines the microbial communities) to specify which environmental factors exert the strongest influence on the microbial communities (Fierer & Jackson, 2006). Understanding the cause of "distance-decay" patterns is an area of great interest. Based on the Beijerinck hypothesis, the distance decay patterning can only be driven by differences in environmental conditions across space.

Spatial ecology and modeling studies have concentrated on aboveground biota and abiotic properties (i.e., biogeochemical data and physical properties) and very little on microbial communities (Ettema & Wardle, 2002). Spatial data have the potential to improve our understanding of the ecological factors that regulate the soil biota and their functional roles (Tsiknia, Paranychianakis, Varouchakis, Moraetis, & Nikolaidis, 2014). Geostatistics has been used in soil science studies to assess and quantify the heterogeneity (patchiness) and variability of the spatial structure (Goovaerts, 1998). Pattern analysis can subsequently help develop, test, and validate prominent ecological theories driving biogeographic differences (Violle, Reich, Pacala, Enquist, & Kattge, 2014). Moreover, these hypotheses lay the foundation for using soil microbial

communities for forensic applications. Geostatistics uses the soil sample's spatial information to model spatial patterns, interpolate to unsampled locations, and assess uncertainty of the predictions (Goovaerts, 1998). This can be a valuable tool in forensics or for intelligence purposes to determine soil origin. Furthermore, Geostatistical analysis is a promising approach to model spatial patterns at various scales to understand whether a soil profiling approach can be applied in a one test fits all model or needs to be tuned depending on a particular soil type or geographic situation (Sensabaugh, 2009).

Soil DNA profiling has great potential as a forensic tool and research to date have been promising. The current ecological hypothesis states that the soil type (e.g., chemical and physical properties) determines which microbes occupy a particular soil and provides the foundation for soil provenance studies. Studies to date have shown the potential and effectiveness of using microbial DNA from soil, not just for comparison, but also for intelligence gathering to geographically pinpoint the origin of the soil (Heath & Saunders, 2006; Horswell et al., 2002; Moreno, Mills, Entry, Sautter, & Mathee, 2006). Previous studies have shown that supervised classification algorithms were able to classify and distinguish soils at multiple spatial scales--soil type, transect, and subplot levels--with high accuracy (Damaso et al., in review). This indicated that there are hidden patterns within the microbial profiles that can be discerned by the mathematical-based tools (Yang et al., 2006)

However, some layers of data, thus information, are often lost during classification schemes as a generalization is performed in order to organize the data into clear structural vectors. Moreover, these classification methods imply that the soil properties are discontinuous which is not correct as soil processes operate under different

scales (McKinley, 2013). Therefore, the spatial scale is important consideration for forensic applications (Sensabaugh, 2009). Understanding the spatial variability of the microbial community as a whole and the extent to which other soil variables might shape the structure is important to develop a sampling strategy and to understand the sample size, and the spatial and temporal scale of the collection required in order to representatively sample an ecosystem (Mummey & Stahl, 2003; Sensabaugh, 2009).

III. Processing Overview

In this study, multivariate and geostatistical techniques were used to validate the ecological theories structuring the soil microbial biogeography at multiple spatial scales. A survey was conducted to examine the spatial organization of the microbial community structure at multiple spatial scales across Miami-Dade, Florida. The community structure was compared using four-taxon microbial profiles. Spatial autocorrelation, a term for spatial dependence and queries the resemblance between "neighbors", was used as a function of spatial separation distance. When near neighbors are more similar than those farther away, the data are said to be autocorrelated, and therefore violate the assumption that the data are independent (Ettema & Wardle, 2002). The relative dissimilarity was calculated and compared using geostatistical variograms to observe the spatial patterns as a function of separation distance at different scales by fitting a continuous function to smooth sample fluctuations. The best variogram model would then be used to interpolate soil properties at un-sampled locations.

IV. Materials and Methods

A. Soil Collection

Soil samples (N= 1268) were collected across Miami-Dade County, Florida in

2010-2011 over two season (dry and wet) and one transect (FIU) over 1.5 year (dry-wet-dry) as described in Damaso et al. 2016 in review. The soils were classified into six different soil types according to USDA soil surveys (Noble, Drew, & Slabaugh, 1996): 1-Urban Land-Udorthents, 2- Lauderhill Dania-Pahokee, 3- Rock Outcrop-Biscayne-Chekika, 4- Perrine-Biscayne-Pennsuco, 5- Krome Association, 6- Perrine-Terra Ceia-Pennsuco. All six soil types with 2-4 transects per type were surveyed. Most transects were established in undisturbed sites that had limited public access. Each transect was ≥ 1.6 km distant from the next, transects were 100 m in length and six subplots were randomly sampled along each transect. GPS coordinates for every subplot for each transect were recorded. Within each subplot, six cored samples were taken within a $1m^2$ quadrat. A 5 cm diameter soil corer was used to collect the top 5-10 cm of the soil. The soil samples were transported back to the laboratory and sieved to remove large objects and debris for subsequent processing.

B. Microbial DNA Profiles

In this study, microbial DNA profiles were obtained by first extracting the metagenomic DNA from the soil sample, then amplified using length heterogeneity polymerase chain reaction (LH-PCR). This technique is a rapid, robust, and reliable method that uses universal taxa primer sets that have broad specificity for organisms known to be ubiquitous in soil. In this study, four taxa (i.e., bacteria, Archaea, fungi, and plant) universal DNA markers were amplified and separated by capillary electrophoresis to obtain a DNA profile that provide a unique soil fingerprint.

i. DNA Extraction

Extraction was conducted using the BIO 101 Fast DNA Spin Kit for Soil® and

FastPrep®-24 System homogenizer (MP Bio, Solon, OH). Quantification was conducted using the Qubit® Assay kit on the Qubit 2.0 Fluorometer (Invitrogen, Carlsbad, CA). Samples were diluted to a 20 ng/µl working stock.

ii. Length Heterogeneity-PCR (LH-PCR)

DNA was amplified as described in Damaso et al. 2016, in review using two PCR duplexes: (1) bacteria and fungi, and (2) plant and Archaea as well as each taxa separately. PCR reaction mixtures were: 1X reaction buffer, 2.5 mM $MgCl_2$, 0.25 mM dNTPs (Promega, Madison, WI), 1% BSA (fraction V, Fisher Scientific, Pittsburgh, PA), 1% DMSO (Promega, Madison, WI), various concentrations of primers (bacteria=0.5 µM, Archaea=0.4 µM, fungi=0.4 µM, plant=0.6 µM), 40 ng microbial DNA, 0.5 U AmpliTaq Gold® DNA Polymerase (Applied Biosystems, Foster City, CA). Universal primers were used for the following genomic regions for each taxa: 16S rRNA for bacteria (27-F, 355-R) (Suzuki, Rappe, & Giovannoni, 1998) and Archaea (21-F, 518-R) (Cocolin, Manzano, Cantoni, & Comi, 2001; DeLong, 1992) ribosomal internal transcribed spacer region (ITS) for fungi (ITS5-F, ITS2-R) (White, Bruns, Lee, & Taylor, 1990) and chloroplast *trn*l intergenic region for plant (trnL-F, trnL-R) (Taberlet, Gielly, Pautou, & Bouvet, 1991). Forward primers were labeled with 6-FAM fluorescent dye. Each duplex was amplified with the same program using the ABI 9700™ thermocycler (Applied Biosystems, Foster City, CA) with the following parameters: initial 10 minute denaturing step at 95°C, 25 cycles of denaturation at 95°C annealing at 54°C and extension at 74°C each for 30 seconds with a final extension at 74°C for 10 minutes.

iii. Capillary Electrophoresis

Fragment analysis was conducted using the ABI Prism™ 310 (Applied

Biosystems, Foster City, CA) using Performance Optimized Polymer 4 (POP4) (Applied Biosystems, Foster City, CA). Samples from the two duplexes were co-loaded where 1 µl of each PCR product was added to a mixture of 11.5 µl Hi-Di™ Formamide (Applied Biosystems, Foster City, CA) and 0.65 µl internal size standard, GeneScan LIZ®600 (Applied Biosystems, Foster City, CA), denatured by heating for 2 min at 95°C and then snap-cooled on ice for 2 min. Raw data were analyzed using the GeneMapper™ v 4.0 (Applied Biosystems, Foster City, CA). Local Southern size calling was used for the analysis parameters with a minimum threshold of 50 relative fluorescent units (RFUs). The relative ratios were calculated by normalizing the heights of each peak in the genotype to the total peak intensities resulting in the ratio for each peak height as a decimal value from zero to one using the Galaxy ABI Data Formatting tool found in http://usegalaxy.org/ (Afgan et al., 2016).

C. Assessment of Spatial Variability

Multivariate and geostatistical techniques were used to validate the ecological theories that shape the soil microbial biogeography at multiple spatial scales. Spatial autocorrelation using Mantel Test was determined, which is a term for spatial dependence and queries the resemblance between "neighbors" as a function of spatial separation distance. The relative dissimilarity using Bray Curtis Similarity was calculated and compared using geostatistical semivariograms to observe the spatial patterns as a function of separation distance at different scales by fitting continuous function to smooth out sample fluctuations. For the first set of analyses, the relationship between all samples in Miami-Dade were considered to obtain an average overall spatial variability in the plot. Then subsets of the data were analyzed to quantify the autocorrelation and spatial

variability at different spatial scales by varying the maximum separation distance. The scales were based on relative size: 1) plot scale (all samples in Miami-Dade County), 2) soil type scale (i.e., same physical and chemical properties as defined by USDA and >1.6km), and 3) transect scale (100m). Because of the relatively small number of samples (i.e., 6 samples per subplot), correlation analysis did not analyze subplot scale ($1m^2$). The best variogram model was used to interpolate soil properties at un-sampled locations.

i. Spatial Autocorrelation (Mantel Test)

Mantel tests were performed using R programming language with the ade4 library (Dray & Dufour, 2007). Two distance matrices were tested: geographic distance and genetic distance with data imported as binary data (presence/absence). The Mantel tests were performed and plotted using the function mantel.randtest in the ade4 package and calculated based on the random permutation using the Monte Carlo method. This method relies on repeated random sampling (using 999 permutations) to compute the results so that no assumptions regarding the statistical distributions of samples in the matrix were needed. The rows and columns of one matrix were randomly permutated followed by recalculation of the correlation after each permutation, thereby testing the significance.

ii. Statistical Analysis (Dissimilarity Percentages)

All analyses were conducted using Primer-E v.7 software (PRIMER E Ltd., Plymouth Marine Laboratory, Plymouth, U.K.). Dissimilarity percentages were obtained from Bray-Curtis similarity matrices that were generated on relative abundance ratios that had been square-root transformed prior to analysis. SIMPER analysis was conducted to determine the percent dissimilarity within and between samples at multiple spatial scales

(i.e., soil type, transect, subplot) and seasonal differences (i.e., wet, dry). The dissimilarity percentages were used as the input data in the semivariogram tool to assess the variability in the microbial community composition at various spatial scales.

iii. Geographic Information Systems (Semivariograms)

Geostatistics tool within GIS was utilized to observe the spatial patterns of the soil microbial profile at different scales. The theoretical semivariogram model fitting is usually expressed by three parameters: nugget, sill, and range. The nugget represents the measurement errors or spatial dependence at scales not explicitly sampled. The sill represents the variance of the correlated measurements. The range shows the extent of heterogeneity (i.e., zone of influence or distance of dependence) (Ettema & Wardle, 2002; Goovaerts, 1998). Since microbial profiles generate multivariate relative abundance data, it is not possible to calculate the semi-variance between sample pairs. Instead pseudo-variograms were created using 'relative dissimilarity' values calculated from the Bray Curtis similarity index. These pseudo-variograms are constructed and analyzed the same way as the traditional variograms (Franklin & Mills, 2003). Semivariograms were constructed that represented the spatial variability in terms of dissimilarity between observations as a function of geographic distance (Goovaerts, 1998). The semivariograms show the hypothetically observed distance class (filled circles) and the fitted model (solid line). Three different variogram models: spherical, exponential, and Gaussian models were conducted and the best model was chosen based on the Root Mean Square Error (RMSE). An exponential semivariogram best describes a spatial structure where a variable displays abrupt changes at all distances. A Gaussian variogram fits the spatial patterns where the variable has a continuous, gradually varying

structure. Lastly, the spherical semivariogram describes spatial structure that has no clearly defined abrupt boundaries (Franklin, Blum, McComb, & Mills, 2002).

Prior to constructing the variograms, it was necessary to segregate the data into distance classes by calculating the appropriate number of bins and appropriate bin width (i.e., lag distance) because this allows for the maximum resolution to be obtained at small distances without being misled by structural artifacts. This technique allowed dominant spatial patterns at each scale to be quantified and obscured the autocorrelation structure at smaller distances. The lag distance was calculated by considering the maximum separation distance between sample pairs as discussed in Franklin et al. (2002 & 2003). The appropriate number of bins for each analysis was determined by Sturge's rule, which states that appropriate number of classes $= 1 + 3.3\log(m)$, where m is the number of points in either the upper or lower triangle matrix. Furthermore, variograms are not valid beyond half of the maximum distance between samples and so the appropriate lag distance (distance increment for each class) was calculated as maximum pair distance divided by 2 and then subdivided into number of equal classes as described in (Franklin et al., 2002).

V. Results

A. Spatial Autocorrelation Analysis: Mantel Test

Autocorrelations were conducted using Mantel tests which linked microbial DNA profiles to soil type as well as to specific transects within a soil type with strong accuracy (Table 6). Therefore, spatial autocorrelation observed in the soil samples illustrate that the microbial communities that are closer geographically are closer genetically. The four taxonomic profiles of each soil type displayed significant positive autocorrelation ranging

from 0.22-0.83 for both seasons (Table 6). At the transect level, six transects had non-significant correlations (Table 6). Some sites were found to have been previously disturbed (e.g., fire, illegal dumping, agricultural disturbance).

Table 6. The Mantel test results for all of Miami-Dade County's six soil types, transects within each soil type, for each season (wet and dry). Numbers represent the Mantel coefficient (positive correlation>0; negative correlation<0; random=0).

Soils	Wet	Dry
All Miami-Dade	0.32*	0.35*
By soil type		
Soil Type 1	0.24*	0.29*
Soil Type 2	0.53*	0.42*
Soil Type 3	0.22*	0.44*
Soil Type 4	0.43*	0.47*
Soil Type 5	0.38*	0.48*
Soil Type 6	0.83*	0.79*
By transect#		
1–FIU	0.43*	0.18*
1–OSP1	-0.04	0.09
1–OSP2	0.24*	-0.08
1–OSP3	0.19*	0.22*
2–CC6	0.29*	0.55*
2–KNT	0.43*	0.16*
2–KS8	0.41*	0.53*
2–NW137	0.25*	-0.01
3–CH	0.03	0.42*
3–KK	0.35*	0.46*
4–CS	0.17*	0.35*
4–PE	-0.17	0.22*
5–HA	0.46*	0.13*
5–TREC	0.21*	0.16*
5–USDA1	-0.05	-0.17
5–USDA2	0.42*	0.31*
6–USDA3	0.15*	0.37*
6–FC	0.51*	0.43*

#: Soil samples are identified by a soil type number, followed by a transect descriptor (e.g., 1-FIU corresponds to soil type 1, transect FIU).
*Represents significant ($p \leq 0.05$) spatial autocorrelation.

