Foley, Brianna A., Use of ATCC® MSA-2002™ for Validation of Extraction and Amplification Techniques in 16S Microbial Community Profiling. Master of Science (Biomedical Sciences), May 2018, 47 pp., 5 tables, 9 figures, 34 references.

The implementation of microbiome analysis is an emerging area of focus in forensic research. However, microbiome analysis is not well validated for use in forensic analyses and there is no standard protocol in place. In this research a comprehensive analysis of extraction and amplification techniques employed during investigation of microbial communities was performed using ATCC® MSA-2002™ as a mock microbial community. Comparison of DNA extraction protocols was performed followed by an analysis of commercially-available polymerases. Samples were pooled for sequencing of the V4 region of 16S ribosomal RNA gene using the Illumina MiSeq System, and subsequently analyzed for community composition. The results were compared with the known genomic data of the mock microbial community and statistical methods were employed to determine the extent of deviation.

USE OF ATCC® MSA-2002™ FOR VALIDATION OF EXTRACTION AND AMPLIFICATION TECHNIQUES IN 16S MICROBIAL COMMUNITY PROFILING

THESIS

Presented to the Graduate Council of the

Graduate School of Biomedical Sciences

University of North Texas

Health Science Center at Fort Worth

in Partial Fulfillment of the Requirements

For the Degree of

MASTER OF SCIENCE

By

Brianna A. Foley, B.S.

Fort Worth, Texas

May 2018

ACKNOWLEDGEMENTS

I would like to express my appreciation to my major professor, Dr. Michael Allen, for all of his help and guidance and for providing me with the opportunity to conduct this research. I would also like to thank my committee members: Dr. Joseph Warren, Dr. Yan Zhang, and Dr. Scott Maddux for their support and feedback throughout this process. A special thanks to Viviana Mancilla and Elizabeth Mitchell for making my transition into their lab such a smooth process and for always being there when I had questions. They both went above and beyond when I needed help and I am not sure what I would have done without them. Finally, a big thank you to my Excel guru and graphic designer, Patrick Manion. I thank you for always being there for me and for keeping me motivated and focused during this journey. You have been my rock, and for that I am forever grateful.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION AND BACKGROUND

Background on Microbial Forensics

Microbial forensics has been defined as the discipline of applying scientific methods to the analysis of microbial evidence in criminal and civil cases for investigative purposes (*1*). A microbiome is an aggregate of microorganisms in a particular environment, whether it be from soil, a body of water, or bacteria that resides on or within the human body. By analyzing the microbiome of an environmental or biological specimen that has been left behind at the scene of the crime, a profile can be generated based on the classification of sequencing reads to specific taxa which are unique to the source of the specimen. In recent years, technical advancements through the introduction of next-generation sequencing (NGS), also known as high-throughput sequencing (HTS), have revolutionized the study and application of genomics and molecular biology. The implementation of next-generation sequencing offers the possibility of a new form of trace evidence to be utilized for criminal investigations. Microbiome analysis can be utilized in conjunction with human DNA analysis to improve trace evidence options for forensic investigations by expanding on current genetic testing abilities and linking perpetrators to the crime scene.

The NIH-funded Human Microbiome Project has led to a significant increase in the public and scientific recognition of the importance of microbial communities and their relationships with their human hosts (*2*). The human body contains approximately as many bacterial cells as it does human cells, with types of bacterial colonization varying based on body site (*3*). When an item is touched by an individual, the bacterial community that is transferred from their fingertips could potentially be used to identify the person based on the bacterial residue (*4*). Similarly, bacterial populations found in soil residue can be individualizing by supplying a microbiome profile that is specific to the plot of land from which the soil originated (*5*).

Massively–parallel, "Next-generation" sequencing platforms such as the Illumina MiSeq System (Illumina Inc., San Diego, California) offer a method of sequencing the 16S rRNA gene amplicons from a sample to identify and compare bacteria from complex microbiomes and environments (*6*). The 16S rRNA gene encodes the ribosomal RNA component of the bacterial 30S small ribosomal subunit (*7*). Sequencing of this gene is commonly used for bacterial phylogenetic and taxonomic studies due to its high level of conservation between bacterial species (*7*). Unlike shotgun sequencing where random fragments of genomes are sequenced, 16S rRNA gene sequencing can be specifically targeted against bacteria, can provide very deep coverage of complex communities, and can be used in circumstances where only trace amounts of bacterial DNA are present (*8*). For these reasons, 16S rRNA gene sequencing offers a key advantage for forensic microbiome analysis.

Potential Sources of Error

While the implementation of 16S rRNA gene sequencing for microbiome analysis has clear and notable advantages in the field of forensic science, the use of this methodology has yet to be validated in a controlled laboratory setting for forensic microbial analysis. Specific sources of errors must be addressed such as differential or biased extraction of nucleic acids and errors or bias in PCR amplification (*7*).

Bacteria are highly diverse but are broadly classified as either Gram-positive or Gram-negative based on the structure of their cell wall. The name refers to the gram staining method that was developed by Christian Gram in 1884 (*7*). When sampling a complex bacterial population, these physiological differences in the outer cell walls can cause differential cell lysis within a sample. While Gram-negative cells have a thin lipopolysaccharide cell wall, and are generally structurally weaker, the cell wall of Gram-positive organisms is made up of a thick peptidoglycan outer cell wall that is more resistant to lysis, which leads to differential extraction (*7*). Some Gram-positive bacteria are also capable of forming endospores in response to adverse conditions. Sporulation gives rise to heavy impervious layers of protein which make the bacteria resistant to threats such as high temperature, radiation, desiccation, and enzymatic destruction (*9*).

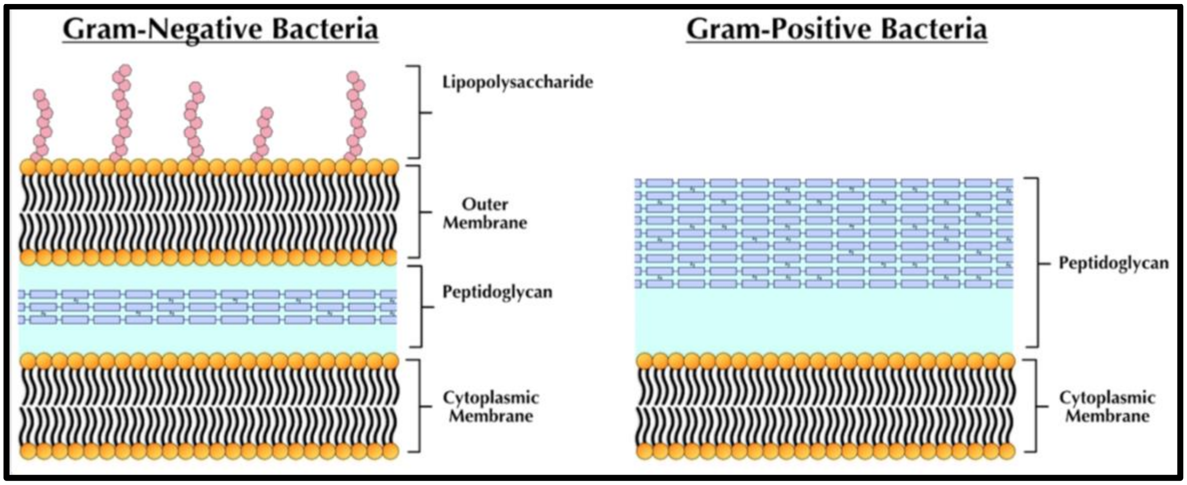

Figure 1. Gram-negative Versus Gram-positive Cell Wall (*10*). Gram-negative bacteria (left) generally possess a thin layer of peptidoglycan located between two outer membranes. Gram-positive bacteria (right) generally have a single outer membrane surrounded by a thick layer of peptidoglycan.

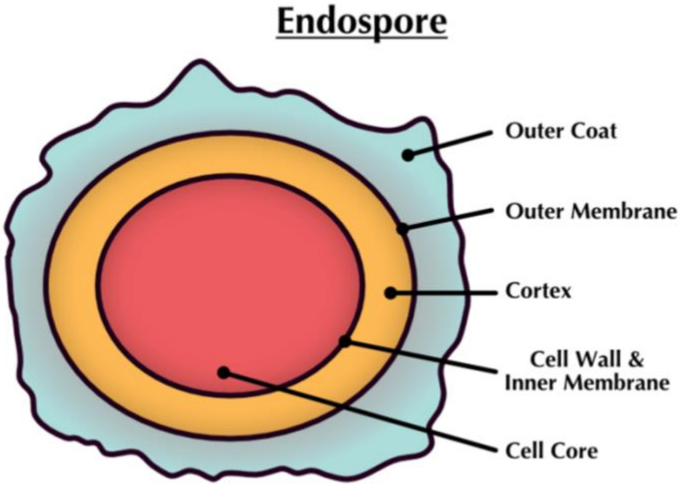

Figure 2. Endospore Structure (*9*).