B. Multivariate Statistics: Dissimilarity Percentages

SIMPER analysis conducted at different scales—soil type, transect, subplot and season illustrated the dissimilarities between and within each scale (Table 7). For example, at the level of soil types (labeled "Between"), Table 7 shows the average dissimilarity of one soil type (i.e., soil type 1) when compared to the other five soil types, while the "Within" comparisons consider the average dissimilarity of the 2-4 transects within one soil type. In contrast, under transects, the "Between" column compares the average dissimilarity of one transect (i.e., FIU) when compared to the other seventeen transects, while the "Within" column compares the average dissimilarities of the six subplots within the transect. "Season" represents the average dissimilarity between wet and dry season for each site. Overall, the average dissimilarity between site comparisons was greater than within sites. For example, between soil type dissimilarities ranged from 80-88% and between transects dissimilarities ranged from 74-92%, while their within site dissimilarities ranged from 50-80% and 28-65% for soil type and transect, respectively. Seasonal dissimilarity varied based on soil type and transect with transect level, KS8 having the lowest season dissimilarity of 28% while, PE had the largest seasonal dissimilarity of 76%. This can be attributed to the soil physical characteristics.

Table 7. SIMPER analysis illustrating the average dissimilarity between and within each soil type and transect (± is the SE of the mean % dissimilarity).

Soils	Between	Within	Season
Soil Type			
1	80.27 ± 3.34	67.52 ± 2.62	64.57
2	81.82 ± 2.68	54.40 ± 1.06	51.08
3	81.78 ± 2.93	50.82 ± 0.00	54.9
4	88.39 ± 2.13	80.84 ± 0.00	79.07
5	82.67 ± 1.58	66.01 ± 2.74	61.71

6	81.12 ± 0.74	65.73 ± 0.00	59.89
Transect#			
1–FIU	78.07 ± 2.13	62.33 ± 1.15	63.24
1–OSP1	77.79 ± 2.23	52.92 ± 2.26	51.09
1–OSP2	74.14 ± 2.14	42.41 ± 0.88	46.54
1–OSP3	80.01 ± 1.87	52.95 ± 2.80	54.90
2–NW	75.03 ± 2.84	42.72 ± 1.84	44.78
2–KNT	77.27 ± 2.86	53.93 ± 2.40	59.45
2–KS8	76.62 ± 2.98	28.54 ± 0.86	28.57
2–CC6	74.64 ± 2.73	31.73 ± 0.91	35.01
3–KK	79.16 ± 2.35	46.86 ± 0.91	57.81
3–CH	78.95 ± 2.41	47.74 ± 2.26	52.45
4–CS	85.44 ± 1.09	64.10 ± 1.80	70.00
4–PE	92.18 ± 1.49	60.27 ± 0.58	76.06
5–HA	75.95 ± 2.04	31.16 ± 1.06	31.70
5–TREC	80.05 ± 2.05	52.50 ± 1.09	54.78
5–USDA1	81.88 ± 1.77	53.94 ± 1.19	63.94
5–USDA2	78.74 ± 1.72	45.16 ± 1.13	46.43
6–USDA3	79.92 ± 1.40	46.08 ± 0.71	64.16
6–FC	80.64 ± 1.31	41.80 ± 1.26	42.65

#: Soil samples are identified by a soil type number, followed by a transect descriptor (e.g., 1–FIU corresponds to soil type 1, transect FIU).

C. Geographic Information Systems: Semivariograms

Miami-Dade County samples illustrated limited spatial autocorrelation (Figure 7A). At the soil type scale, spatial variability varied with soil type 2, 4, 5 showing limited spatial autocorrelation and soil type 1, 3, 6 showing no spatial structure at the extent studied (Figure 7B). At transect scale, most transects demonstrated no spatial structure (i.e., PE in Figure 7C) at this scale, while OSP3 and FC illustrated a limited spatial autocorrelation. The three models tested (i.e., spherical, exponential, and Gaussian) displayed similar results (data not shown). All samples had a nugget effect ranging from 26-751 (Table 8 & 9). Those that showed a spatial autocorrelation in the semivariograms, illustrated in the table to have a partial sill, while those that showed no

spatial structure had a partial sill of 0. Partial sill (calculated by subtracting the sill and nugget) shows the variance of spatial autocorrelation without any nugget effect. The root mean square standardized (RMS std) for all samples was close to 1 and ranged from 0.75-1.01 (Table 8 & 9). All semivariograms tested had a RMS std less than one except Soil Type 1 that had a value of 1.01. A RMS std less than one illustrated a potential overestimation of the variability in the predictions. The Mean standardized error was close to 0 ranging from -0.17-0.18 which means that the predictions are unbiased and centered on the true values.

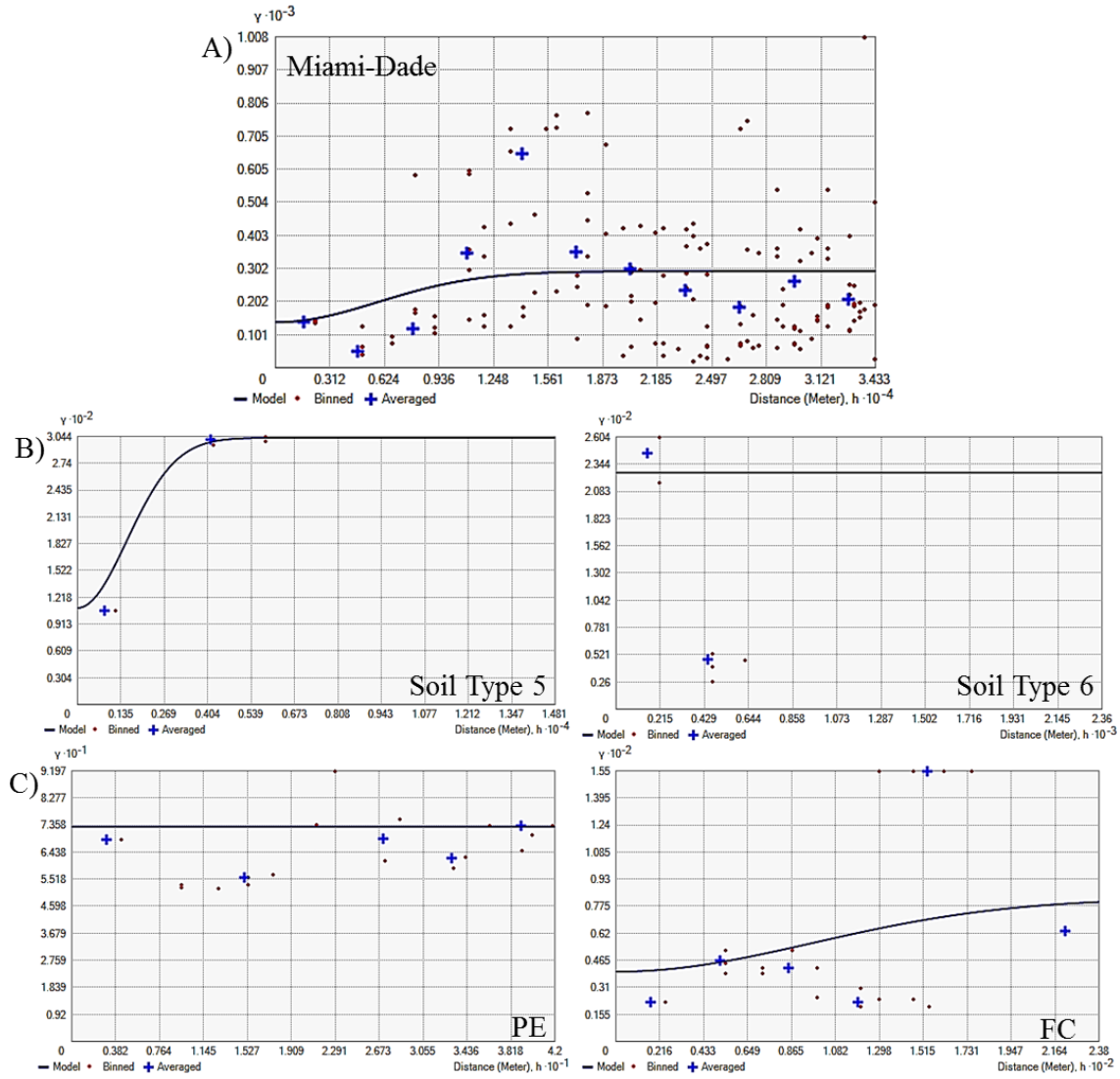

Figure 7. Semivariogram illustrating A) Miami-Dade spatial variability, B) soil type level spatial variability, C) Transect level variability.

Table 8. GIS semivariogram results of Miami-Dade County samples (plot) and soil type (ST) scales. Number of sample points per site/scale and maximum distance between samples were used to calculate number of lags and lag size for the semivariogram. Root mean square error (RMS), mean standardized error (Mean Std), root mean square standardized (RMS Std), and average standard error (Avg SE) was used to determine the best model. Nugget, partial sill, and range was used to determine the spatial variability at the extent.

Scale	# point	# Lags	Max pair dist (m)	Lag size (m)	RMS	Mean Std	RMS Std	Avg SE	Nugget	Partial Sill	Range (m)
Plot	1269	11	68669	3121	11.06	-0.02	0.91	12.22	138.68	155.33	14263

ST 1	303	9	30902	1717	18.22	-0.08	1.01	18.13	307.67	0	15453
ST 2	282	9	8010	445	8.24	-0.03	0.88	9.37	81.00	184.35	1519
ST 3	136	8	11083	693	13.37	0.02	0.95	14.12	181.11	0	5544
ST 4	134	8	12193	762	16.60	-0.02	0.90	18.36	302.77	58.84	48
ST 5	276	9	29627	1646	10.02	0.04	0.92	10.90	109.83	193.16	3729
ST 6	138	8	4718	295	14.85	0.01	0.95	15.58	226.37	0	2360

Table 9. GIS semivariogram results at transect scale. Number of sample points per site/scale and maximum distance between samples were used to calculate number of lags and lag size for the semivariogram. Root mean square error (RMS), mean standardized error (Mean Std), root mean square standardized (RMS Std), and average standard error (Avg SE) was used to determine the best model. Nugget, partial sill, and range was used to determine the spatial variability at the extent.

Scale	# point	# Lags	Max pair dist (m)	Lag Size (m)	RMS	Mean Std	RMS Std	Avg SE	Nugget	Partial Sill	Range (m)
FIU	100	8	105.69	7	25.61	-0.17	0.90	28.45	750.95	0	56
OSP1	65	7	229.10	16	11.67	0.02	0.95	12.27	140.26	0	112
OSP2	72	7	141.14	10	10.71	0.10	0.98	10.91	110.95	0	70
OSP3	66	7	372.00	27	10.47	0.07	0.80	13.05	148.34	187.89	189
NW	72	7	227.84	16	10.37	-0.01	0.92	11.34	118.25	6.90	112
KNT	67	7	172.76	12	11.01	0.14	0.79	14.07	176.23	3.22	38.39
KS8	72	7	47.41	3	4.33	0.18	0.82	5.29	25.82	0	21
CC6	71	7	306.83	22	5.84	-0.06	0.90	6.48	38.48	0	154
KK	67	7	83.25	6	14.90	0.01	0.95	15.71	221.14	0	42
CH	69	7	182.16	13	10.94	0.05	0.98	11.30	116.34	0	91
CS	69	7	98.76	7	23.49	0.09	0.91	25.88	613.30	0	49
PE	65	7	85.11	6	8.63	0.10	0.97	8.87	72.92	0	42
HA	72	7	89.46	6	9.81	0.10	0.88	11.08	113.94	0	42
TREC	64	7	102.09	7	9.36	0.02	0.90	10.41	99.26	0	49
USDA1	72	7	89.55	6	12.71	-0.08	0.82	15.62	221.55	0	42
USDA2	68	7	120.56	9	6.22	-0.07	0.87	7.16	46.60	0	63
USDA3	71	7	74.37	5	19.75	0.14	0.97	20.36	384.56	0	35
FC	67	7	476.90	34	5.15	-0.10	0.76	6.77	39.97	42.06	238

VI. Discussion

Three issues need to be addressed for sampling the soil spatial variability: location of sample points, size of sample, and total number of samples to be collected (Lin, Wheeler, Bell, & Wilding, 2005b). In sampling theory, spatial scale is defined by the

grain size, sampling interval, and extent. The grain size is the size of the sampling unit, sampling interval is the average distance between sampling units, and the extent is the total area included in the study (Franklin et al., 2002). Intuitively, researchers know that a large number of widely spaced samples are likely to be a better measurement of the spatial mean of a soil property rather than few samples located close to one another; however what is "a large number" or "widely spaced" (Loescher, Ayres, Duffy, Luo, & Brunke, 2014). Adequate spacing will ultimately minimize costs with increasing distance separation. However, it is also important to understand different sites spatial variability to see if different sampling designs are needed to accurately depict the soil site (Loescher et al., 2014). Previous studies have illustrated the differences of within site variability between homogeneous grassland over shrubland (Mummey & Stahl, 2003). Mummey & Stahl (2003) showed that homogeneous grasslands had a highly similar bacterial community and lower within site variability compared to the shrubland (Mummey & Stahl, 2003). In this study, variability within sites varied from 28-65% illustrating that the level of heterogeneity differed based on site. For example, site KK had a 46% within site dissimilarity compared to FIU that had a 62%. This can be attributed to the similar and almost identical microbial flora and fauna within some transects. Local heterogeneity can be due to different soil properties and multiple environmental factors such as unique plant species, sunlight amount, and differing moisture content (Franklin & Mills, 2003; Lenz & Foran, 2010).