It has been well documented that variations in amplification efficiencies create inaccuracies and bias in multi-template PCR reactions (*11*). Polymerase base substitution, template-switching, and PCR-mediated recombination are all sources of error that have been observed in PCR products (*12*). Additionally, chimeras are formed when two or more biological sequences are fused together in a PCR reaction. This event has the potential to lead to problems when interpreting assay results such as overestimation of community diversity and the presence of non-existing microorganisms in a sample (*7*).

Research Significance

Validation and standardization are essential in the field of forensics, as law enforcement officials rely on the accuracy and reliability of results obtained from forensic analyses (1*2*). An important consideration to keep in mind is that the results of such analyses have the potential to impact the life and freedom of the individuals involved. Validation guidelines and standards also have an influence on the admissibility of evidence in a court of law. The *Daubert* standard is used by judges in order to make preliminary assessments of whether an item of evidence or methodology will be admissible in a trial hearing. Under this standard, factors are called into question such as whether the theory or technique in question can be tested, whether it has been subjected to peer review, its known or potential error rate, and whether there are standards controlling the technique (*13*).

Validation of microbial profiling for forensic use is important to determine accurate result interpretation and limitations, as well as to establish issues such as varying differences between extraction methods, error rates of polymerases, and sources of technical variability (*1*). To address these issues, this research provides a comprehensive analysis of various techniques using

ATCC® 20 Strain Even Mix Whole Cell Material (ATCC® MSA-2002™) (*14*) as a mock microbial community. The data obtained in this study evaluate how these different techniques affect downstream methods such as next-generation sequencing as well as provide useful suggestions that others can use for microbial analysis in the forensic setting.

CHAPTER II

MATERIALS AND METHODS

Experimental Design

The purpose of this research focused on the need for development of the best standards and practices to be put into place for forensic microbial profiling, as well as the identification of sources of error, including differences between extraction and amplification methods and other sources of technical variability. This was tested through four experimental phases. The first phase was to compare differences in extraction techniques by performing three different extraction methods using ATCC MSA-2002 as a mock microbial community. The second phase involved completing library preparations. This began with analyzing PCR amplification by testing three different polymerases on each of the three extraction samples, followed by PCR clean-up, index PCR, quantification, and normalization. The third phase was to perform next-generation sequencing of the prepared libraries, using the Illumina MiSeq System. In the fourth and final phase, analysis and interpretation of the genomic data was performed in order to determine correlation between the obtained data and the known composition of the mock microbial community. The results of this phase were also used to compare differences between extraction methods and polymerases.

ATCC® MSA-2002™ Mock Microbial Community

For the purposes of this research, ATCC® 20 Strain Even Mix Whole Cell Material (ATCC® MSA-2002™) purchased from the American Type Culture Collection served as the mock microbial community. This product is prepared as a mixture of Gram-positive and Gram-negative whole cells and contains fully sequenced, characterized, and authenticated cultures which enable the optimization of metagenomic workflows and microbiome research (*14*). The contents of the product are shown in Figure 3. Upon receipt of ATCC MSA-2002, the lyophilized pellet containing cells was resuspended in 1 mL of cold phosphate-buffered saline and allowed to dissolve for two minutes. The vial was kept on ice during this step to prevent cell lysis. After reconstitution, the product was gently mixed, aliquoted into five Eppendorf tubes (200 μL each), and centrifuged for 10 minutes at 10,000xg at 4°C. The supernatant was then carefully discarded without disturbing the pellet. Tubes were stored on ice until ready for extraction. Two vials of ATCC MSA-2002 were purchased for this study, yielding a total of 10 aliquot tubes. This preparation was performed according to the manufacturer’s instructions.

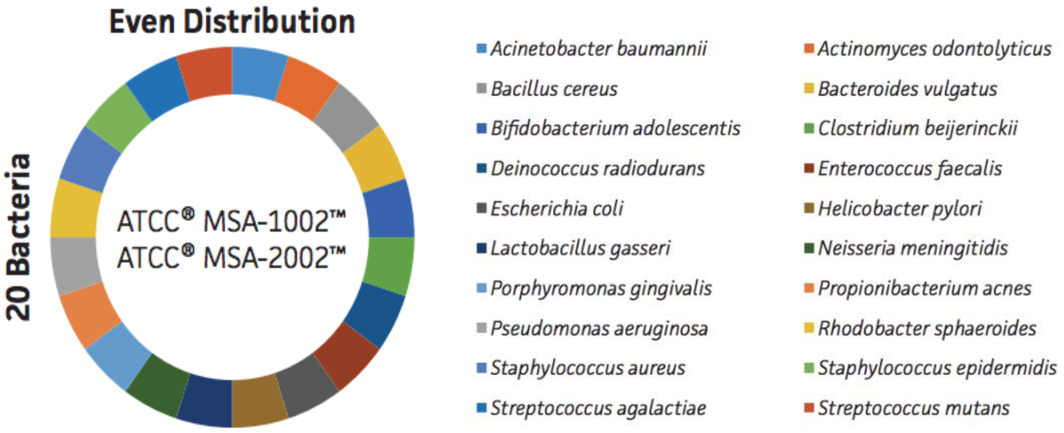

Figure 3. ATCC® 20 Strain Even Mix Whole Cell Material (ATCC® MSA-2002™) (*14*). Fully sequenced mock microbial community that mimics mixed metagenomic sample. Each microorganism represents 5% of the total sample.

16S V4 Amplicon Primers and Adapters

The gene-specific primers used during this research target the 16S V4 region. These sequences were selected from those described in Caporaso et al. (2011) and are detailed in **Table 1** (*27*, *28*). The gene-specific primer sequences are as follows:

16S rRNA 515F Illumina V4 (5'-GTGCCAGCMGCCGCGGTAA-3')

16S rRNA 806R Illumina V4 (5'-GGACTACHVGGGTWTCTAAT-3')

Forward Primer	
Name	16S rRNA 515F Illumina V4
Sequence Structure	5'-(forward adapter overhang)-forward primer-3'
Forward Adapter Overhang and Primer Sequence	5'-(TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG)-GTGCCAGCMGCCGCGGTAA-3'
Reverse Primer	
Name	16S rRNA 806R Illumina V4
Sequence Structure	5'-(reverse adapter overhang)-forward primer-3'
Reverse Adapter Overhang and Primer Sequence	5'-(GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG)-GGACTACHVGGGTWTCTAAT-3'

Table 1. Amplicon Primers and Overhang Adapter Sequences used to Target the 16S V4 Region (*15*, *27, 28*)**.**

Illumina overhang adapter nucleotide sequences (**Table 1**) were added to the gene-specific primers according to the 16S Metagenomics Sequencing Library Preparation protocol which was used for this study (*15*). The gene-specific primer sequences were attached to template DNA in order for the 16S V4 region to be targeted. The overhang adapter sequences must be added to the gene-specific primers so that multiplexing indices can be attached during index PCR (**Figure 4**).

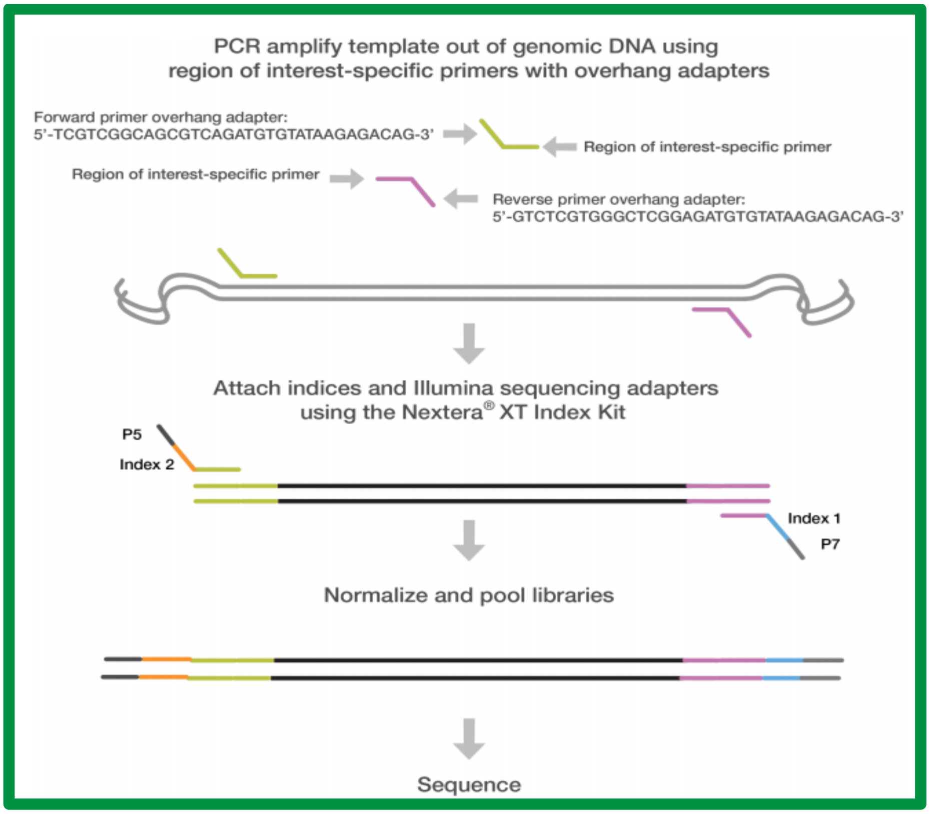
PCR amplify template out of genomic DNA using region of interest-specific primers with overhang adapters
Forward primer overhang adapter:
5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-3'
Region of interest-specific primer
Region of interest-specific primer
Reverse primer overhang adapter:
5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG-3'
Attach indices and Illumina sequencing adapters using the Nextera® XT Index Kit
P5
Index 2
Index 1
P7
Normalize and pool libraries
Sequence

Figure 4. 16S V4 Amplicon Workflow (*15*). Region of interest-specific amplicon primers and overhang adapter sequences were used to amplify templates from genomic DNA. Following this, a subsequent amplification step was performed in order to attach multiplexing indices and Illumina sequencing adapters onto template DNA.

After the initial amplification step was performed, multiplexing indices were added to each sample using the Nextera® XT Index Kit v2, followed by a subsequent limited-cycle amplification step. This is performed in order to attach indices onto template DNA.
The PCR step adds index adapter sequences on both ends of the DNA, which enables dual-indexed sequencing of pooled libraries on the Illumina sequencing platform (*16*). Indices allow each sample to receive its own unique index combination, giving it a type of distinctive barcode.

It is necessary for each sample to have a distinctive barcode to individualize samples when they are pooled together for sequencing.

Illumina MiSeq System

Next-generation sequencing was performed using the Illumina MiSeq System (Illumina Inc., San Diego, California) which integrates cluster generation, amplification, sequencing, and data analysis into a single instrument. Libraries were prepared using indexed or bar-coded adapters that allow the possibility of including up to 96 samples to be sequenced in a single run (*17*). The system utilizes Illumina sequencing by synthesis chemistry (SBS) uses a reversible terminator-based method that detects single bases as they are incorporated into massively parallel DNA strands (*17*). Fluorescent terminator dyes are imaged as each dNTP is added and then cleaved to allow incorporation of the next base. Base calls are made directly from signal intensity measurements during each cycle (*17*). Data could then be directly uploaded from the instrument to the Illumina genomic analysis platform BaseSpace, which enables real-time data management analysis (*18*).

Phase 1: Comparison of Extraction Methods

The first phase of this research compared variations in extraction techniques by performing three different extraction methods using ATCC MSA-2002 aliquots as a mock microbial community. The extraction kits that were tested include FastDNA™ Spin Kit for Soil (MP Biomedicals, LLC, Santa Ana, California), PowerSoil DNA Isolation Kit (MO BIO Laboratories, Inc., Carlsbad, California), and E.Z.N.A.® Mollusc DNA Kit (Omega Bio-tek, Inc., Norcross, Georgia). Each extraction method was performed in triplicate according to manufacturer guidelines with any variations noted. A reagent blank was processed alongside each extraction method. All extracted samples were stored at -20°C until ready for PCR amplification.

For the FastDNA™ extraction, 978 μL of Sodium Phosphate Buffer (provided in extraction kit) was added to the aliquot tube containing the mock microbial community pellet. The tube was gently pulse vortexed and briefly spun down using a microcentrifuge. The entire volume of the tube was then transferred to Lysing Matrix B tube (provided in extraction kit). From this point the manufacturer protocol was followed beginning at the MT Buffer step.

The mock microbial community pellet for the PowerSoil extraction was resuspended by adding 60 μL of Solution C1 (provided in extraction kit) to the aliquot tube. The tube was gently pulse vortexed and briefly spun down using a microcentrifuge. The entire volume of the tube was then transferred to the PowerBead Tube (provided in extraction kit) and gently mixed. From this point the manufacturer protocol was followed beginning at the MO BIO Vortex Adapter step.

For the E.Z.N.A.® Mollusc extraction, 350 μL ML1 Buffer (provided in extraction kit) was added to the aliquot tube containing the mock microbial community pellet. The tube was gently pulse vortexed and briefly spun down using a microcentrifuge in order to resuspend cells. The entire volume of the tube was transferred to a bead tube with the addition of 25 μL Proteinase K Solution (provided in extraction kit). The bead tube was then placed on the FastPrep bead beating instrument for 30 seconds at a speed setting of 6.0 meters per second for homogenization. From this step, the manufacturer protocol was followed beginning at the 60°C incubation step.

Phase 2: 16S Metagenomic Sequencing Library Preparation

Sample sets from each extraction method underwent PCR amplification in duplicate to minimize PCR random error. Three different polymerases were analyzed during this phase. These were: AccuPrime™ *Taq* DNA Polymerase (Thermo Fisher Scientific Corporation, Carlsbad, California), NEB Q5 DNA Polymerase (New England Biolabs, Inc., Ipswich, Massachusetts), and Platinum™ Hot Start (Thermo Fisher Scientific Corporation, Carlsbad, California). A positive control in the form of a known *Lactobacillus acidophilus* NCFM strain, that was previously isolated and identified, was used during each amplification process. Molecular grade water was used as the negative control.

Reaction conditions for amplification using AccuPrime™ *Taq* DNA Polymerase included 2.5 µL of 10X AccuPrime PCR Buffer II, 1.0 µL 16S rRNA 515F Illumina V4 (10 µM), 1.0 µL 16S rRNA 806R Illumina V4 (10 µM), 2.5 µL 10X BSA (1 mg/µL), 0.1 µL AccuPrime *Taq* High Fidelity (5 U/µL), 4 µL extracted DNA, and 13.9 µL molecular grade water per tube for a total volume of 25 µL. Reactions were amplified using the Bio-Rad C1000 Touch™ thermal cycler under the following cycling conditions: initial activation of *Taq* polymerase at 94°C for 2 minutes, followed by 25 cycles of 30 seconds of denaturation at 94°C, 40 seconds of annealing at 55°C, and 40 seconds of extension at 68°C, followed by a final 5 minute extension at 68°C and a 4°C indefinite hold.

Reaction conditions for NEB Q5 DNA Polymerase included 5.0 µL of 5X Q5 Reaction Buffer, 1.25 µL 16S rRNA 515F Illumina V4 (10 µM), 1.25 16S rRNA 806R Illumina V4 (10 µM), 2.5 µL of 10X BSA (1 mg/µL), 2.0 µL of 2.5 mM dNTP, 0.25 µL Q5 High Fidelity

Polymerase, 4 μL extracted DNA, and 8.75 μL of molecular grade water per tube to yield a total volume of 25 μL. Reactions were amplified using a Bio-Rad C1000 Touch™ thermal cycler under the following cycling conditions: initial activation of polymerase at 98°C for 30 seconds, followed by 25 cycles of 30 seconds of denaturation at 98°C, 40 seconds of annealing at 55°C, and 40 seconds of extension at 72°C, followed by a final 2 minute extension at 72°C and a 4°C indefinite hold.

Amplification mix for Platinum™ Hot Start included 12.5 μL of Platinum Hot Start 2X Master Mix, 0.5 μL 16S rRNA 515F Illumina V4 (10 μM), 0.5 μL 16S rRNA 806R Illumina V4 (10 μM), 2.5 μL 10X BSA (1 mg/μL), 4 μL extracted DNA, and 5.0 μL of molecular grade water per tube to produce a total volume of 25 μL. Reactions were amplified using the Bio-Rad C1000 Touch™ thermal cycler under the following cycling conditions: initial activation of *Taq* polymerase at 94°C for 2 minutes, followed by 25 cycles of 30 seconds of denaturation at 94°C, 40 seconds of annealing at 55°C, and 40 seconds of extension at 72°C, with a final 2 minute hold at 72°C.

After cycling, all amplification products, reagent blanks, and controls underwent agarose gel electrophoresis using Thermo Scientific 6X Orange Loading Dye and Phenix Research 100 bp DNA ladder. This was performed to ensure proper amplification took place, and that sample PCR products were between 300-400 bp. Duplicate amplification samples were pooled after determining that all samples were within the optimal 300-400 bp range.

A PCR purification process was performed using the Beckman Coulter Agencourt AMPure XP system, which is a solid-phase paramagnetic bead technology that uses magnetic separation in order to purify DNA (*19*). AMPure XP uses an optimized buffer that works by selectively binding DNA fragments that are 100 bp and larger to paramagnetic beads. Excess primers, nucleotides, salts, and enzymes are then washed away from the PCR product. PCR clean-up was performed according to manufacturer guidelines.

A subsequent limited-cycle amplification step was completed using Nextera® XT Index Kit v2 to attach Index 1 adapters (N7XX), Index 2 adapters (S5XX), and multiplexing indices. Reaction conditions for samples previously amplified using AccuPrime™ *Taq* DNA Polymerase included 5.0 μL of 10X AccuPrime PCR Buffer II, 5.0 μL Index 1 adapters (N7XX), 5.0 μL Index 2 adapters (S5XX), 0.2 μL AccuPrime *Taq* High Fidelity (5 U/μL), 5 μL DNA, and 29.8 μL molecular grade water per tube.