Semivariograms can provide a useful tool for designing robust sampling strategies by estimating the variance (sill) that can be used to inform sample size in future studies as well as estimate the minimum distance required for samples to be considered spatially

independent (range) that can be used to inform sample spacing as explained by (Lin et al., 2005b). In this study, the multivariate, non-parametric LH-PCR microbial community profiles could not be calculated using the traditional autocorrelation function. Instead, the relative dissimilarity values were plotted with distance and took the form of the traditional correlogram as was described in various studies (Franklin & Mills, 2003; Mummey & Stahl, 2003). The semivariograms illustrated spatial autocorrelation for plot scale (all Miami Dade County samples), three soil types (2,4,5) and two transects (OSP3 and FC). For these sites, there was small scale heterogeneity observed that could be attributed to many small and sharply discontinuous distinct patches (Ettema & Wardle, 2002). However, the other three soil types (1,3,6) and most transects showed a pure nugget effect in where no spatial structure (zero partial sill) was observed at the spatial extent studied (Figure 7). Therefore, at the finer scales such as the PE site, no spatial variability was observed within the 42 meters extent and therefore, the interpolation (i.e., prediction) designated the samples within this site to have the same microbial communities in contrast to FC site that showed a spatial variability within the 238 meters and could distinguish samples within that distance (Figure 7C). Similar results have been found in previous studies by (Franklin & Mills, 2003; Loescher et al., 2014). Franklin & Mills (2003) found that the spatial distribution of the community was different based on the scales used with finer scales not able to detect spatial patterns. This can occur due to random sampling variance or variability that is occurring at other spatial scales not examined in the spatial extent. For instance, more samples at greater distances (>42 meters) are needed to see the spatial variability at the PE site. The decreased number of samples at the transect scale (N=65-100), compared to the soil type (N=134-303) and plot

scale (N=1269) could also limit our ability to detect fine scale relationships. A 'rule of thumb' in geostatistics suggest that each class should contain at least thirty pair of points (Franklin & Mills, 2003). The greater the number of points, the greater the statistical reliability is. This could be a reason as to why spatial variability was not observed at the finer resolution (i.e., transect) as only 12 samples were collected for each of the six locations.

For forensics, it is important to understand the scale at which the soil microbial community must be measured to create a sampling design that will result in sound discrimination Geostatistics can assist in assessing the spatial variability and offer an index to quantify the magnitude and scale of spatial variation in a soil property (i.e., microbial community profiles). The significance of this for forensics links back to the issue of soil variability at the crime scene and how realistic it is to expect a soil sample collected by an investigator to be similar to a questioned sample (Lark & Rawlins, 2008). GIS is increasingly being used to integrate and analyze data. However, robust databases and sampling schemes are needed for forensic purpose and spatial resolution, amount of material, and condition of sample collection need to be addressed (McKinley, 2013). This can be a useful technology that can capture a wide range of useful soil properties and incorporate the results into a common format that can be quantitatively measured at low cost. Kriging interpolation methods under GIS can be then used to predict values at unsampled locations using the theoretical semivariograms depicting the spatial variability. Kriging estimates linear combinations of the data with weights from the model semivariogram (Tsiknia et al., 2014).

Further research is needed to understand the number of samples needed to represent the population and the discriminatory capacity to determine if one test fits all or if the model needs to be tuned to fit particular soil types or geographic situations. This study recommends a hierarchical sampling approach similar to the one used in this study to catalog the soil spatial variability at multiple scales so that an understanding the soil variability across different landscapes, but also at what scale the variability is most likely to occur (Lin et al., 2005b). These results showed a snapshot of the relationship at various soil sites at a single time, and therefore, did not consider the temporal variability or its interaction with spatial heterogeneity in determining the community pattern (Franklin & Mills, 2009). Further studies are needed to examine both spatial and temporal scales simultaneously to determine the usefulness of this technique over time. Currently there is no comprehensive soil microbial community profiling database and very few published attempts to develop databases of soil properties (i.e., chemical and physical) specifically with forensic application in mind (Pye & Blott, 2009). Soil databases can be useful and suitable for forensic inferences; however, as (McKinley, 2013) states they need to be consistent, compatible, and applicable to be useful in forensic cases.

VII. References

Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Cech, M., Chilton, J., Clements, D., Coraor, N., Eberhard, C., Cruning, B., Guerler, A., Hillman-Jackson, J., Von Kuster, G., Rasche, E., Soranzo, N., Turaga, N., Taylor, J., Nekrutenko, A., Goecks, J. (2016). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Research, 44*(W1), W3-W10. doi:10.1093/nar/gkw343 [doi]

Cocolin, L., Manzano, M., Cantoni, C., & Comi, G. (2001). Denaturing gradient gel electrophoresis analysis of the 16S rRNA gene V1 region to monitor dynamic

changes in the bacterial population during fermentation of Italian sausages. *Applied and Environmental Microbiology, 67*(11), 5113-5121.

Damaso, N., Mendel, J., Mendoza, M., Oliva, Y., Diaz, A., von Wettberg, E. J., Narasimhan, G., Mills, D. (in review). A comparison of machine learning algorithms for the classification and provenance of soil samples based on biotic content. *Forensic Science International,*

DeLong, E. F. (1992). Archaea in coastal marine environments. *Proceedings of the National Academy of Sciences, 89*(12), 5685-5689.

Dray, S., & Dufour, A. B. (2007). The ade4 package: Implementing the duality diagram for ecologists. *Journal of Statistical Software, 22*(4), 1-20.

Ettema, C. H., & Wardle, D. A. (2002). Spatial soil ecology. *Trends in Ecology & Evolution, 17*(4), 177-183.

Fierer, N., & Jackson, R. B. (2006). The diversity and biogeography of soil bacterial communities. *Proceedings of the National Academy of Sciences, 103*(3), 626-631.

Franklin, R. B., & Mills, A. L. (2009). Importance of spatially structured environmental heterogeneity in controlling microbial community composition at small spatial scales in an agricultural field. *Soil Biology and Biochemistry, 41*(9), 1833-1840.

Franklin, R. B., Blum, L. K., McComb, A. C., & Mills, A. L. (2002). A geostatistical analysis of small-scale spatial variability in bacterial abundance and community structure in salt marsh creek bank sediments. *FEMS Microbiology Ecology, 42*(1), 71-80. doi:S0168-6496(02)00320-3 [pii]

Franklin, R. B., & Mills, A. L. (2003). Multi-scale variation in spatial heterogeneity for microbial community structure in an eastern Virginia agricultural field. *FEMS Microbiology Ecology, 44*(3), 335-346. doi:S0168-6496(03)00074-6 [pii]

Goovaerts, P. (1998). Geostatistical tools for characterizing the spatial variability of microbiological and physico-chemical soil properties. *Biology and Fertility of Soils, 27*(4), 315-334.

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

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

Lark, R., & Rawlins, B. (2008). Can we predict the provenance of a soil sample for forensic purposes by reference to a spatial database? *European Journal of Soil Science, 59*(5), 1000-1006.

Lenz, E. J., & Foran, D. R. (2010). Bacterial profiling of soil using genus-specific markers and multidimensional scaling. *Journal of Forensic Sciences, 55*(6), 1437-1442.

Lin, H., Wheeler, D., Bell, J., & Wilding, L. (2005a). Assessment of soil spatial variability at multiple scales. *Ecological Modelling, 182*(3), 271-290.

Lin, H., Wheeler, D., Bell, J., & Wilding, L. (2005b). Assessment of soil spatial variability at multiple scales. *Ecological Modelling, 182*(3), 271-290.

Loescher, H., Ayres, E., Duffy, P., Luo, H., & Brunke, M. (2014). Spatial variation in soil properties among North American ecosystems and guidelines for sampling designs. *PloS One, 9*(1), e83216.

Martiny, J. B., Eisen, J. A., Penn, K., Allison, S. D., & Horner-Devine, M. C. (2011). Drivers of bacterial beta-diversity depend on spatial scale. *Proceedings of the National Academy of Sciences of the United States of America, 108*(19), 7850-7854. doi:10.1073/pnas.1016308108 [doi]

McKinley, J. (2013). How useful are databases in environmental and criminal forensics? *Geological Society, London, Special Publications, 384*(1), 109-119.

Moreno, L. I., Mills, D. K., Entry, J., Sautter, R. T., & Mathee, K. (2006). Microbial metagenome profiling using amplicon length heterogeneity-polymerase chain reaction proves more effective than elemental analysis in discriminating soil specimens. *Journal of Forensic Sciences, 51*(6), 1315-1322.

Mummey, D. L., & Stahl, P. D. (2003). Spatial and temporal variability of bacterial 16S rDNA-based T-RFLP patterns derived from soil of two Wyoming grassland ecosystems. *FEMS Microbiology Ecology, 46*(1), 113-120. doi:10.1016/S0168-6496(03)00208-3

Noble, C. V., Drew, R. W., & Slabaugh, J. D. (1996). Soil survey of Dade County area, Florida. *USDA NRCS, Gainesville, FL,*

Pye, K., & Blott, S. J. (2009). Development of a searchable major and trace element database for use in forensic soil comparisons. *Science & Justice, 49*(3), 170-181.

Sensabaugh, G. F. (2009). Microbial community profiling for the characterisation of soil evidence: Forensic considerations. *Criminal and Environmental Soil Forensics* (pp. 49-60) Springer.

Suzuki, M., Rappe, M. S., & Giovannoni, S. J. (1998). Kinetic bias in estimates of coastal picoplankton community structure obtained by measurements of small-subunit rRNA gene PCR amplicon length heterogeneity. *Applied and Environmental Microbiology, 64*(11), 4522-4529.

Taberlet, P., Gielly, L., Pautou, G., & Bouvet, J. (1991). Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant Molecular Biology, 17*(5), 1105-1109.

Tsiknia, M., Paranychianakis, N. V., Varouchakis, E. A., Moraetis, D., & Nikolaidis, N. P. (2014). Environmental drivers of soil microbial community distribution at the Koiliaris critical zone observatory. *FEMS Microbiology Ecology, 90*(1), 139-152. doi:10.1111/1574-6941.12379 [doi]

Violle, C., Reich, P. B., Pacala, S. W., Enquist, B. J., & Kattge, J. (2014). The emergence and promise of functional biogeography. *Proceedings of the National Academy of Sciences of the United States of America, 111*(38), 13690-13696. doi:10.1073/pnas.1415442111 [doi]

White, T. J., Bruns, T., Lee, S., & Taylor, J. (1990). Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. *PCR Protocols: A Guide to Methods and Applications, 18*, 315-322.

Yang, C., Mills, D., Mathee, K., Wang, Y., Jayachandran, K., Sikaroodi, M., Gillevet, P., Entry, J., Narasimhan, G. (2006). An ecoinformatics tool for microbial community studies: Supervised classification of amplicon length heterogeneity (ALH) profiles of 16S rRNA. *Journal of Microbiological Methods, 65*(1), 49-62.

Chapter 3: Assessing temporal variability and DNA marker selection for forensic soil provenance applications

This chapter has been formatted for: Forensic Science International.

Natalie Damaso[1,2], Yu Cheung[2], and DeEtta Mills[1,2*]

[1]Department of Biological Sciences, Florida International University, Miami, Florida, United States of America

[2]International Forensic Research Institute, Florida International University, Miami, Florida, United States of America

I. Introduction

Soil particles can provide valuable information when recovered from a crime scene. Its value is due to its prevalence, high transferability, and retention probability that can easily adhere to objects such as footwear and car tires that are often overlooked by a suspect in attempts to conceal evidence [1]. With the advances in molecular genomics, the forensic potential for using microorganisms to query soil provenance for intelligence assessments or to establish an evidentiary link between suspect and soil evidence have been increasing [1-2]. Microorganisms are abundant and ubiquitous in all environments and can therefore serve as a powerful source of trace evidence [2-6]. Microbial community profiling can be done using a very small sample size (~50-500 mg) with the DNA expertise and instrumentation already employed in many laboratories [7]. Previous research has shown the potential to use microbial community profiling in forensics to link soil evidence samples to its origin [8-11]. Although microbial soil profiling has been promising, its application is still in its infancy and further research needs to be conducted to develop standard operating procedures (SOPs) for the collection, analysis, and interpretation of microbial forensic evidence for it to be acceptable in a court of law [2].

For a robust tool to be applied in forensic application, an understanding of the uncertainty associated with any comparisons and the parameters that can significantly influence variability in profiles needs to be determined. These issues include selecting suitable microbial markers and the influence of temporal variability on the DNA profile. Most often soil forensic analyses have exclusively looked at bacteria [10,16,17]. However, fungi have been recently shown to be robust for soil forensic discrimination

and resistant to biological, chemical, and mechanical degradation [6,11,14]. As in human identification, the more DNA markers queried the greater the discrimination power. Previously, MacDonald et al. (2008) illustrated a multiplex approach that analyzed bacteria, archaea, and fungi, which led to better site discrimination [14]. Bacteria provided greater resolution between two sites, but were more susceptible to air-drying, and sensitive to dehydration pressures that lead to population shifts. Fungi were less altered by air-drying, resilient to desiccation, tolerant to a wide range of pH (i.e., persist in acidic soils), and provided discrimination between sites [18]. Lastly, Archaea were useful to identify saline or water logged soil environments. Therefore, a multi-taxon approach can provide a different level of discrimination [14] and has the potential for forensics to link soil evidence to its origin. The first objective of this study was to analyze four taxa (bacteria, fungi, archaea, and plant) individually and compare to a multi-taxon approach to determine which would provide the highest degree of discrimination between and within sites.

An assumption underlying the use of microbial profiling is that there should be limited temporal variability as soil and its biotic communities should not change substantially over time in order to use pattern modeling for forensic application [12]. Therefore, the reliability of this approach needs to be tested to determine if major spatio-temporal changes in a soil's microbial community could have an effect on its probative value [2]. Soils are extremely complex environments that exhibit substantial spatial and temporal heterogeneity [13]; however, it has been shown that spatial variability is more significant than temporal variability [12,14]. If the microbial community changes substantially over time, origin of evidentiary soil may be excluded in error and it may be

more difficult to link soil from a suspect to a specific location if not analyzed within a certain time frame [12,15].

Previous results have assessed short-term (1-1.5 yr) spatio-temporal variability of soil communities and showed that biotic content was correlated to soil type and to specific transects with strong accuracy using pattern analyses and machine learning algorithms (Damaso et al. 2016 In Review). However, if archived data and training sets are to be useful long term, temporal variability (> 2 yr) also needs to be considered. Unlike human identification, soil environment is dynamic and changes over time. Therefore, it is important to see if meaningful comparisons and links can still be made between soil evidence deposited at the crime and archived reference data previously collected (> 2 yr) from a site can still be classified [7]. The second objective of this study was to determine if there was temporal site variability observed in the microbial communities from freshly sampled soils after a four-year time span (2010 to 2014). This is vital as microbial communities need to be stable enough over a reasonable time span if they are to be useful for forensic purposes. The goal was to characterize the temporal dynamics of microbial communities from three previously sampled sites to establish how variable the communities may be over time.

II. Materials and Methods

In this study, universal primer sets were selected that have broad specificity for organisms known to be ubiquitous in soil. Bacteria, archaea, fungi, and plant universal DNA markers were PCR amplified, separated by capillary electrophoresis, and queried across three soil types in Miami-Dade County, Florida over two seasons (dry and wet) in

2010 and again four years later in 2014. Abiotic information such as pH, organic matter content, moisture content, and soil texture was also obtained from the soil.