Reaction conditions for samples previously amplified with NEB Q5 DNA Polymerase included 10.0 μL of 5X Q5 Reaction Buffer, 5.0 μL Index 1 adapters (N7XX), 5.0 μL Index 2 adapters (S5XX), 4.0 μL of 2.5 mM dNTP, 0.50 μL Q5 High Fidelity Polymerase, 5 μL DNA, and 20.5 μL of molecular grade water per tube. Reaction conditions for samples previously amplified using Platinum™ Hot Start included 25 μL of Platinum Hot Start 2X Master Mix, 5.0 μL Index 1 adapters (N7XX), 5.0 μL Index 2 adapters (S5XX), 5 μL extracted DNA, and 10.0 μL of molecular grade water per tube.

All sample reaction tubes had a final volume of 50 μL per tube. Samples were prepared according to the Nextera XT DNA Library Prep Reference Guide and centrifuged at 280 × *g* at 20°C for 1 minute before being loaded on to the thermocycler (*16*). Reactions were amplified using the Bio-Rad C1000 Touch™ thermal cycler under the following cycling conditions: 72°C for 3 minutes, 95°C for 30 seconds, 12 cycles of 95°C for 10 seconds, 55°C for 30 seconds, and 72°C for 30 seconds, followed by 72°C for 5 minutes and a final hold at 10°C. Following amplification, a subsequent PCR clean-up step was performed using the previously described Beckman Coulter Agencourt AMPure XP system.

Libraries were quantified using the Qubit® 2.0 Fluorometer. This step was performed so that the concentration of each sample in the library could be normalized to ensure more equal sample representation in the pooled sequencing library. The Qubit® quantifies DNA using the highly sensitive and accurate fluorescence-based Qubit™ quantitation assays. A 1:100 dilution of each sample was made and total double-stranded DNA (dsDNA) concentration was quantified using the Qubit® dsDNA HS Assay Kit (*19*). From these results, samples were normalized and pooled together in preparation for sequencing with the MiSeq system.

Phase 3: Next-Generation Sequencing using the MiSeq System

The MiSeq reagent cartridge, flow cell, and reagent bottles were prepared according to the MiSeq Sequencing System Guide (*21*). The instrument uses a double-sided, single-lane flow cell and reagent cartridge supplied in kit form (*17*). The required 600 μL of pooled libraries were loaded onto the reagent cartridge and sequencing was performed per manufacturer guidelines as described in the MiSeq Sequencing System Guide. During cluster generation, sequencing templates are immobilized by oligonucleotides on the surface of the flow cell and clusters are formed by way of bridge amplification. During bridge amplification, the bound DNA strand folds over and attaches to the oligonucleotide that is complementary to its adapter sequence. Polymerases then generate a complementary strand, forming a double stranded bridge. The bridge is then denatured, resulting in two single stranded copies of the molecule. This process is repeated numerous times and occurs simultaneously across the flow cell.

Sequencing begins after cluster generation is complete. With each sequencing cycle, fluorescently tagged nucleotides compete for addition to the growing chain with only one nucleotide being incorporated at a time based on the sequence of the template. After the addition of each nucleotide, clusters are excited by a light source and a fluorescent signal is emitted that is exclusive to each of the four fluorescently-labeled ddNTPs. The base call is determined according to the emission wavelength and the signal intensity. An image is then captured, and the identity of the base is recorded. This process is repeated for each sequencing cycle (*21*).

Phase 4: Bioinformatics Analysis

The genomic data that was generated from the MiSeq was processed using mothur v.1.39.4 following the MiSeq SOP (*24*). mothur is an open-sourced software package that is used for bioinformatics data analysis (*23*). Overall, paired end sequences were constructed, the primers were trimmed, and sequences were excluded from the data set if they were short (< 100bp) or of low quality (homopolymers > 8). The SILVA reference database was used to build several sequence alignments. Any redundant sequences that were found were minimized by applying the unique.seqs and precluster (diffs=2) command. Chimeras were identified and removed using UCHIME (*25*). The Operational Taxonomic Units (OTUs) from the dataset were assigned with a 97% sequence similarity based on the average neighbor clustering algorithm. Taxonomic classification was performed using the Greengenes database. Greengenes is a 16S rRNA gene database that provides chimera screening, standard alignment, and taxonomic classification using multiple published taxonomies (*22*). Percent relative abundance was calculated, and the obtained taxonomic results were compared to the known genomic data of the ATCC mock microbial community. The sample variance, S^2, was calculated in order to determine the spread of the data sets. Statistical methods were also employed to determine the extent of deviation among microbial communities in different samples by performing UniFrac (*30*) and Principal Coordinate Analysis (PCoA) (*31*).

Sample variance was calculated as a measure of the spread of the data sets in relation to the expected results. The variance of a sample is defined mathematically as the average of the squared differences from the mean (34). Variance was calculated by subtracting the mean from each number in the obtained data set and then squaring the result. The squared differences were then divided by the sample size minus one (n-1). Because Staphylococcus sp. and Streptococcus sp. were unable to achieve species-level taxonomic resolution, the sample size for the data sets was n=18.

The taxonomic classifications that were obtained from the Greengenes database were compared to the known genomic data from the mock microbial community and evaluated for divergence. The raw data was converted to percent relative abundance based on the total number of sequencing reads per sample. A conditional formatting rule was applied to exclude reads that equaled less than one percent. Samples were grouped for comparison based on the type of polymerase and extraction method. An insignificant number of reads were detected in some of the PCR negative controls. This was attributed to background noise and was accounted for by subtracting the number of reads found in the negative control from its corresponding sample before converting the raw data to percent relative abundance.

Data was normalized in order to correct for copy number variation of the 16S rRNA gene within each species. This was accomplished by dividing the number of sequencing reads by the 16S rRNA gene copy number relative to each species. An expected distribution was calculated based on the target gene copy number in each organism, and assuming 100% extraction and amplification efficiency. Values were then recalculated as total percent and displayed in table format.

CHAPTER III

RESULTS AND DISCUSSION

Overview of Taxonomic Classifications

It was noted that *Propionibacterium acnes* was only present upon amplification with NEB Q5 DNA Polymerase and could not be identified past the family level. There was no evidence of detection at the species or genus level. It was speculated that this organism may have had complications with lysis during the extraction phase, as indicated by its near-total absence from all extraction kits, as well as issues with amplification based on the raw data. However, it is also possible that the primers used in this study are poorly suited for the amplification of the 16S gene from this organism. However, as stated, some reads were detected that corresponded to this organism, and were highest when using the FastDNA extraction kit and Q5 polymerase.

There was also difficulty with the classification of the *Pseudomonas aeruginosa* strain. Although the organism was detected, the Greengenes database had trouble classifying this bacterium at the genus and species level. Instead, it was classified in the family *Pseudomonadaceae* and then identified as “unclassified” at the genus level. This unclassified result was able to be tracked back to *Pseudomonadaceae* by using the taxonomic rank ID. The raw data were converted to percent relative abundance and samples were graphically grouped based on polymerase type (**Figure 5**) and by method of extraction (**Figure 6**) for comparison.

Expected Species	Identified Species	Expected Species	Identified Species
Acinetobacter baumannii	*Acinetobacter sp.*	*Lactobacillus gasseri*	*Lactobacillus sp.*
Actinomyces odontolyticus	*Actinomyces sp.*	*Neisseria meningitidis*	*Neisseria sp.*
Bacillus cereus	*Bacillus cereus*	*Porphyromonas gingivalis*	*Porphyromonas sp.*
Bacteroides vulgatus	*Bacteroides sp.*	*Propionibacterium acnes*	*Propionibacterium sp.*
Bifidobacterium adolescentis	*Bifidobacterium adolescentis*	*Pseudomonas aeruginosa*	*Pseudomonas sp.*
Clostridium beijerinckii	*Clostridium sp.*	*Rhodobacter sphaeroides*	*Rhodobacter sphaeroides*
Deinococcus radiodurans	*Deinococcus sp.*	*Staphylococcus aureus*	*Staphylococcus spp.*
Enterococcus faecalis	*Enterococcus sp.*	*Staphylococcus epidermidis*	*Staphylococcus spp.*
Escherichia coli	*Escherichia coli*	*Streptococcus agalactiae*	*Streptococcus agalactiae*
Helicobacter pylori	*Helicobacter pylori*	*Streptococcus mutans*	*Streptococcus spp.*

Table 2. Expected Species vs. Identified Species. **“**Expected species” consisted of the known members of the ATCC mock community product. “Identified species” represent the closest taxonomic prediction generated from the analyses of the samples.