A. Soil Collection

In 2010-2011, sites across Miami-Dade County, Florida were sampled during the dry and wet seasons (Damaso et al. 2016 In Review). Three sites were again sampled during the dry and wet seasons in 2014 and 216 samples were collected and analyzed. Sites included (as categorized by USDA-NRCS [19]): FIU as soil type 1 (Urban Land-Udorthents), CC6 from soil type 2 (Lauderhill Dania-Pahokee), and KK from soil type 3 (Rock Outcrop-Biscayne-Chekika). Sites were established in undisturbed sites that had limited public access in the three different soil types. Each site was at least ≥ 1.6 km distant from the next, 100 m in length and six subplots within each site were randomly sampled. GPS coordinates for every subplot were recorded. Within each subplot, six cored samples were taken within a 1 m^2 quadrat using a 5 cm diameter soil corer to collect the top 5-10 cm of the soil. The soil samples were transported back to the laboratory and sieved to remove large objects and debris.

B. Abiotic Analysis

Soil texture (% sand, silt, and clay) was obtained for each site for the wet season in 2014. Percent moisture, percent total organic content (TOC), and pH were obtained for subplot level for each season (dry and wet) in 2014. Soil texture was obtained using the Bouyocos hydrometer method. The pH was measured in soil-water (1:2) solutions. Soil slurries were made by adding 3 g of soil to 6 ml of distilled water, stirred, and measured using the calibrated electrode/digital pH meter (LaMotte, Chestertown, MD). The moisture content was determined gravimetrically by oven drying the soils at ~55°C for 24

hours. The TOC was determined gravimetrically by measuring the difference between the dry weight and the ash-free dry weight that was obtained by igniting the dry soils in an ash oven at ~550°C for 4-5 hours. The percent carbon is 50% of the ash free dry mass for plant matter. Student two-sample t-tests were conducted to observe if there was a significant difference seasonally for each of the abiotic parameters.

C. DNA Extraction

Extraction was conducted using the BIO 101 Fast DNA Spin Kit for Soil® and FastPrep®-24 System homogenizer (MP Bio, Solon, OH). Quantification was performed using the Qubit® Assay kit on the Qubit 2.0 Fluorometer (Invitrogen, Carlsbad, CA). Samples were diluted to a 20 ng/µl working stock.

D. Length Heterogeneity-Polymerase Chain Reaction

Two samples per subplot from each season (dry and wet) for the three sites (FIU, CC6, and KK) were used to test the temporal variability of the soil (n=72). DNA was amplified as described in Damaso et al. 2016 In Review, using two PCR duplexes: (1) bacteria and fungi, and (2) Archaea and plant. PCR reaction mixtures were: 1X reaction buffer, 2.5 mM $MgCl_2$, 0.25 mM dNTPs (Promega, Madison, WI), 1% BSA (Fraction V, Fisher Scientific, Pittsburgh, PA), 1% DMSO (Promega, Madison, WI), various concentrations of primers (bacteria=0.5 µM, fungi=0.4 µM, Archaea=0.4 µM, plant=0.6 µM), 40 ng DNA, and 0.5 U AmpliTaq Gold® DNA Polymerase (Applied Biosystems, Foster City, CA). Universal primers were used for the following genomic regions for each taxa: 16S rRNA for bacteria (V1 +V2 domains, 27-F, 355-R) [20] and Archaea (V1-V3 domains, 21-F, 518-R) [21,22], ribosomal internal transcribed spacer region (ITS) for fungi (ITS5-F, ITS2-R) [23], and chloroplast *trn*L intergenic region for plant (trnL-F,

trnL-R) [24]. Forward primers were labeled with 6-FAM fluorescent dye. Each duplex was amplified with the same program using the ABI 9700™ thermocycler (Applied Biosystems, Foster City, CA) with the following parameters: initial 10 min denaturing step at 95°C, 25 cycles of denaturation at 95°C, annealing at 54°C, and extension at 74°C each for 30 sec with a final extension at 74°C for 10 min. In addition to the duplexes described above, each individual taxon was also amplified separately to determine the discrimination power of each taxon.

E. Capillary Electrophoresis

Fragment analysis was conducted using the ABI Prism™ 3130xl (Applied Biosystems, Foster City, CA) using Performance Optimized Polymer 7 (POP7) (Applied Biosystems, Foster City, CA). Samples from the two duplexes were co-loaded where 1 µl of each PCR product was added to a mixture of 11.5 µl Hi-Di™ Formamide (Applied Biosystems, Foster City, CA) and 0.65 µl internal size standard, GeneScan LIZ®600 (Applied Biosystems, Foster City, CA), denatured by heating for 2 min at 95°C and then snap-cooled on ice for 2 min. CE preparation and separation were conducted using the same parameters as the multiplex approach without co-loading the samples for the individual taxa.

Raw data were analyzed using GeneMapper™ v 4.0 (Applied Biosystems, Foster City, CA). Local Southern size calling was used for the analysis parameters with a minimum threshold of 50 relative fluorescent units (RFUs). The relative ratios were calculated by normalizing the heights of each peak in the profile to the total peak intensities resulting in the ratio for each peak height as a decimal value from zero to one using the Galaxy ABI 310 Data Formatting tool found in http://usegalaxy.org/ [25].

F. Statistical Analysis

All analyses were conducted using Primer-E v.7 software (PRIMER E Ltd., Plymouth Marine Laboratory, Plymouth, U.K.). Bray-Curtis similarity matrices were generated on relative abundance ratios that had been square-root transformed prior to analysis. Analysis of Similarity (ANOSIM) was used to determine the significant effect of time as well as the significant differences between sites based on individual taxa and the combined four-taxa profiles. ANOSIM reports the level of dissimilarity between samples groups (Global R) and the associated level of significance (p) to provide statistical pair-wise comparisons between designated groups. The ANOSIM R-statistic indicates the level of discrimination between groups (sites), with a value close to one indicating complete group discrimination and a value close to zero implying no differences exist between groups [26]. The associated significance level (p) is equivalent to the p-value where 0.1%, 1%, and 5% is equal to $p<0.001$, $p<0.01$, and $p<0.05$, respectively. Non-metric Multidimensional Scaling (nMDS) was used to visualize the site heterogeneity, temporal and seasonal variability, and discrimination power of each taxon to distinguish sites apart. Non-metric Multidimensional Scaling (nMDS) allows complex datasets to be easily visualized, with more similar samples grouping together. The level of confidence in the 2-D plot is indicated by the stress, i.e. <0.2 provides good representation of the fit [1]. Similarity Percentages (SIMPER) analysis was used to identify the LH-PCR peaks contributing to the dissimilarity between sites and temporal variability as well as the percent contribution each amplicon provided for the overall dissimilarities.

G. Random Forest

Random Forest analysis was conducted in R programming language using the randomForest package [27] for soil classification. To determine the temporal effect, the 2010-2011 dataset was used for training and the 2014 four-taxon dataset was used for validation/testing. When determining the individual taxon (bacteria, Archaea, fungi, and plant) as well as the four-taxa discrimination power, two thirds of the 2014 dataset were used for training the algorithm and one third was used for testing each replicate run for each dataset. For reproducibility, the datasets were re-tested by randomly selecting a different training and testing set three different times. Classification accuracy was used to determine the performance of the classification method for each taxon and the multi-taxa. Classification accuracy calculated the percent of samples correctly classified. Student two-sample T-tests were conducted to determine significant differences between different taxon classifications.

III. Results

A. Taxa Discrimination (2014)

To test which taxa would give the best site discrimination, all taxa were examined individually as well as in combination. The nMDS showed a discrete spatial separation between sites using fungi (Figure 8C). Bacteria was able to group KK site and FIU based on seasons (Figure 8B). Four taxa were able to show discrete spatial separation between sites as well as the seasons within the sites (Figure 8A). These results show that four taxa combined the site discrimination of fungi and the seasonal discrimination of bacteria. ANOSIM statistic also supported nMDS results illustrating that fungal profiles were significantly influenced by soil site ($R=0.47$, $p=0.1\%$), whereas four taxa profiles were

significantly influenced by soil sites and seasons within sites (R=0.58, p=0.1%). Archaea and plant profiles were poor, observing few LH-PCR peaks with low signal intensity and therefore, did not provide a clear distinction between soil sites with exception of KK site that grouped using Archaea markers (Figure 8D). Even though nonparametric tests could not differentiate sites clearly using the individual taxa other than fungi and four-taxa, Random Forest was able to classify the soils based on site and classified with accuracy of 72%, 36%, 90%, 98%, and 95% for Archaea, plant, bacteria, fungi, and four-taxa, respectively (Figure 9). Archaea, bacteria, fungi, and four-taxa did not show a significant difference in classification accuracy ($p>0.07$). Plant had significantly lower classification accuracy (36% accuracy, $p<0.009$) than the other individual taxa as well as the four-taxa combined.

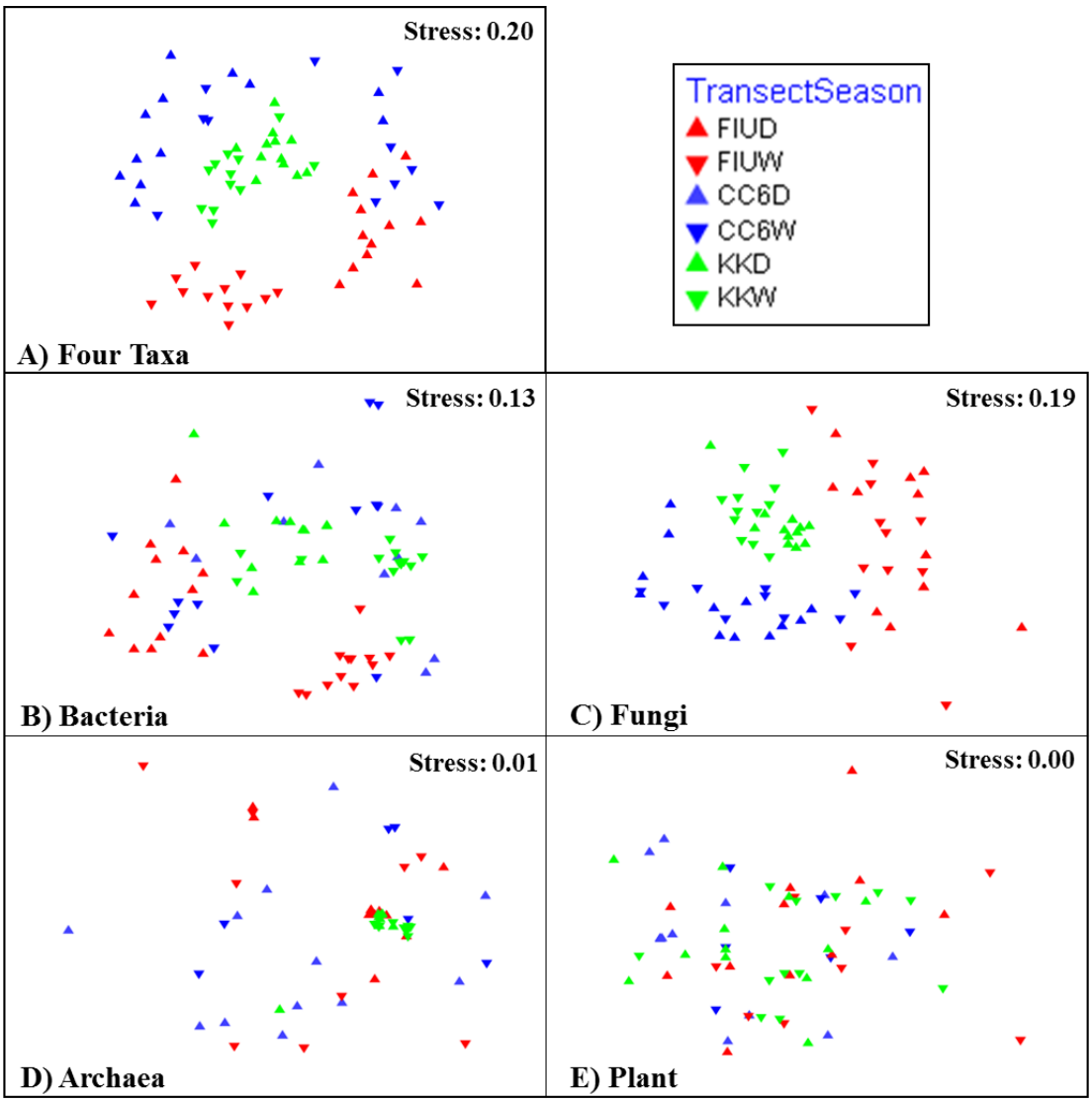

Figure 8. Nonmetric Multidimensional Scaling 2-D plots illustrating the discrimination power to distinguish three sites (Red=FIU, Blue= CC6, Green= KK) and season (▲Dry/▼Wet). A) Four-taxa was able to discriminate sites and seasons within a site. B) Bacteria marker was able to group KK and group FIU based on season. C) Fungi was the best marker to discriminate the three sites. D) Archaea was unable to discriminate FIU and CC6; however, it was able to group KK. E) Plant was unable to distinguish the three soil sites apart.

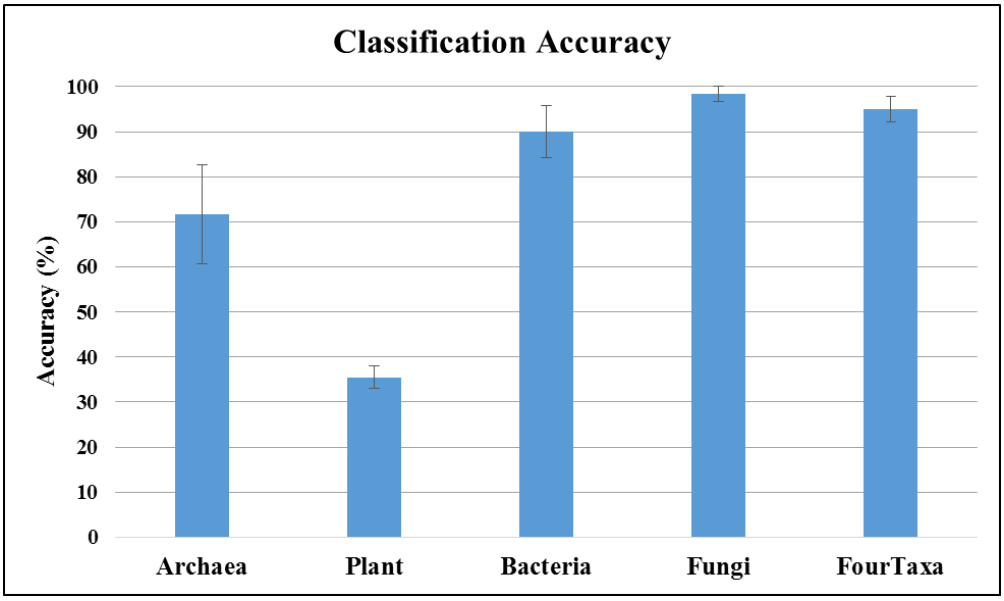

Figure 9. Random Forest classification accuracy using individual taxa and four-taxa to discriminate three sites (FIU, CC6, and KK). Archaea, bacteria, fungi, and four-taxa did not show a significant difference in classification accuracy ($p>0.07$).

B. Abiotic Seasonal Variability

Soil texture, pH, moisture content, and total organic content were collected to determine the characteristics of the soil samples and the abiotic seasonal (dry and wet) variability. Results indicate that there were no significant differences seasonally for pH and total organic content of the soils (Table 10). However, moisture was significantly different seasonally with exception of CC6 that did not have a significant difference between dry and wet season (Table 10). Both FIU and KK are sandy loam while CC6 is sandy clay loam in texture (Table 10).

Table 10. Abiotic seasonal (dry and wet) variability for three sites (FIU, CC6, and KK). Soil texture classification based on the % sand, silt, and clay for each site collected in 2014. pH and total organic content illustrated no significant difference between seasons. Moisture was significantly different seasonally except for CC6. Parenthesis represent standard error.

		FIU	**CC6**	**KK**
Soil Texture		Sandy Loam	Sandy Clay Loam	Sandy Loam
pH	**Dry**	7.48 (± 0.12)	7.68 (± 0.07)	7.88 (± 0.02)

	Wet	7.47 (± 0.12)	7.64 (± 0.10)	7.91 (± 0.03)
Moisture (%)	**Dry**	17.31 (± 2.34)	44.93 (± 4.16)	21.07 (± 1.80)
	Wet	23.67 (± 4.38)	48.02 (± 6.46)	29.53 (± 1.78)
Organic (%)	**Dry**	9.86 (± 1.71)	24.54 (± 4.21)	13.17 (± 0.47)
	Wet	12.34 (± 4.80)	26.80 (± 6.31)	13.38 (± 0.82)

C. Biotic Temporal and Seasonal Variability

Temporal variability was observed for the three sites (FIU, CC6, and KK) between 2010 and 2014, however sites still grouped based on location with exception of CC6 (Figure 10). ANOSIM results showed that there was a significant temporal variability (p=0.1%) with a global R of 0.68, 0.81, 0.68 for FIU, CC6, and KK, respectively. SIMPER results showed that FIU, CC6, and KK sites were 79%, 96%, and 84% dissimilar across time (between 2010 and 2014 profiles), respectively. Nonmetric Multidimensional Scaling (nMDS) illustrated the temporal variability observed for each site, however FIU and KK still grouped regardless of the four year time span (Figure 10).

When the *combined 2010 dataset* (all six soil types and seasons) was used as the Random Forest training set to classify the 2014 sites to their origin, the algorithm classified FIU collected in 2014 to its proper origin with only 16% accuracy, while KK and CC6 were not able to be classified at all. Using just the *soil type subsets* (including both seasons) of the 2010 dataset to train (i.e., soil type 1-2010 for FIU, soil type 2-2010 for CC6, soil type 3-2010 for KK), classification accuracy increased to 83% for FIU and 71% for KK; however, CC6 was still unable to be correctly classified. When the *same soil type and season were used as the training set* (i.e., soil type 1 samples collected in 2010 wet season for FIU samples collected in the wet season in 2014), FIU classification accuracy increased to 100% for wet season and 67% for dry season. KK accuracy was

8% for wet season and 42% for dry season. CC6 however, was not able to be correctly classified. Google Earth images and ground truthing showed that FIU's above ground mixed forest plant community had not changed over the four year time span; however, KK had an increase in vegetation over time and CC6 had a major disturbance within the four year time span.

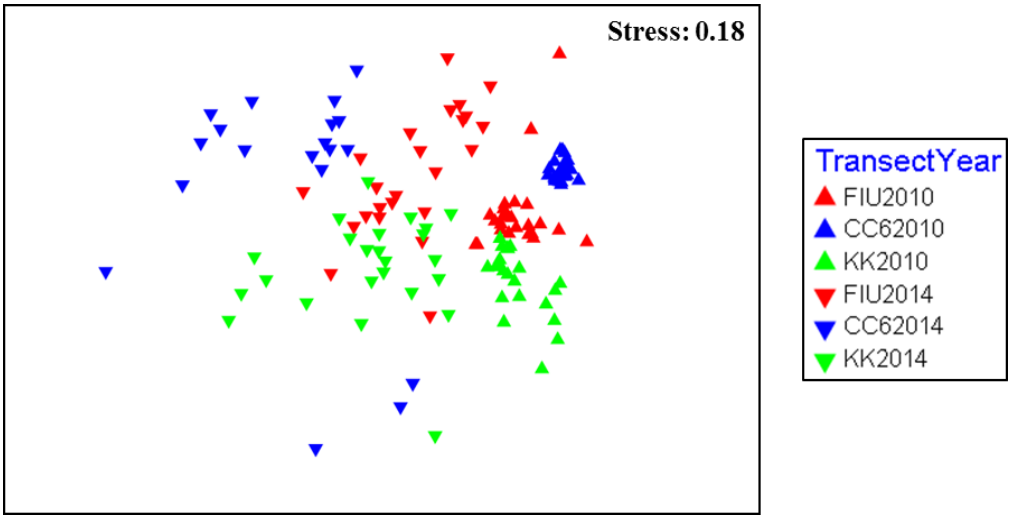

Figure 10. Temporal variability within three sites (Red=FIU, Blue=CC6, and Green=KK) across a four year time-span (▲2010/▼2014) based on Nonmetric Multidimensional Scaling (nMDS) analysis using Bray-Curtis similarity coefficient.

IV. Discussion

Microbial community profiling studies have been promising using bacterial markers alone to distinguish between soil types and has been the standard marker [8,11,12,16]. However, additional taxonomic groups can provide further discriminatory power and requires further investigation. As in human identification, the more DNA markers queried the greater the discrimination power. In this study, we assessed the resolution of bacteria, archaea, fungal, and plant community profiles independently and combined to determine the best marker or markers for forensic comparison of soil evidence. Ideal markers for soil provenance should be sufficiently variable to

discriminate among samples but conserved enough to be less variable within than between species, be robust, and have highly reliable DNA amplifications [41].

A majority of the studies using soil microbial profiling for forensic applications involve the bacterial 16S rRNA genes using T-RFLP [14,31,42] or LH-PCR [8], ribosomal internal transcribed spacer region (ITS) for fungi [43], Archaea 16S rRNA, and[43], the chloroplast *trnL* intergenetic region for plant [41]. In this study, fungi provided clear discrimination between sites using nMDS and ANOSIM as did the multi-taxon approach (Figure 1C). However, the four taxa concatenated were able to better discriminate between transects and seasonally within a transect (Figure 1A). Unlike fungi, bacteria did not provide a clear discrimination between sites as these markers have a high level of variability within and among sites due to their heterogeneous nature [11,15,30]. Targeting only bacteria have been shown to produce too much noise to be useful in differentiating between sites [12] as bacteria form micro-spatial niches within micro-aggregates of soil particles; therefore, their distribution is more heterogeneous than non-bacterial taxa [6,12]. Bacteria can generate a high site-specific DNA profile as their structure can be influenced by soil type, seasonal variation, site management, vegetation cover, and environmental conditions [1,17,44]. Plant and Archaea were not able to distinguish the transects and had a low number of OTU as was found in previous studies [6,14]. Even though Archaea and plant had low informational value within the profiles and were unable to differentiate between the three sites, they may still be useful markers in specific cases. For example, *trnL* marker can be useful when detecting the presence of a certain plant species. Wetland sites are found to have significantly higher Archaea [6] and the 16S rRNA Archaea marker can be useful, as shown with the KK site, to identify

and distinguish soils from water logged environments where Archaea have a greater presence [14].

Using Random Forest machine learning tool, the soils were able to be classified with over 95% accuracy using fungi and the four-taxa. Moreover, even though nonparametric tests, such as nMDS, could not differentiate sites clearly using bacteria and Archaea, Random Forest was able to classify the soils based on site with high accuracy and was not significantly different than fungi and four taxa in their classification accuracies (Figure 2). This is supported by previous research that found that multiplexing bacteria, fungi, and Archaea led to better discrimination compared to soil color and single taxa profiling, but was not significant when classifying using linear discriminant analysis [14]. In conclusion, even though fungi look promising for single taxon soil discrimination, the additional markers can help discriminate between a wide variety of soil types. Plants can assist to link a certain plant species to the site, Archaea can indicate water logged or extreme environments, and the core bacteria can be useful at site specific and when seasonal discrimination is needed [1,45].

This study shows the potential benefit of utilizing classification tools and comprehensible reference database to distinguish soil samples and determine their geographic origin. However, it is also important to understand the manner and level that the communities change temporally if they are going to be used as markers. This is critical to understand how frequently a reference dataset needs to be updated. Extensive study is required using different ecosystems to evaluate the stability of microbial DNA profiles to determine the maximum time that can elapse between sample deposition and reliable comparison with collected samples [16]. Difference in profiles with time is

expected as temporal and seasonal fluctuations such as rainfall and temperature, may impact the microbial community causing population shift [10,35]. Previous studies have tested the temporal variability of the soil; however, they were restricted within one year and looked at the presence/absence similarity [10,15,16]. Lenz & Foran 2010 found that known soil samples can potentially be collected well after a crime occurred throughout a one-year period as the time/season did not have a substantial negative influence on the ability to group soils [15]. Horswell et al. 2002 also found that soil samples collected eight months apart had somewhat dissimilar bacterial TRFLP profiles; however, they still showed a high degree of similarity (70% (8 months) compared to 90% (time of collection)) [10]. The apparent stability of the bacterial profiles could be attributed to the bacterial mechanisms of resistance and dormancy. Even though different species alternate between growth and dormancy based on environmental changes, the cells/DNA will still be present in the soil and can be detected with DNA profiling methods such as TRFLP or LH-PCR [16].

In this study, four year temporal variability as well as the relative abundance similarities across three different soil types were examined. Results indicated that using ANOSIM, there was a significant temporal variability observed between 2010 and 2014 for all sites. Multidimensional scaling (nMDS) also illustrated a temporal variability for each site; however, FIU and KK still grouped regardless of the four-year time span (Figure 3). Using Random Forest machine learning tool, which finds hidden patterns of the microbial communities, it was able to correctly classify the soils based on location regardless of the temporal variability when no disturbances occurred during the time span. One of the three sites, CC6, was unable to be correctly classified as a result of

major disturbances throughout the years. Other two sites were able to be classified with FIU having higher classification accuracy compared to KK as FIU's above ground mixed forest plant community had not changed over the four year time span while KK had an increase in vegetation over time. Overall this study showed that stable profiles may allow comparison between evidence and a possible crime scene despite the time lapse between sample collections. However, this is dependent on the analysis method, site, vegetation, and level of disturbance [16]. Therefore, temporal variability of the soil microbial communities and how this variability compares among different soil types is important to understand.

Research suggests that microbial communities exhibit a wide range of discernable temporal patterns that reflect underlying biotic and abiotic processes [36]. Meta-analysis by Shade et al. (2013) showed that microbial communities' temporal dynamics are dependent on habitat type. Previous study by Lauer et al. (2013) found that land use type and vegetation dynamics played a large role in modulating the temporal variability of the soil bacterial community [37]. Soil texture has also been found to influence temporal variations; Pereria e Silva et al. (2012) found that temporal variations were higher in clayey soils than in sandy ones for archaeal and bacterial communities. Therefore, the temporal variability of CC6 can be attributed to the soil texture of CC6 that is a sandy clay loam soil unlike, FIU and KK that are sandy loam soils. In this study, abiotic properties such as pH, moisture, organic content, and soil texture were collected at each season in 2014 to determine the seasonal variability of the physical properties of the different soils and relate it to the microbial community. Results indicated that there were no significant differences seasonally in pH or total organic content. However, moisture

did have a significant difference seasonally with exception to CC6 that did not have a seasonal difference. This can be attributed to the soil texture of CC6 as soil texture has been shown to have a relationship with moisture content as smaller soil particles such as clay have a larger surface area and therefore have a higher water holding capacity than larger sand particles [29]. Therefore, as CC6 has more clay particles than the other transects, it should have a higher water holding capacity and would not experience a significant difference seasonally. CC6 also had higher organic content compared to the other two transects (average ~25% for CC6 compared with 10% and 12% for FIU and KK, respectively). Humic acid, a known PCR inhibitor, have been shown to be present in higher quantities in soils with high organic matter and can introduce bias and result in lower diversity estimates [16]. The results suggest that abiotic and biotic factors determine the community assembly of these communities.

Understanding the temporal patterns of the communities have been fundamental in ecology to anticipate the responses of ecosystems to global change and disturbance [38,39]. Small disturbances can affect the soil microbial community at different temporal and spatial scales. Even though microorganisms are ubiquitous, abundant, and have critical roles in ecosystems, their temporal dynamics is largely unknown. This study determined that although temporal variability was observed throughout a four-year time span, without drastic disturbance, the soils were still able to be classified. Therefore, there is a great potential of establishing a permanent training set or database to determine soil provenance. This study attempted to address some of the most obvious uncertainties of soil provenance applications, marker selection and temporal variability. However, more data and tests, especially in forensically relevant settings, are required to offer

reliable support for forensic investigators. Moreover, further temporal studies are needed to determine the maximum amount of time lapse that can occur between collections for it to be a viable database for searching as well as further studies examining the possible limitations are needed.

V. References

[1] J.M. Young, L.S. Weyrich, J. Breen, L.M. Macdonald, A. Cooper, Predicting the origin of soil evidence: High throughput eukaryote sequencing and MIR spectroscopy applied to a crime scene scenario, Forensic Science International. 251 (2015) 22-31.

[2] A. Gunn, S.J. Pitt, Review Paper Microbes as forensic indicators, Tropical Biomedicine. 29 (2012) 311-330.

[3] R.M. Morgan, P.A. Bull, Forensic geoscience and crime detection, Minerva Medicolegale. 127 (2007) 73-89.

[4] A. Ruffell, Forensic pedology, forensic geology, forensic geoscience, geoforensics and soil forensics, Forensic Science International. 202 (2010) 9-12.

[5] K. Pye, Geological and soil evidence: Forensic Applications, CRC press 2007.

[6] J.M. Young, L.S. Weyrich, A. Cooper, Forensic soil DNA analysis using high-throughput sequencing: a comparison of four molecular markers, Forensic Science International: Genetics. 13 (2014) 176-184.

[7] G.F. Sensabaugh, Microbial community profiling for the characterisation of soil evidence: forensic considerations, Criminal and Environmental Soil Forensics, Springer, 2009, pp. 49-60.

[8] L.I. Moreno, D.K. Mills, J. Entry, R.T. Sautter, K. Mathee, Microbial metagenome profiling using amplicon length heterogeneity-polymerase chain reaction proves more effective than elemental analysis in discriminating soil specimens, Journal of Forensic Science. 51 (2006) 1315-1322.

[9] S.A. Larson, Developing a high throughput protocol for using soil molecular biology as trace evidence, (2012). Theses and Dissertations in Biochemistry. Paper 9.