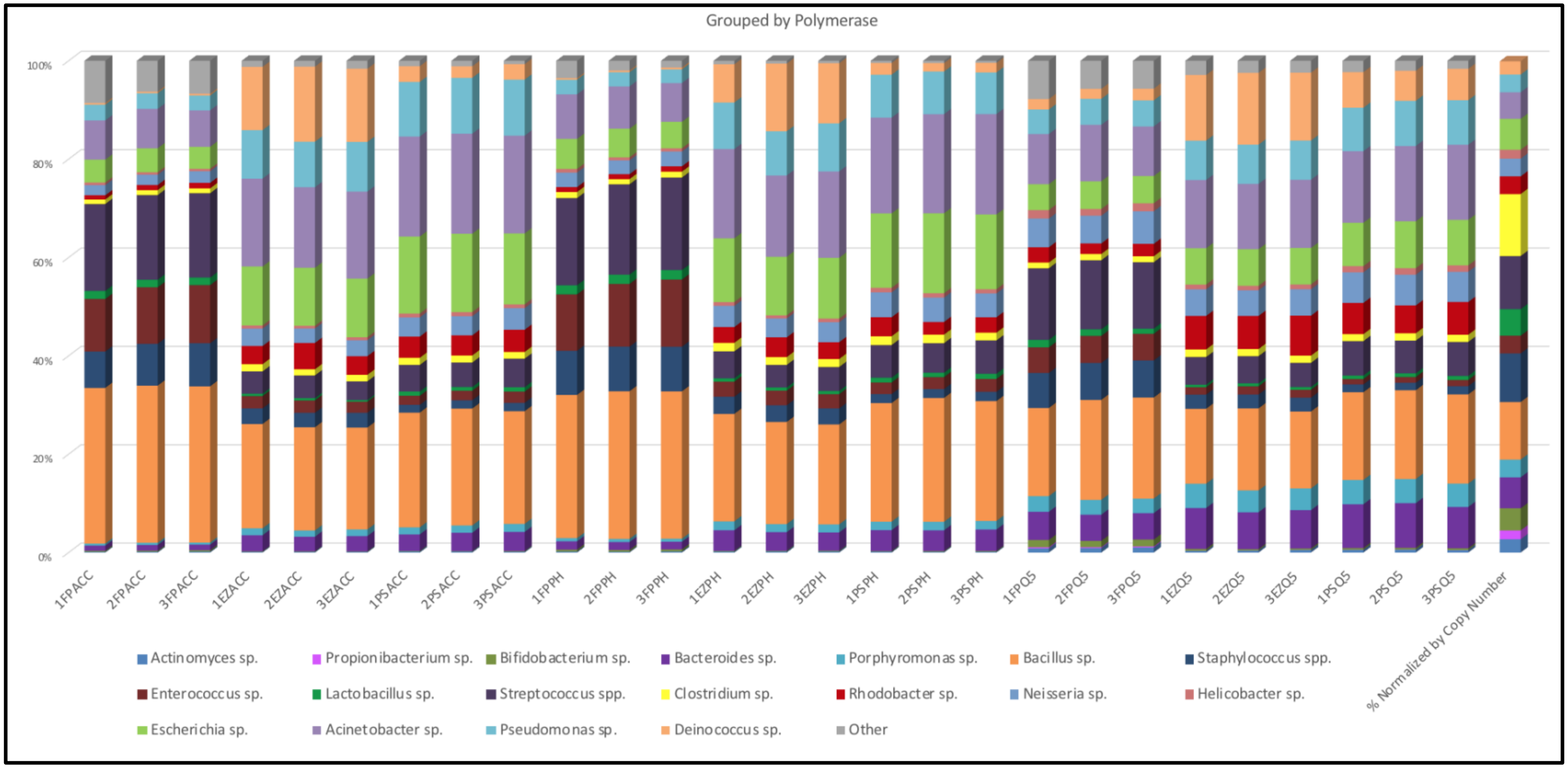

Figure 5. **Percent Relative Abundance of Mock Microbial Community Grouped by Polymerase**. Each extraction method was performed in triplicate (replicates shown) and each group of extracted samples was amplified with each of the three polymerases. Sample names correspond to: (replicate number 1-3)(Extraction method)(polymerase). Extraction methods: FP = FastDNA, EZ = E.Z.N.A., PS = PowerSoil. Polymerase enzymes: ACC = AccuPrime *Taq* DNA Polymerase, PH = Platinum Hot Start, Q5 = NEB Q5 DNA Polymerase.

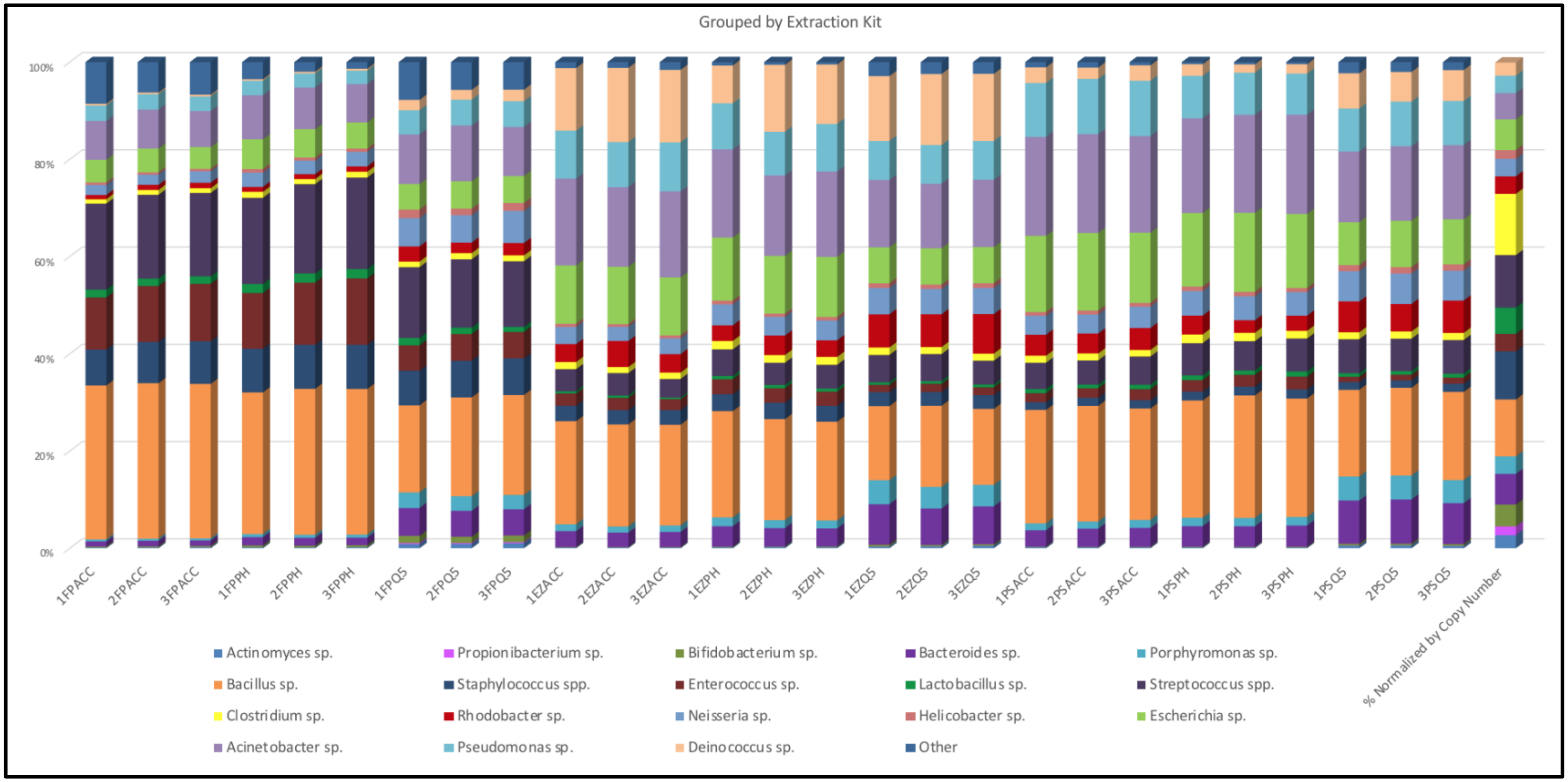

Figure 6. **Percent Relative Abundance of Mock Microbial Community Grouped by Extraction Method**. Each extraction method was performed in triplicate (replicates shown) and each group of extracted samples was amplified with each of the three polymerases. Sample names correspond to: (replicate number 1-3)(Extraction method)(polymerase). Extraction methods: FP = FastDNA, EZ = E.Z.N.A., PS = PowerSoil. Polymerase enzymes: ACC = AccuPrime *Taq* DNA Polymerase, PH = Platinum Hot Start, Q5 = NEB Q5 DNA Polymerase.

Table 3 depicts the expected species in the sample and the number of sequencing reads corresponding to each species as a percentage of the total. These values were normalized in order to correct for copy number variation of the 16S rRNA gene within each species and the results are shown in **Table 4**. Differences in 16S rRNA gene copy number have the potential to skew data. For instance, an organism with a high number of sequence reads could represent a high copy organism that is present in low abundance or a low copy organism that is present in high abundance. Performing normalization based on the 16S rRNA gene copy number assists in adjusting for this error source (*26*).