[10] J. Horswell, S.J. Cordiner, E.W. Maas, T.M. Martin, K.B.W. Sutherland, T.W. Speir, zB. Nogales, A.M. Osborn, Forensic comparison of soils by bacterial community DNA profiling, Journal of Forensic Science. 47 (2002) 350-353.

[11] C.A. Macdonald, R. Ang, S.J. Cordiner, J. Horswell, Discrimination of soils at regional and local levels using bacterial and fungal T-RFLP profiling, Journal of Forensic Science. 56 (2011) 61-69.

[12] L.M. Macdonald, B.K. Singh, N. Thomas, M.J. Brewer, C.D. Campbell, L.A. Dawson, Microbial DNA profiling by multiplex terminal restriction fragment length polymorphism for forensic comparison of soil and the influence of sample condition, Journal of Applied Microbiology. 105 (2008) 813-821.

[13] L.E. Heath, V.A. Saunders, Assessing the potential of bacterial DNA profiling for forensic soil comparisons, Journal of Forensic Science. 51 (2006) 1062-1068.

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

[15] G.A. Evdokimova, N.P. Mozgova, The effect of drying of soil samples on the number of bacteria and fungi, Eurasian Soil Science. 36 (2003) 546-549.

[16] M.S. Meyers, D.R. Foran, Spatial and temporal influences on bacterial profiling of forensic soil samples, Journal of Forensic Science. 53 (2008) 652-660.

[17] N. Fierer, A.S. Grandy, J. Six, E.A. Paul, Searching for unifying principles in soil ecology, Soil Biology and Biochemistry. 41 (2009) 2249-2256.

[18] E.J. Lenz, D.R. Foran, Bacterial profiling of soil using genus-specific markers and multidimensional scaling, Journal of Forensic Science. 55 (2010) 1437-1442.

[19] C.V. Noble, R.W. Drew, J.D. Slabaugh, Soil survey of Dade County area, Florida, USDA NRCS, Gainesville, FL. (1996).

[20] M. Suzuki, M.S. Rappe, S.J. Giovannoni, Kinetic bias in estimates of coastal picoplankton community structure obtained by measurements of small-subunit rRNA gene PCR amplicon length heterogeneity, Applied Environmental Microbiology. 64 (1998) 4522-4529.

[21] E.F. DeLong, Archaea in coastal marine environments, Proceedings of the National Academy of Sciences. 89 (1992) 5685-5689.

[22] L. Cocolin, M. Manzano, C. Cantoni, G. Comi, Denaturing gradient gel electrophoresis analysis of the 16S rRNA gene V1 region to monitor dynamic changes in the bacterial population during fermentation of Italian sausages, Applied Environmental Microbiology. 67 (2001) 5113-5121.

[23] T.J. White, T. Bruns, S. Lee, J. Taylor, Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics, PCR protocols: a guide to methods and applications. 18 (1990) 315-322.

[24] P. Taberlet, L. Gielly, G. Pautou, J. Bouvet, Universal primers for amplification of three non-coding regions of chloroplast DNA, Plant Molecular Biology. 17 (1991) 1105-1109.

[25] E. Afgan, D. Baker, M. van den Beek, D. Blankenberg, D. Bouvier, M. Cech,J. Chilton, D. Clements, N. Coraor, C. Eberhard, B. Gruning, A. Guerler, J. Hillman-Jackson, G.V. Kuster, E. Rasche, N. Soranzo, N. Turaga, J. Taylor, A. Nekrutenko, J. Goecks, The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update, Nucleic Acids Research. 44 (2016) W3-W10.

[26] A. Ramette, Multivariate analyses in microbial ecology, FEMS Microbiology Ecology. 62 (2007) 142-160.

[27] A. Liaw, M. Wiener, Classification and regression by randomForest, R News. 2 (2002) 18-22.

[28] P. Taberlet, E. Coissac, F. Pompanon, L. Gielly, C. Miquel, A. Valentini,T. Vermat, G. Corthier, C. Brochmann, E. Willerslev, Power and limitations of the chloroplast *trnL* (UAA) intron for plant DNA barcoding, Nucleic Acids Research. 35 (2007) e14.

[29] K. Smalla, M. Oros-Sichler, A. Milling, H. Heuer, S. Baumgarte, R. Becker, G. Neuber, S. Kropf, A. Ulrich, C.C. Tebbe, Bacterial diversity of soils assessed by DGGE, T-RFLP and SSCP fingerprints of PCR-amplified 16S rRNA gene fragments: Do the different methods provide similar results? Journal of Microbiological Methods. 69 (2007) 470-479.

[30] F.C.A. Quaak, I. Kuiper, Statistical data analysis of bacterial t-RFLP profiles in forensic soil comparisons, Forensic Science International. 210 (2011) 96-101.

[31] C.L. Schoch, K.A. Seifert, S. Huhndorf, V. Robert, J.L. Spouge, C.A. Levesque, W. Chen, Fungal Barcoding Consortium, Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi, Proceedings of the National Academy of Sciences. 109 (2012) 6241-6246.

[32] Z. Pasternak, A. Al-Ashhab, J. Gatica, R. Gafny, S. Avraham, S. Frenk, D. Minz, O. Gillor, E. Jurkevitch., Optimization of molecular methods and statistical procedures for forensic fingerprinting of microbial soil communities, International Research Journal of Microbiology. 3 (2012) 363-372.

[33] M.J. Johnson, K.Y. Lee, K.M. Scow, DNA fingerprinting reveals links among agricultural crops, soil properties, and the composition of soil microbial communities, Geoderma. 114 (2003) 279-303.

[34] S. Kasel, L.T. Bennett, J. Tibbits, Land use influences soil fungal community composition across central Victoria, South-eastern Australia, Soil Biology and Biochemistry. 40 (2008) 1724-1732.

[35] R.D. Bardgett, R.D. Lovell, P.J. Hobbs, S.C. Jarvis, Seasonal changes in soil microbial communities along a fertility gradient of temperate grasslands, Soil Biology and Biochemistry. 31 (1999) 1021-1030.

[36] A. Shade, J.G. Caporaso, J. Handelsman, R. Knight, N. Fierer, A meta-analysis of changes in bacterial and archaeal communities with time, The ISME Journal. 7 (2013) 1493-1506.

[37] C.L. Lauber, K.S. Ramirez, Z. Aanderud, J. Lennon, N. Fierer, Temporal variability in soil microbial communities across land-use types, The ISME Journal. 7 (2013) 1641-1650.

[38] E.B. Alexander, Soils in Natural Landscapes, CRC Press 2013.

[39] B.K. Singh, R.D. Bardgett, P. Smith, D.S. Reay, Microorganisms and climate change: terrestrial feedbacks and mitigation options, Nature Reviews Microbiology. 8 (2010) 779-790.

[40] M.C. Pereira e Silva, A.C.F. Dias, J.D. van Elsas, J.F. Salles, Spatial and temporal variation of archaeal, bacterial and fungal communities in agricultural soils, PLoS One. 7 (2012) e51554.

Chapter 4: Analysis of the microbial functional diversity using iron genes across different soil types in Miami-Dade County, FL

This chapter has been formatted for: Forensic Science International.

Natalie Damaso[1,2], Yu Cheung[2], Priyanka Kushwaha[3], and DeEtta Mills[1,2*]

[1]Department of Biological Sciences, Florida International University, Miami, Florida, United States of America

[2]International Forensic Research Institute, Florida International University, Miami, Florida, United States of America

[3]Department of Chemistry and Biochemistry, Florida International University, CP 304, 11200 SW 8th Street Miami, FL 33199, USA.

I. Introduction

Over the past decade, a shift in research has been observed to study the functional diversity of an ecosystem versus the taxonomic diversity. Biodiversity is usually defined as the species abundance or richness in an environment. However, the Millennium Ecosystem Assessment group (http://www.millenniumassessment.org/en/index.html) termed biodiversity as the genetic diversity, number (abundance) of species, and functional traits present in an ecosystem [1]. Under global threats such as climate change (drought, flooded, etc.), major alterations of ecosystems are predicted, which can lead to substantial microbial community compositional changes affecting the ecosystem functioning and biogeochemical cycles. Biodiversity has shown to influence ecosystem stability and resilience toward stress and disturbance. However, the relationship between the biotic diversity and microbial guild function in soil is understudied [2]. Ecological equivalence and functional dissimilarity are two contrasting hypotheses that have been the subject of debate. Ecological equivalence hypothesis assumes that under similar environments the microbial communities will display functional redundancy. In contrast, functional dissimilarity assumes that the community functions are dissimilar and not attributed to the environmental conditions but rather linked to the diversity of the microbes present in the system [3]. Ecological equivalence hypothesis has also been related to the biological insurance hypothesis, which states that redundancy within functional groups due to increase diversity will result in overall ecosystem performance and stability [4]. This hypothesis assumes (a) that microbial communities under similar environments are more functionally similar across space and time; and (b) that highly diverse systems support a healthy ecosystem because many taxonomically unrelated

organisms have intrinsic functional redundancy that buffer ecosystem services when environmental stress is experienced [5].

Therefore, more studies should be conducted to understand the regulating forces behind specific functional guilds to determine if soil type drives function or if other environmental factors (e.g., moisture) structure their biogeographical patterns. Correlating the abiotic factors of soil can help understand ecological factors that could regulate the soil biota and their functional guilds. In the current study, the functional gene diversity, specifically the genes related to iron cycling, were queried. Iron is an essential element in organisms and is important in cellular metabolism, respiration, photosynthesis, and other processes. Even though iron is the fourth most abundant element on earth, this transitional element is not readily available for biotic assimilatory or dissimilatory uptake in many environments [6]. Under anaerobic and neutral pH conditions, soluble ferrous iron (Fe^{2+}) is easily accessible and can be taken up by organisms. However, under aerobic and acidic conditions, (Fe^{2+}) is rapidly converted to ferric iron (Fe^{3+}) leading to reduced levels of bioavailable iron for microorganisms. Moreover, intracellular ferrous iron has to be strictly regulated as large quantity can result in cellular toxicity. Microorganisms play a vital role in regulating the transformation and uptake of bioavailable iron under aerobic and anaerobic conditions. Many have also evolved to have the ability to use this terminal electron acceptor (Fe^{3+}) in respiration when oxygen is absent [7,8].

Using functional gene markers could be valuable in forensics to discriminate soils. In this study, the relationship between the abiotic conditions and the functional guilds related to the iron cycle was observed to determine if soil type influenced function.

One of the discriminatory iron genes (*feoB*) detected on the microarray platform, GeoChip, was used to design novel degenerate primers and ultimately make functional diversity profiles to determine if it adds to the discrimination for soil provenance. This approach can potentially reduce the complexity of assaying all bacteria that lead to high level of variability within and among habitats by profiling specific functional markers to discriminate the soils.

II. Materials and Methods

A. Soil Collection

Soil samples (N=168) were collected from five sites across Miami-Dade County, Florida during the wet season in 2014. Sites included: FIU categorized by USDA-NRCS [9] as soil type 1 (Urban Land-Udorthents), CC6 and KNT from soil type 2 (Lauderhill Dania-Pahokee), KK from soil type 3 (Rock Outcrop-Biscayne-Chekika), and CS from soil type 4 (Perrine-Biscayne-Pennsuco). Sites were established in undisturbed areas that had limited public access in the four different soil types. Each site was at least $\geq$ 1.6 km distant from the next, 100 m in length and six subplots within each site were randomly sampled. GPS coordinates for every subplot were recorded. Within each subplot, six cored samples were taken within a 1-m^2 quadrat using a 5 cm diameter soil corer to collect the top 5-10 cm of the soil. The soil samples were transported back to the laboratory and sieved to remove large objects and debris.

B. Abiotic Analysis

Soil texture (% sand, silt, and clay), pH, percent moisture, and percent total organic content (TOC), was obtained for each site. Soil texture was obtained using Bouyocos hydrometer method. The pH was measured in soil-water (1:2) solutions. Soil

slurries were made by adding 3 g of soil to 6 ml of distilled water, stirred, and measured using the calibrated electrode/digital pH meter (LaMotte, Chestertown, MD). The moisture content was determined gravimetrically by oven drying the soils at ~55°C for 24 hours. The TOC was determined gravimetrically by measuring the difference between the dry weight and the ash-free dry weight that was obtained by igniting the dry soils in an ash oven at ~550°C for 4-5 hours. The percent carbon is 50% of the ash free dry mass for plant matter. Lamotte Model STH-14 Outfit (Code 5010-01) (LaMotte, Chestertown, MD) was used to analyze the soil for ferric iron. The methods involve addition of potassium thiocyanate that reacts with iron to give the colored ferric thiocyanate.

C. GeoChip 5.0 Preparation/Analysis

Extraction was conducted using the BIO 101 Fast DNA Spin Kit for Soil® and FastPrep®-24 System homogenizer (MP Bio, Solon, OH). All samples (N=18) per site were pooled and then precipitated with 100% ethanol and 0.3 M sodium acetate. DNA quantity and purity (A260/280~1.8 and A260/230 >1.7) were assessed using UV absorbance. DNA samples were then dried down using the vacufuge before shipping to Institute of Environmental Genomics (IEG) at University of Oklahoma (Norman, OK) for analysis using the GeoChip 5.0. The data were obtained as normalized signal intensity depicting all positive probes detected in each sample and were queried for the iron-related genes. One-way analysis of variance (ANOVA) and Tukey's multiple comparison tests were used to determine iron discriminative genes to distinguish between soil types.

D. Primer Design

Degenerate primers were designed for the *feoB* gene that showed a significant

difference between sites based on the GeoChip analysis. GeoChip data was queried to obtain the GenBank sequences from the *feoB* gene (most discriminative gene) and aligned using Jalview [10] to visualize multiple sequence alignments and show the consensus sequence. Highly conserved areas were targeted to design optimal primers using Primer-BLAST [11]. Novel degenerate primers are listed in Table 11.

Table 11. Degenerate primers designed to amplify the *feoB* gene fragments.

Primer	Sequence (5'-3')	Location
feoB_157F	CCG AAC DBS GGC AAG A	157-172
feoB_555R	CCD BGT CSA NCA TGT TCA	564-581

E. Length Heterogeneity-Polymerase Chain Reaction

DNA was amplified using novel *feoB* primers designed based on the GeoChip results. PCR reaction mixtures were: 1X reaction buffer, 2.5 mM $MgCl_2$, 0.25 mM dNTPs (Promega, Madison, WI), 0.1% BSA (Fraction V, Fisher Scientific, Pittsburgh, PA), 0.1% DMSO (Promega, Madison, WI), 0.6 µM of primers, 40 ng DNA, 0.5 U AmpliTaq Gold® DNA Polymerase (Applied Biosystems, Foster City, CA). Reverse primer was labeled with 6-FAM fluorescent dye. The Bio-Rad C1000 Touch™ thermocycler (Bio-Rad, Hercules, CA) with the following parameters: initial 10 min denaturing step at 95°C, 35 cycles of denaturation at 95°C, annealing at 56°C, and extension at 72°C each for 30 sec with a final extension at 72°C for 10 min.