Expected Species	**Total number of sequencing reads (%)**	**Expected Species**	**Total number of sequencing reads (%)**
Acinetobacter baumannii	20.0%	*Helicobacter pylori*	1.0%
Actinomyces odontolyticus	0.0%	*Lactobacillus gasseri*	1.0%
Bacillus cereus	23.3%	*Neisseria meningitidis*	4.0%
Bacteroides vulgatus	3.7%	*Porphyromonas gingivalis*	1.3%
Bifidobacterium adolescentis	0.0%	*Propionibacterium acnes*	0.0%
Clostridium beijerinckii	1.0%	*Pseudomonas aeruginosa*	11.0%
Deinococcus radiodurans	2.6%	*Rhodobacter sphaeroides*	4.1%
Enterococcus faecalis	2.0%	*Staphylococcus spp.*	2.0%
Escherichia coli	15.3%	*Streptococcus spp.*	5.3%

Table 3. **Percentage of Sequence Reads per Organism using Data from PowerSoil/AccuPrime Analysis.**

Expected Species	**Sequencing reads (%)**	**16S rRNA Copy Number**	**Expected number of reads corrected by copy number**	**Sequencing reads normalized by copy number (%)**
Acinetobacter baumannii	20.0%	6	3025	20.2%
Actinomyces odontolyticus	0.0%	3	58	0.4%
Bacillus cereus	23.3%	13	1613	10.8%
Bacteroides vulgatus	3.7%	7	473	3.2%
Bifidobacterium adolescentis	0.0%	5	19	19.0%
Clostridium beijerinckii	1.0%	14	90	0.6%
Deinococcus radiodurans	2.6%	3	886	5.9%
Enterococcus faecalis	2.0%	4	460	3.1%
Escherichia coli	15.3%	7	1961	13.1%
Helicobacter pylori	1.0%	2	355	2.4%
Lactobacillus gasseri	1.0%	6	127	0.8%
Neisseria meningitidis	4.0%	4	923	6.2%
Porphyromonas gingivalis	1.3%	4	345	2.3%

Propionibacterium acnes	0.0%	2	1	0.0%
Pseudomonas aeruginosa	11.0%	4	2542	17.0%
Rhodobacter sphaeroides	4.1%	4	976	6.5%
Staphylococcus spp.	2.0%	5	300	2.0%
Streptococcus spp.	5.3%	6	819	5.5%

Table 4. **Sequence Reads per Organism after Normalization Based on 16S rRNA Copy Number**. Sequence reads per organism using an average of replicates within the set from PowerSoil/AccuPrime analysis. Values were normalized in order to correct for copy number variation of the 16S rRNA gene within each species. 16S gene copy numbers were obtained using records from *rrn*DB (*26*).

Sample Variance

Sample variance was calculated as a measure of the spread of the data sets in relation to the expected result. The expected sample variance was 0%. The variance between samples ranged from 0.22% to 0.64%, with samples with a combination of FastDNA and NEB Q5 showing the least amount of variance and FastDNA and AccuPrime combination showing the greatest amount of variance.

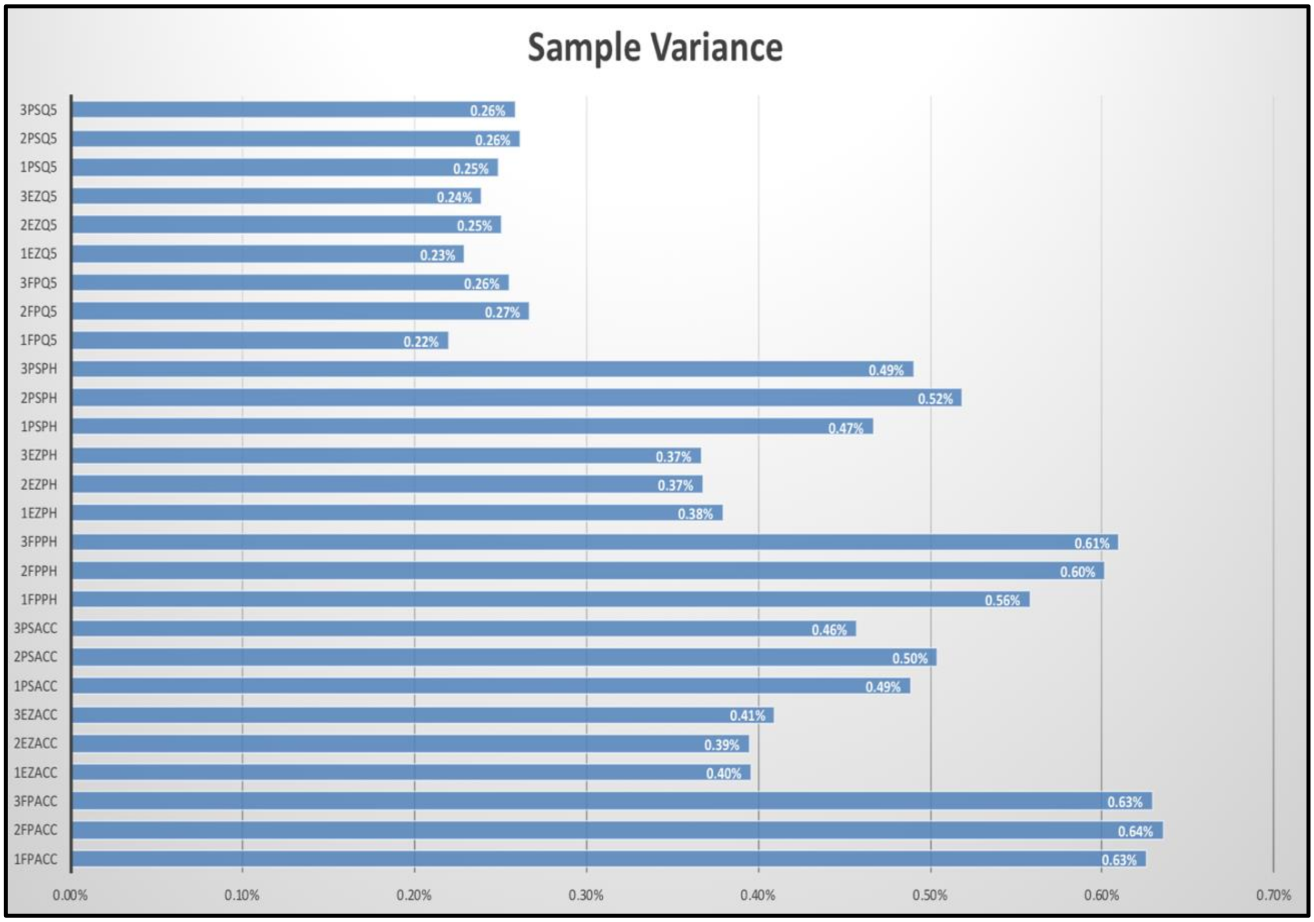

Figure 7. Sample Variance. Sample variance was compared between all samples, which was determined for each sample by subtracting the mean from each number in the obtained data set and then squaring the result. The squared differences were then divided by the sample size minus one (n-1). Each extraction method was performed in triplicate (replicates shown) and each group of extracted samples was amplified with each of the three polymerases. Sample names correspond to: (replicate number 1-3)(Extraction method)(polymerase). Extraction methods: FP = FastDNA, EZ = E.Z.N.A., PS = PowerSoil. Polymerase enzymes: ACC = AccuPrime *Taq* DNA Polymerase, PH = Platinum Hot Start, Q5 = NEB Q5 DNA Polymerase.

Principal Coordinate Analysis (PCoA)

The Principal Coordinate Analysis was used to visualize similarities and dissimilarities in the obtained data. By using PCoA individual or group differences can be visualized, and outliers can be shown. Weighted analysis accounts for abundance of the observed organisms, while unweighted analysis only considers whether the organisms are present or absent in the sample. Based on the unweighted PCoA graph for polymerase grouping (**Figure 7**) it can be seen that AccuPrime *Taq* DNA Polymerase (Group 1) and Platinum Hot Start (Group 2) amplify similarly. This was also apparent when looking at the raw data for the sequence reads where NEB Q5 seemed to outperform the other enzymes based on the number of organisms that were amplified with this enzyme. Likewise, in the weighted PCoA graph (**Figure 8**) that was grouped based on extraction method, E.Z.N.A. and PowerSoil are shown to extract comparably, while the FastDNA kit proves to be considerably different.