F. Capillary Electrophoresis

Fragment analysis was conducted using the ABI Prism™ 3130xl (Applied Biosystems, Foster City, CA) using Performance Optimized Polymer 7 (POP7) (Applied Biosystems, Foster City, CA). Samples were prepared adding 2 µl of each PCR product to a mixture of 11.5 µl Hi-Di™ Formamide (Applied Biosystems, Foster City, CA) and

0.65 µl internal size standard, GeneScan LIZ™ 600 (Applied Biosystems, Foster City, CA), denatured by heating for 2 min at 95°C and then snap-cooled on ice for 2 min.

Raw data were analyzed using the GeneMapper™ v 4.0 (Applied Biosystems, Foster City, CA). Local Southern size calling was used for the analysis parameters with a minimum threshold of 50 relative fluorescent units (RFUs). The relative ratios were calculated by normalizing the heights of each peak in the profile to the total peak intensities resulting in the ratio for each peak height as a decimal value from zero to one using the Galaxy ABI 310 Data Formatting tool found in http://usegalaxy.org/ [12].

G. Statistical Analysis

All analyses were conducted using Primer-E v.7 software (PRIMER E Ltd., Plymouth Marine Laboratory, Plymouth, U.K.). Bray-Curtis similarity matrices were generated on relative abundance ratios that had been square root transformed prior to analysis. Analysis of Similarity (ANOSIM) was used to determine the significant differences between sites. ANOSIM reports the level of dissimilarity between samples groups (Global R) and the associated level of significance (p) to provide statistical pair-wise comparisons between designated groups. The ANOSIM R-statistic indicates the level of discrimination between groups (site), with a value close to one indicating complete group discrimination and a value close to zero implying no differences exist between groups. The associated significance level (p) is equivalent to the p-value where 0.1%, 1%, and 5% is equal to $p<0.001$, $p<0.01$, and $p<0.05$, respectively. Non-metric Multidimensional Scaling (nMDS) was used to visualize the site heterogeneity and discrimination power of the iron primers to distinguish sites apart. nMDS allows complex datasets to be easily visualized, with more similar samples grouping together.

The level of confidence in the 2-D plot is indicated by the stress, i.e. <0.2 provides good representation of the fit [13]. Similarity Percentages (SIMPER) analysis was used to identify the LH-PCR peaks contributing to the dissimilarity between sites as well as the percent contribution each amplicon provided for the overall dissimilarities. Canonical correspondence analyses (CCA) was performed to determine the effect of the abiotic factors (i.e., moisture, TOC, pH, Fe^{3+}) on the *feoB* functional diversity.

III. Results

A. Abiotic Results

Across the four soil types, samples from soil types 1 and 3 were similar in abiotic content while, soil types 2 and 4 were similar (Table 12). Soil type 1 and 3 had less moisture and organic content compared to soil type 2 and 4. FIU and KK from soil type 1 and 3 respectively were both sandy loam texture, less than 30% moisture, and had 13% organic content. While KNT and CS from soil type 2 and 4, respectively, were both loam textures had over 70% moisture, and over 25% organic content. CC6 from soil type 2 also had high moisture (48%) and organic content (27%) similar to KNT from soil type 2 however, it was sandy clay loam. All sites had similar ferric iron concentration (<2.5 ppm) except for KNT that had 6.5 ppm.

Table 12. Soil texture, moisture percent, organic content percent, pH, and ferric iron concentration for each site (FIU, CC6, KNT, KK, CS). Soil samples are identified by a soil type number followed by a transect descriptor (e.g., 1-FIU corresponds to soil type 1, transect FIU). Soil texture classification based on the % sand, silt, and clay for each site. Parenthesis represent standard error.

	Soil Texture	**Moisture (%)**	**Organic (%)**	**pH**	**Fe^{3+} (ppm)**
1-FIU	Sandy Loam	23.67 (± 4.38)	12.34 (± 4.80)	7.47 (± 0.12)	<2.5 (± 0.00)
2-CC6	Sandy Clay Loam	48.02 (± 6.46)	26.80 (± 6.31)	7.64 (± 0.10)	<2.5 (± 0.00)
2-KNT	Loam	74.61 (± 4.43)	24.40 (± 2.66)	7.43 (± 0.03)	6.5 (± 1.00)
3-KK	Sandy Loam	29.53 (± 1.78)	13.38 (± 0.82)	7.91 (± 0.03)	<2.5 (± 0.00)

4-CS	Loam	79.55 (± 2.51)	41.04 (± 4.47)	7.44 (± 0.05)	<2.5 (± 0.00)

B. GeoChip Results

GeoChip results revealed forty-seven associated iron cycle genes in the soil types queried across Miami-Dade County (Figure 11). Two were involved with iron storage, 16 for iron transport and 29 for iron uptake (Figure 11). Archaea were involved in iron storage and transport, while fungi were involved exclusively for iron transport and uptake (Figure 11). Bacteria had the most iron cycle associated genes relative to Archaea and fungi on the microarray and the genes were associated with all three functions (storage, transport, and uptake) (Figure 11). Eight genes were unique to the soil's moisture content: *ccm, frgA, fyuA, hasA, hhu, ira, pch, pyoverdin_pvcC* were exclusive to KNT and CS sites that had >75% moisture, while *ira* was present in FIU and KK sites that had <30% moisture. Two genes were exclusive to a soil type: *mbtD* exclusive to KK site and *psn* exclusive to CS site. Three genes were found to be significantly different ($p<0.05$) across soil types based on ANOVA and Tukey HSD test: *dps, cirA, feoB*. Dps is one of the three iron storage proteins that are used to enhance growth when external iron supplies are limited. Unlike the other storage proteins, Dps has a lower storage capacity and can be involved in protecting DNA from anti-redox agents. The other two genes (*feoB* and *cirA*) are involved with iron transport. The *feoB* produces a ferrous iron transporter that is highly conserved in many bacteria. This transporter is important during low oxygen conditions when ferrous is stable and dominates over ferric iron. In contrast, the *cirA* gene is an important transporter when iron is limited and interacts with *fur* (ferric uptake regulator). Under iron limitation, regulators such as *fur*, uptake ferric iron and convert it to the bioavailable form [14]. One of the discriminatory iron genes

detected in GeoChip (*feoB*) was used to make functional diversity profiles to determine if it adds to the discrimination for soil provenance.

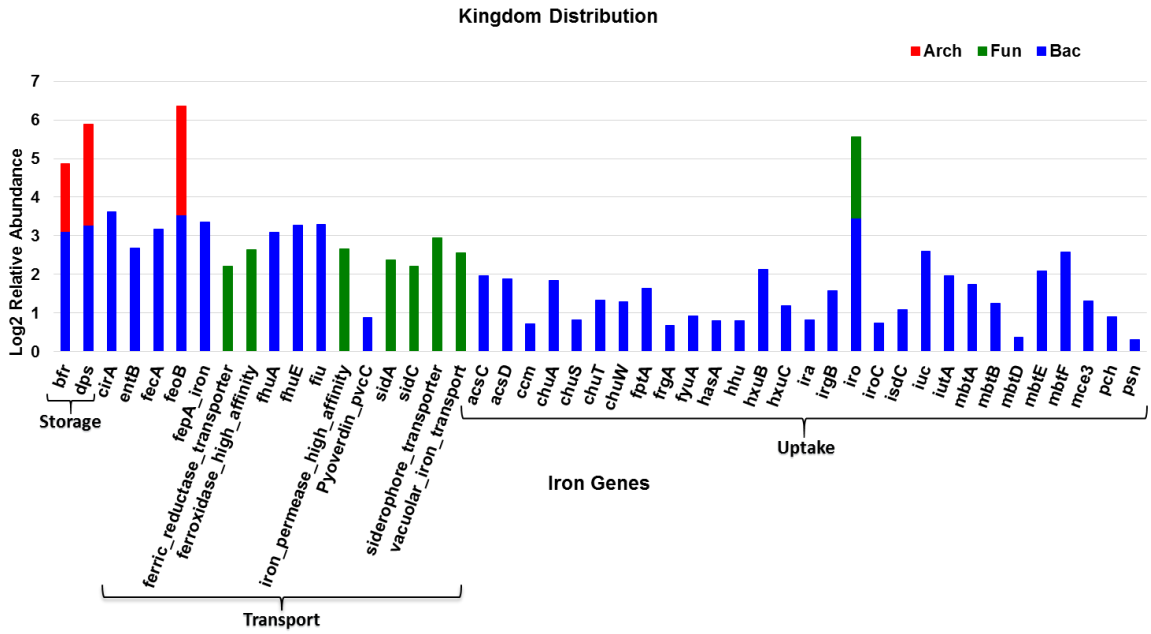

Figure 11. GeoChip results illustrating the Kingdom distribution across the four soil types for each iron gene. Out of the 47 iron genes, two were involved with iron storage, 16 for iron transport, and 29 for iron uptake. Archaea (Arch) was involved in iron storage and transport, while fungi (Fun) was involved exclusively for iron transport and uptake. Bacteria (Bac) had the most iron genes and were involved with all three functions, storage, transport and uptake.

C. Soil discrimination using *feoB* degenerate primers

The *feoB* degenerate primers were able to produce profiles for all the sites except for KNT. Further troubleshooting and optimization is ongoing for the KNT transect. Therefore, transect KNT was not included in further statistical analysis. Similarity percentages (SIMPER) were used to show the variability (dissimilarity) of the functional diversity among and within the sites (Table 13). The "Among" column compares the average dissimilarity of one site (i.e., FIU) when compared to the other three sites (CC6, KK, CS), while the "Within" column compares the average dissimilarities of the six

subplots within the site. Overall, the average dissimilarity among sites was greater than within sites. Moreover, the within site average dissimilarity varied compared to site. For example, KK had the lowest within site dissimilarity of 50%, whereas CC6 had the largest within site dissimilarity of 97%. This can be attributed to the site soil and above ground biomass heterogeneity (Table 12). This was also visualized in the nMDS plot where KK grouped together, while CC6 and FIU subplots did not all group together (Figure 12). ANOSIM statistic also supported nMDS results illustrating that the profiles were significantly influenced by site ($R=0.44$, $p=0.1\%$). However, some sites were not significantly different such as FIU and CC6. This was also shown by nMDS in where some subplots grouped together. For instance, FIU subplot 3 & 4 grouped with CC6 subplot 1, 3-5 (data not shown). Based on canonical correlation analysis (CCA), no significant relationship ($p>0.05$) between the abiotic parameters measured (i.e., moisture, TOC, pH, Fe^{3+}) and the microorganism's *feoB* functional diversity profiles was shown (data not shown).

Table 13. SIMPER analysis illustrating the average dissimilarity between and within each site (± is the SE of the mean % dissimilarity).

	Dissimilarity (%)	
Transect	**Among**	**Within**
FIU	93.61 (±3.27)	86.00 (±8.67)
CC6	96.90 (±1.14)	96.86 (±12.81)
KK	94.16 (±3.38)	50.24 (±4.83)
CS	96.48 (±1.18)	71.16 (±10.49)

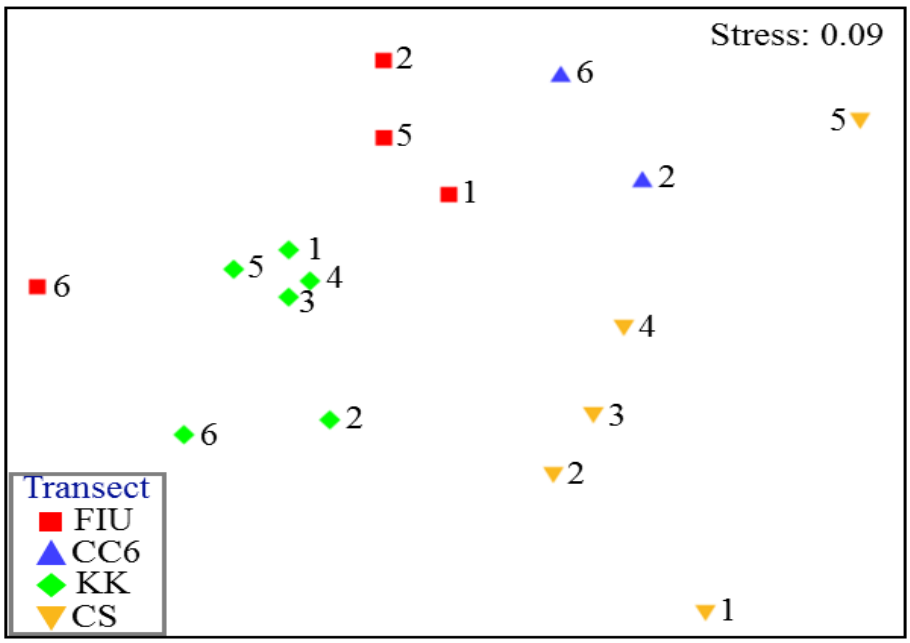

Figure 12. Non-Metric Multidimensional Scaling 2-D plot illustrating the discrimination power to distinguish five sites (Red=FIU, Blue=CC6, Green=KK, Yellow=CS) using novel degenerate *feoB* primers. Numbers represent the subplots for each site.

IV. Discussion

Ecological equivalence and functional dissimilarity are two contrasting hypotheses that have been the subject of debate over the past decade. Most previous research has supported the ecological functional redundancy hypothesis [15]; however, others such as Strickland et al. (2009) found that soil microbial communities in the same environment are not functionally equivalent as the rates of carbon dioxide production from the litter decomposition were dependent on the microbial inoculum [3]. Studies by Desai et al. (2012) showed that in phototrophic organisms, there was a clear influence of the ecological niche on the diversity of Fe uptake systems and that Fe uptake and homeostasis mechanisms differed significantly across marine niches defined by temperature and bioavailable Fe concentrations. This was linked to the distribution of microbial taxa in these niches [16]. In the present study, eight genes (*ccm, frgA, fyuA,*

hasA, hhu, ira, pch, pyoverdin_pvcC) were unique to some sites and were related to the environmental conditions, specifically the moisture content. These genes were found in sites that had a >75% moisture, except for *ira* that was also found in drier soils with <30% moisture. This supports the ecological equivalence hypothesis that assumes that the microbial community under similar environments will function more similarly. This can also be useful in a forensic setting to assist in determining origin of the soil sample.

Soils can be a great source of evidence to assist in criminal investigations as they are highly individualistic, have high probability of transfer and retention, and nearly invisible so suspect will often overlook the evidence [17]. Several forensic soil studies have shown the potential of using microbial profiles to distinguish soils and also to identify origin of the soil sample; however, assaying all bacteria can lead to complex and noisy data due to their high level of variability within and among habitats [18-20]. Previously, Lenz & Foran (2010) profiled *recA* gene specific to nitrogen fixing bacteria rhizobia that lead to less complex TRFLP profiles than bacterial 16S rRNA gene [20]. However, Angermeyer et al. (2015), showed that distance-decay which is commonly observed in structural genes (i.e., 16S rRNA) was not observed in a sulfate-reduction gene (*dsrA*) [21]. In this study, some iron related genes were unique to an environment (i.e., eight genes unique to moisture content). Moreover, three genes (*dps, cirA, feoB*) were found to be significantly different based on geographic location. One of the discriminatory iron genes detected in GeoChip (*feoB*) was used to make functional diversity profiles to determine if it adds to the discrimination for soil provenance. Currently, there is *feoB* primers have been species specific. In this study, novel degenerate 'universal' primers were designed to target a vast array of species.