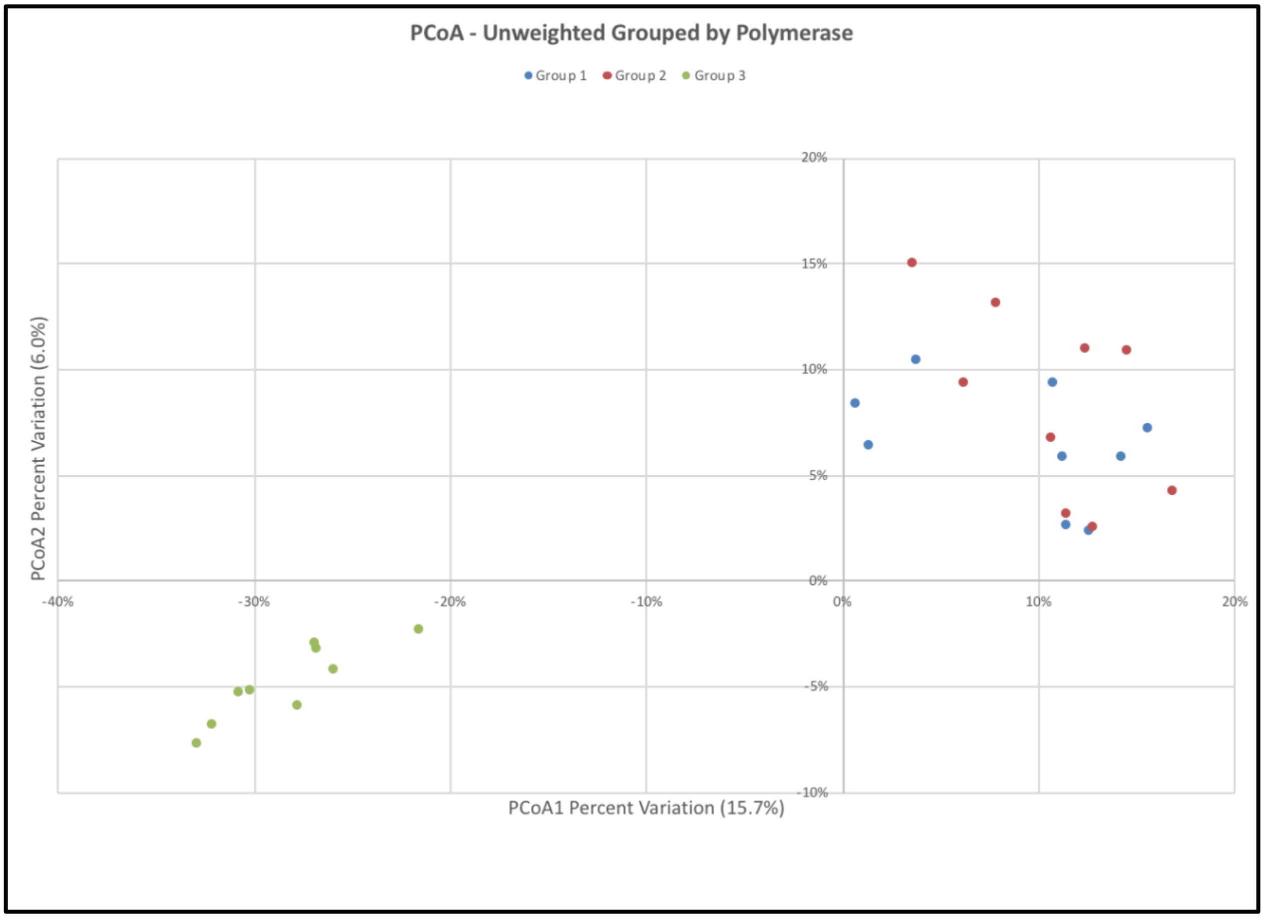

Figure 8. PCoA - Unweighted Grouped by Polymerase. Plots were generated for unweighted UniFrac distances using data generated from mothur. These two components explain 21.7% of the variance. Samples in Group 1 were amplified with AccuPrime *Taq* DNA Polymerase and are shown in blue; Group 2 samples were amplified with Platinum Hot Start and are shown in red; and Group 3 samples, amplified using NEB Q5 DNA Polymerase, are shown in green.

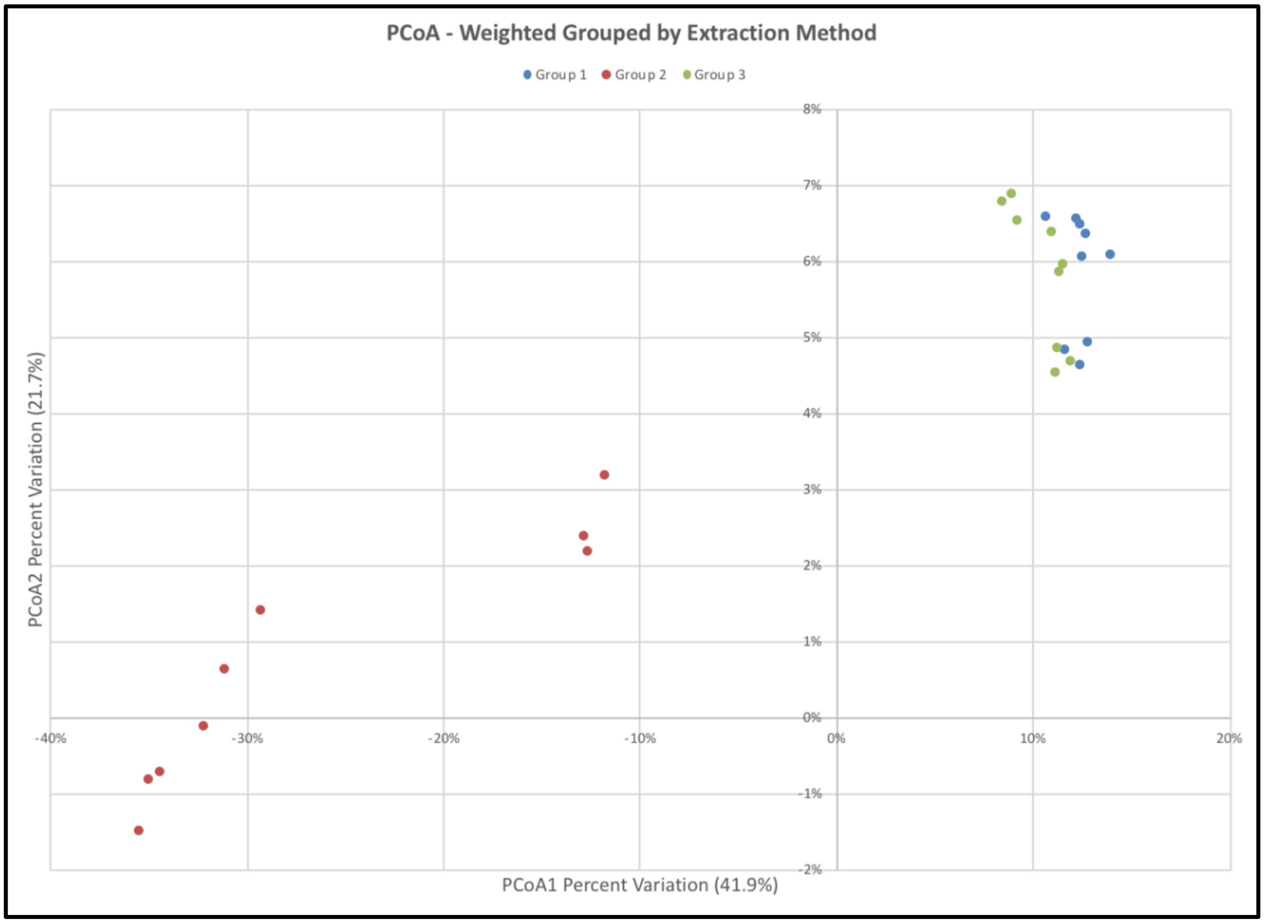

Figure 9. PCoA - Weighted Grouped by Extraction Method. Plots were generated for weighted UniFrac distances using data generated from mothur. These two components explain 63.6% of the variance. Samples in Group 1 were extracted with E.Z.N.A. and are shown in blue; Group 2 samples were extracted with FastDNA and are shown in red; and Group 3 PowerSoil extractions are shown in green.

CHAPTER IV

CONCLUSIONS

This study began to address the need for validation and standardization of microbial profiling for forensic use and set out to address specific sources of errors that could occur when implementing microbiome analysis using next-generation sequencing. Differential extraction and PCR bias were two major areas of concern focused on by the study. Discrepancies were found to exist between the data that were obtained from sequencing the mock microbial community and the expected genomic data that were provided by ATCC. One major concern was that although each bacterial strain was reported to compose 5% of the whole cell mixture, sequencing data from each of the 27 samples failed to reflect this composition. This was due in part to the variation in gene copy number targeted by the PCR process used in this study. Therefore, an expected distribution was calculated based on the target gene copy number in each organism, and assuming 100% extraction and amplification efficiency.

Extraction Methods

Resultant analyses revealed very different performances between extraction kits, with FastDNA preferentially extracting more Gram-positive organisms compared to the other two extraction kits. Looking at the data for *Bacillus sp.* as a representative of Gram-positive organisms, differences in extraction methods ranged from an average of 30% relative abundance with FastDNA, to 21% with PowerSoil, and 18% with E.Z.N.A. coming in last. The E.Z.N.A. and PowerSoil kits did result in increased numbers of Gram-negative bacteria in general; however, this may be an artefact due to relative lack of lysis of Gram-positive cells, leaving more Gram-negative DNA available for PCR amplification. It was suspected that the bead beating method would facilitate lysis of higher numbers of Gram-positive bacteria, but this hypothesis was not able to be directly tested in this study due to all of the extraction kits incorporating some form of bead beating process into their respective protocols, including the E.Z.N.A. kit, which was modified to include a FASTPrep bead-beating protocol similar to that used in the FastDNA extraction method. Nevertheless, based on the data from the weighted and unweighted PCoA, differences were seen across the extraction kits, with E.Z.N.A. and PowerSoil extracting comparably.

Enzyme Performance

When looking at the performance of the polymerases in relation to extraction methods, there is a clear indication that NEB Q5 outperformed both AccuPrime and Platinum Hot Start in regard to the number of organisms that were amplified per sample. On average the samples with a combination of FastDNA and NEB Q5 amplified approximately 5-6 more species compared to other kit/enzyme combinations. However, this may not necessarily be a positive outcome due to the fact that Q5 could be leading to other non-specific amplification and causing bias. This was manifested in additional bacteria, shown as "other", that were not known to be in the sample. *Propionibacterium acnes* was only detected at low numbers in the raw data when amplified with NEB Q5 DNA Polymerase, suggesting that this particular bacterium either did not extract efficiently, amplify efficiently, or both. It was also noted that although NEB Q5 outperformed other enzymes in terms of the number of species amplified, on average both AccuPrime and Platinum Hot Start yielded higher overall sequencing reads per sample.

Bioinformatics

The Greengenes database was unable to successfully classify *Pseudomonas aeruginosa* in any of the analyzed samples. Upon reanalysis of the raw data using the RDP database (*29*) it was observed that *Pseudomonas sp.* was classified with a high number of reads. This may be due to the V4 region of the 16S rRNA gene being indiscriminate for *P. aeruginosa,* or it could be that the database was not curated properly for this organism. The 16S rRNA gene is highly conserved, and this key feature undoubtedly contributed to the limited taxonomic discrimination achieved in this research. Indeed, one of the main drawbacks of performing microbiome analysis on one portion of the 16S rRNA gene as a single target is that it often lacks species-level taxonomic resolution (*7*).

The results of this study indicate that there is a clear need for validation of microbial forensic methods on a much larger scale, which will enable further evaluation of factors that could affect downstream methods such as next-generation sequencing. Putting better standards in place will not only provide data and potential recommendations that others can use for microbial analysis in the forensic setting, but also assist in ensuring the accuracy and reliability of future microbial forensic analyses.

APPENDIX

Kingdom	**Phylum**	**Order**	**Family**	**Genus**
Bacteria	Proteobacteria	Pseudomonadales	Moraxellaceae	*Acinetobacter*
Bacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	*Actinomyces*
Bacteria	Firmicutes	Bacillales	Bacillaceae	*Bacillus*
Bacteria	Bacteroidetes	Bacteroidales	Bacteroidaceae	*Bacteroides*
Bacteria	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	*Bifidobacterium*
Bacteria	Firmicutes	Clostridiales	Clostridiaceae	*Clostridium*
Bacteria	Deinococcus-Thermus	Deinococcales	Deinococcaceae	*Deinococcus*
Bacteria	Firmicutes	Lactobacillales	Enterococcaceae	*Enterococcus*
Bacteria	Proteobacteria	Enterobacteriales	Enterobacteriaceae	*Escherichia*
Bacteria	Proteobacteria	Campylobacterales	Helicobacteraceae	*Helicobacter*
Bacteria	Firmicutes	Lactobacillales	Lactobacillaceae	*Lactobacillus*
Bacteria	Proteobacteria	Neisseriales	Neisseriaceae	*Neisseria*
Bacteria	Bacteroidetes	Bacteroidales	Porphyromonadaceae	*Porphyromonas*
Bacteria	Actinobacteria	Propionibacteriales	Propionibacteriaceae	*Propionibacterium*

Bacteria	Proteobacteria	Pseudomonadales	Pseudomonadaceae	*Pseudomonas*
Bacteria	Proteobacteria	Rhodobacterales	Rhodobacteraceae	*Rhodobacter*
Bacteria	Firmicutes	Bacillales	Staphylococcaceae	*Staphylococcus*
Bacteria	Firmicutes	Lactobacillales	Streptococcaceae	*Streptococcus*

Table 5. Taxonomic Classification of Mock Microbial Community (*32*, *33*).

REFERENCES

1. Schmedesa, Sarah E., Antti Sajantilaab, "Expansion of Microbial Forensics." *Journal of Clinical Microbiology*, 1 Aug. 2016, jcm.asm.org/lookup/doi/10.1128/JCM.00046-16.
2. *NIH Human Microbiome Project*, hmpdacc.org/hmp/.
3. Sender, Ron, et al. "Revised Estimates for the Number of Human and Bacteria Cells in the Body." *PLOS Biology*, vol. 14, no. 8, 2016, doi: 10.1371/journal.pbio.1002533.
4. Hampton-Marcell, Jarrad T., et al. "The human microbiome: an emerging tool in forensics." *Microbial Biotechnology*, vol. 10, no. 2, 2017, pp. 228–230., doi:10.1111/1751-7915.12699.
5. Dawson, Jim. "Solving Crimes with Soil Bacteria." *National Institute of Justice*, 14 Sept. 2017, www.nij.gov/topics/forensics/evidence/trace/Pages/solving-crimes-with-soil-bacteria.aspx.
6. 16S rRNA Sequencing, Illumina, Inc., www.illumina.com/areas-of-interest/microbiology/microbial-sequencing-methods/16s-rrna-sequencing.html.
7. Allen, Michael, and Michael LaMontagne. "Microbial Genetics and Systematics." *Carrion Ecology, Evolution, and Their Applications*, June 2015, pp. 403–420., doi:10.1201/b18819-22.

8. Salipante, Stephen J., et al. "Performance Comparison of Illumina and Ion Torrent Next-Generation Sequencing Platforms for 16S rRNA-Based Bacterial Community Profiling." *Applied and Environmental Microbiology*, vol. 80, no. 24, 2014, pp. 7583–7591., doi:10.1128/aem.02206-14.

9. Bacterial Endospores, https://micro.cornell.edu/research/epulopiscium/bacterial-endospores.

10. Cornell, B., Gram staining, http://ib.bioninja.com.au/options/untitled/b1-microbiology-organisms/gram-staining.html.

11. Kalle, Elena, et al. "Multi-Template Polymerase Chain Reaction." *Biomolecular Detection and Quantification*, vol. 2, 2014, doi: 10.1016/j.bdq.2014.11.002.

12. Peter McInerney, Paul Adams, and Masood Z. Hadi, "Error Rate Comparison during Polymerase Chain Reaction by DNA Polymerase," Molecular Biology International, vol. 2014, Article ID 287430, 8 pages, 2014. doi:10.1155/2014/287430.

13. Daubert v. Merrell Dow Pharmaceuticals, Inc. *(1993) 509 U.S. 579, 589* https://supreme.justia.com/cases/federal/us/509/579/case.html.

14. 20 Strain Even Mix Whole Cell Material (ATCC® MSA-2002™). *ATCC ® MSA-2002*™, www.atcc.org/products/all/MSA-2002.aspx.

15. 16S Metagenomic Sequencing Library Preparation, Illumina, Inc., https://support.illumina.com/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf.

16. Nextera XT DNA Library Prep Kit Reference Guide, Illumina, Inc., https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_nextera/nextera-xt/nextera-xt-library-prep-reference-guide-15031942-03.pdf.

17. MiSeq System Specification Sheet, Illumina, Inc., www.illumina.com/documents/products/datasheets/datasheet_miseq.pdf.

18. BaseSpace Sequence Hub, Illumina, Inc., www.illumina.com/basespace.

19. Agencourt AMPure XP PCR Purification, Agencourt Bioscience, https://www.beckmancoulter.com/wsrportal/bibliography?docname=Protocol_000387v001.pdf.

20. Qubit 2.0 Fluorometer, Life Technologies, https://tools.thermofisher.com/content/sfs/manuals/mp32866.pdf.

21. MiSeq Sequencing System Guide, Illumina, Inc., https://support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/miseq/miseq-system-guide-15027617-02.pdf.

22. Greengenes Database, http://greengenes.secondgenome.com/.

23. Mothur, Department of Microbiology and Immunology at The University of Michigan, www.mothur.org/.

24. Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., . . . Weber, C. F. (2009). Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial

Communities. *Applied and Environmental Microbiology,75*(23), 7537-7541. doi:10.1128/aem.01541-09.

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

26. Stoddard S.F, Smith B.J., Hein R., Roller B.R.K. and Schmidt T.M. (2015) *rrn*DB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. *Nucleic Acids Research* 2014; doi: 10.1093/nar/gku1201.

27. Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., Noah Fierer, N., & Knight, R. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci USA 108, 4516-4522. http://doi.org/10.1073/pnas.1000080107.

28. Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., Noah Fierer, N., & Knight, R. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci USA 108, 4516-4522. http://doi.org/10.1073/pnas.1000080107.

29. Wang, Q, G. M. Garrity, J. M. Tiedje, and J. R. Cole. 2007. Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. Appl Environ Microbiol. 73(16):5261-5267; doi: 10.1128/AEM.00062-07 [PMID: 17586664].

30. Lozupone C, Lladser ME, Knights D, Stombaugh J, Knight R. UniFrac: an effective distance metric for microbial community comparison. The ISME journal. 2011;5(2):169-172. doi:10.1038/ismej.2010.133.

31. Dray, S., Legendre, P., & Peres-Neto, P. R. (2006). Spatial modelling: A comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). Ecological Modelling, 196(3-4), 483-493. doi:10.1016/j.ecolmodel.2006.02.015.

32. Taxonomy Index. (n.d.). Retrieved April 10, 2018, from https://microbewiki.kenyon.edu/index.php/Taxonomy_Index

33. Wikipedia. (n.d.). Retrieved April 10, 2018, from https://en.wikipedia.org/

34. Sample Variance: Simple Definition, How to Find it in Easy Steps. (n.d.). Retrieved April 18, 2018, from http://www.statisticshowto.com/sample-variance/.