The *feoB* is a ferrous iron transporter that is highly conserved in many bacteria. This transporter is important during low oxygen conditions when ferrous is stable and dominates over ferric iron. Under anoxic conditions, Fe^{2+} is stable and more soluble than Fe^{3+}, this allows the transport without complexation by ligands [22]. Under anaerobic-microerophilic conditions, bacteria use the FeoB pathway to mediate the transport of free Fe^{2+} across the inner membrane to the cytoplasm in a GTP-dependent manner [23]. FeoAB consists of cytosolic protein and inner membrane transporter for uptake of ferrous iron [24]. Ferrous iron transporter is encoded by anaerobically induced, and iron repressed *feoAB* genes that are highly conserved in many bacteria [14]. FeoB systems have no significant difference between Gram positive and Gram negative bacteria; however, high GC Gram positive bacteria generally do not possess ferrous iron uptake systems [25].

This study has shown that using functional markers can be used in forensics to discriminate soils and have the potential to reduce the complexity of assaying all bacteria. Functional diversity profiles using novel *feoB* primers did show both among and within sites variability. As commonly observed with structural diversity (i.e., 16S rRNA) the functional diversity among sites was greater than the within site variability. Moreover, as commonly seen the within site variability differed based on more homogeneous sites (i.e., KK) having lower within site dissimilarity than more heterogeneous sites (i.e., FIU). However, not all sites were distinguishable. Some sites within FIU and CC6 were indistinguishable. In this study, no significant relationship between the abiotic parameters measured (i.e., moisture, TOC, pH, Fe^{3+}) and the microorganism's *feoB* functional diversity profiles was observed. Further research is needed to determine the

factors regulating the microbial organisms' functional diversity and their biogeographical patterns. Moreover, not all sites produced a profile (i.e., KNT). Optimization and troubleshooting is ongoing to determine the sensitivity and effectiveness of the degenerate *feoB* primers for all sites. This study, showed a novel method to query the iron relating genes and ability to design a novel marker that can easily be used to profile the functional diversity of a soil community.

V. References

[1] S. Naeem, Biodiversity, ecosystem functioning, and human wellbeing: an ecological and economic perspective, Oxford University Press, Oxford; New York, 2009.

[2] V. Torsvik, L. Øvreås, Microbial diversity and function in soil: from genes to ecosystems, Current Opinion in Microbiology. 5 (2002) 240-245.

[3] M.S. Strickland, C. Lauber, N. Fierer, M.A. Bradford, Testing the functional significance of microbial community composition, Ecology. 90 (2009) 441-451.

[4] D.C. Coleman, W.B. Whitman, Linking species richness, biodiversity and ecosystem function in soil systems, Pedobiologia. 49 (2005) 479-497.

[5] S.D. Allison, J.B. Martiny, Colloquium paper: resistance, resilience, and redundancy in microbial communities, Proceedings of the National Academy of Science. 105 Suppl 1 (2008) 11512-11519.

[6] K.A. Weber, L.A. Achenbach, J.D. Coates, Microorganisms pumping iron: anaerobic microbial iron oxidation and reduction, Nature Reviews Microbiology. 4 (2006) 752-764.

[7] C. Colombo, G. Palumbo, J. He, R. Pinton, S. Cesco, Review on iron availability in soil: interaction of Fe minerals, plants, and microbes, Journal of Soils and Sediments. 14 (2014) 538-548.

[8] L. Escolar, J. Perez-Martin, V. de Lorenzo, Opening the iron box: transcriptional metalloregulation by the Fur protein, Journal of Bacteriology. 181 (1999) 6223-6229.

[9] C.V. Noble, R.W. Drew, J.D. Slabaugh, Soil survey of Dade County area, Florida, USDA NRCS, Gainesville, FL. (1996).

[10] A.M. Waterhouse, J.B. Procter, D.M. Martin, M. Clamp, G.J. Barton, Jalview Version 2--a multiple sequence alignment editor and analysis workbench, Bioinformatics. 25 (2009) 1189-1191.

[11] J. Ye, G. Coulouris, I. Zaretskaya, I. Cutcutache, S. Rozen, T.L. Madden, Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction, BMC Bioinformatics. 13 (2012) 1.

[12] E. Afgan, D. Baker, M. van den Beek, D. Blankenberg, D. Bouvier, M. Cech, J. Chilton, D. Clements, N. Coraor, C. Eberhard, B. Gruning, A.. Guerler, J. Hillman-Jackson, G. Von Kuster, E. Rasche, N. Soranzo, N. Turaga, J. Taylor, A. Nekrutenko, J. Goecks, The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update, Nucleic Acids Research. 44 (2016) W3-W10.

[13] J.M. Young, L.S. Weyrich, J. Breen, L.M. Macdonald, A. Cooper, Predicting the origin of soil evidence: High throughput eukaryote sequencing and MIR spectroscopy applied to a crime scene scenario, Forensic Science International. 251 (2015) 22-31.

[14] S.C. Andrews, A.K. Robinson, F. Rodriguez-Quinones, Bacterial iron homeostasis, FEMS Microbiological Reviews. 27 (2003) 215-237.

[15] P. Maron, C. Mougel, L. Ranjard, Soil microbial diversity: methodological strategy, spatial overview and functional interest, Comptes Rendus Biologies. 334 (2011) 403-411.

[16] D.K. Desai, F.D. Desai, J. LaRoche, Factors influencing the diversity of iron uptake systems in aquatic microorganisms, Environmental Bioinorganic Chemistry of Aquatic Microbial Organisms. (2012) 103.

[17] R. Fitzpatrick, S. In, J. Siegel, P. Saukko, Soils, (2013).

[18] M.S. Meyers, D.R. Foran, Spatial and temporal influences on bacterial profiling of forensic soil samples, Journal of Forensic Science. 53 (2008) 652-660.

[19] J.M. Young, L.S. Weyrich, A. Cooper, Forensic soil DNA analysis using high-throughput sequencing: a comparison of four molecular markers, Forensic Science International: Genetics. 13 (2014) 176-184.

[20] E.J. Lenz, D.R. Foran, Bacterial profiling of soil using genus-specific markers and multidimensional scaling, Journal of Forensic Science. 55 (2010) 1437-1442.

[21] A. Angermeyer, S.C. Crosby, J.A. Huber, Decoupled distance–decay patterns between *dsrA* and 16S rRNA genes among salt marsh sulfate-reducing bacteria, Environmental Microbiology. (2015).

[22] R. Crichton, Intracellular Iron Metabolism and Cellular Iron Homeostasis, Iron Metabolism: From Molecular Mechanisms to Clinical Consequences, 3rd Edition. (2001) 223-269.

[23] G. Porcheron, A. Garénaux, J. Proulx, M. Sabri, C. Dozois, Iron, copper, zinc, and manganese transport and regulation in pathogenic Enterobacteria: correlations between strains, site of infection and the relative importance of the different metal transport systems for virulence, Frontiers in Cellular and Infection Microbiology. 3 (2013).

[24] E.R. Frawley, F.C. Fang, The ins and outs of bacterial iron metabolism, Molecular Microbiology. 93 (2014) 609-616.

[25] M. Ballouche, P. Cornelis, C. Baysse, Iron metabolism: a promising target for antibacterial strategies, Recent patents on anti-infective drug discovery. 4 (2009) 190-205.

Conclusion

The current ecological hypothesis states that the soil type (e.g., chemical and physical properties) determines which microbes occupy a particular soil and provides the foundation for soil provenance studies. As human profiles are used to determine a match between evidence from a crime scene and a suspect, a soil microbial profile can be used to determine a match between soil found on the suspect's shoes or clothing to the soil at a crime scene. This research showed the potential and effectiveness of using microbial DNA from soil, not just for comparison, but also for intelligence gathering to pinpoint the geographic origin of the soil.

Microbial profiling and bioinformatics analyses of the soil community provided a rapid method for soil provenance that can be informative, easier to perform, and more cost effective than approached using traditional physico-chemical data. To our knowledge, the work presented here is one of the first studies to use bioinformatic tools for soil forensic application using four-taxa and is unique in its consideration of multiple spatial scales. This present study builds on our growing knowledge of spatial relationships in microbial communities by applying the Mantel statistic to this dataset to illustrate that the biotic patterns and their geographic location are indeed spatially auto-correlated in Miami-Dade soils. Based on the four-taxa microbial profiles and Mantel test, correlation between biotic content and geographic location was observed, thus justifying the use of machine learning tools to predict biotic patterns that can be applied for determination of soil provenance. Five supervised machine-learning algorithms were evaluated for their predictive value when using four-taxa biotic profiles for soil classification.

Currently there is no comprehensive soil microbial community profiling database and very few published attempts to develop databases of soil properties (i.e., chemical and physical) specifically with forensic application in mind. Determining the sampling design- number of samples collected and distance between samples across different habitats- to utilize soil microbial profiling for intelligence based forensic investigations and ultimately establish a usable database for soil provenance are needed. This study showed that Geostatistics can assist in assessing the spatial variability and offer an index to quantify the magnitude and scale of spatial variation in a soil property (i.e., microbial community profiles). The significance of this for forensics links back to the issue of soil variability at the crime scene and how realistic it is to expect a soil sample collected by an investigator to be similar to a questioned sample (Lark & Rawlins, 2008).

Moreover, for a robust tool to be applied in forensic application, an understanding of the uncertainty associated with any comparisons and the parameters that can significantly influence variability in profiles needs to be determined. These issues include selecting suitable microbial markers and the influence of temporal variability on the DNA profile. Previous results have assessed short-term (1-1.5 yr) spatio-temporal variability of soil communities however, if archived data and training sets are to be useful long term, temporal variability (> 2 yr) also needs to be considered. Unlike human identification, soil environment is dynamic and changes over time. Therefore, it is important to see if meaningful comparisons and links can still be made between soil evidence deposited at the crime and archived reference data previously collected (> 2 yr) from a site can still be classified.

Overall this study showed that stable profiles may allow comparison between evidence and a possible crime scene despite the time lapse between sample collections, however, this is dependent on the analysis method, site, vegetation, and level of disturbance. Therefore, temporal variability of the soil microbial communities and how this variability compares among different soil types is important to understand. More data and tests, especially in forensically relevant settings, are required to offer reliable support for forensic investigators. Moreover, further temporal studies are needed to determine the maximum amount of time lapse that can occur between collections for it to be a viable database for searching as well as further studies examining the possible limitations are needed. This is critical to understand how frequently a reference dataset needs to be updated.

Lastly, marker selection is an important consideration for microbial profiling. In this study, we assessed the resolution of bacteria, Archaea, fungal, and plant community profiles independently and combined to determine the best marker or markers for forensic comparison of soil evidence. In conclusion, even though fungi look promising for single taxon soil discrimination, the additional markers can help discriminate between a wide variety of soil types. As in human identification, the more DNA markers queried the greater the discrimination power. In this study, functional diversity was also assessed to determine if soil type drives function and their potential to use functional markers for forensic purposes to discriminate soils. This study, showed a novel method to query the iron relating genes and ability to design a novel marker that can easily be used to profile the functional diversity of a soil community.

VITA

NATALIE DAMASO

	Born, Hialeah, Florida
2006-2011	B.S., Biology Florida International University Miami, Florida
2006-2011	Bright Futures Scholarship Florida Department of Education Tallahassee, Florida
2008-2009	National Student Exchange Program Plymouth State University Plymouth, New Hampshire
2010	McNair Post Baccalaureate Achievement Program Florida International University Miami, Florida
2011-2013	M.S., Forensic Science Florida International University Miami, Florida
2012-2016	Doctoral Candidate Florida International University Miami, Florida
2012-2014	McNair Graduate Fellowship Florida International University Miami, Florida
2013-2014	Grants-in-Aid of Research Program Sigma Xi, The Scientific Research Society Research Triangle Park, North Carolina
2016	Dissertation Year Fellowship Florida International University Miami, Florida
2016	Visiting Scientist/Research Participation Program Oak Ridge Institute for Science and Education Oak Ridge, Tennessee

PUBLICATIONS AND PRESENTATIONS

Martin, L., Damaso, N., & Mills, D. (2016). A comparison for detection of single nucleotide polymorphisms (SNP) in equine coat color genes using SNaPshot Multiplex kit or Pluronic F-108 tri-block copolymer and capillary electrophoresis. *Electrophoresis*.

Damaso, N., Kushwaha, P., Cheung, Y. & Mills, D., 2015 "Analysis of the functional guild diversity across different soil types and their application to soil provenance" Poster presentation at the 2015 Annual Meeting-Florida Branch American Society for Microbiology.

Cheung, Y., Damaso, N., Kushwaha, P., & Mills, D., 2015 "A comparison of 16S rRNA genes and *gyrB* as phylogenetic markers: Two metagenomic approaches" Poster presentation at the 2015 Annual Meeting-Florida Branch American Society for Microbiology.

Damaso, N. and Mills, D., 2015 "Machine learning tools for classification of soils using microbial profiles" Oral presentation at the National Cooperative Soil Survey National Conference-NRCS USDA.

Martin, L., Damaso, N. and Mills, D., 2015 "A methodological comparison to detect horse coat color single nucleotide polymorphisms (SNP) utilizing a novel polymer" Oral presentation at the 8th Annual International Veterinary Forensic Science Association Conference.

Cheung, Y., Kushwaha, P., Damaso, N., & Mills, D., 2015 "GeoChip 5.0: Investigating Microbial Functional Diversity for Soil Provenance" Poster presentation at the 4th Annual Forensic Symposium-Florida International University.

Damaso, N., Cheung, Y. & Mills, D., 2015 "Spatial and Temporal Variability in Soils-Their Importance for Intelligence and Forensic Application" Poster presentation at the 2015 Annual Scientific Meeting-American Academy of Forensic Science.

Damaso, N., Martin, L., Kushwaha, P., & Mills, D. (2014). F-108 polymer and capillary electrophoresis easily resolves complex environmental DNA mixtures and SNPs. *Electrophoresis*, *35*(21-22), 3208-3211.

Damaso, N. Mendel, J., Mendoza, M., and Mills, D., 2014 "Bioinformatics for the classification and provenance of soil samples for intelligence and forensic applications" Oral presentation at the 3rd Annual Forensic Symposium-Florida International University.

Damaso, N., Mendoza, M., and Mills, D., 2013 "Spatial autocorrelation of soil biota profiles with soil type across Miami-Dade County" Oral presentation at the 2013 Annual Meeting-Florida Branch American Society for Microbiology.