2020

A Eukaryotic Signature Based Method for Identification of Saliva, Feces, and Urine

Alyssa Daniels
Virginia Commonwealth University

A Eukaryotic Signature Based Method for Identification of Saliva, Feces, and Urine

FRSC 793: Technical Report

Researcher: Alyssa Daniels

Date Approved: May 5, 2020

Partial Fulfillment Statement

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at Virginia Commonwealth University.

Host Laboratory

Room 2001, Harris Hall South

Department of Forensic Science, Virginia Commonwealth University

1015 Floyd Avenue, Richmond, VA 23284-3079

Research Mentor

Dr. Baneshwar Singh, Ph. D

Assistant Professor, Department of Forensic Science

Virginia Commonwealth University

Faculty Mentor

Dr. Sarah Seashols-Willams, Ph. D

Assistant Professor, Department of Forensic Science

Virginia Commonwealth University

Committee Member

Paul Brooks

Associate Professor, School of Business

Virginia Commonwealth University

Contents

Abstract

The conception of the Human Microbiome Project advanced the understanding of microbial communities in the human body and previous research has established that unique microbial signatures can help distinguish each body fluid. While these signatures have been developed for the prokaryotic microbiome, the next step is the examination of the eukaryotic microbiome. Eukaryotic signatures could provide a greater specificity and statistical weight when discerning between body fluids. These microbial markers can be implemented to develop a confirmatory assay for body fluid identification that works in tandem with other DNA based methods in the forensic workflow. Using a VCU approved IRB protocol, samples of urine (n=100), feces (n=72), and saliva (n=77) were collected. DNA was isolated and quantified using Qiagen's QIAamp DNA Investigator and DNA Micro kits. The V9 region of the 18S rDNA was amplified using dual-index strategy and samples were sequenced on the Illumina MiSeq FGx. The sequences were analysed using Mothur (v 1.39.5) and RStudio (v 3.6.3). The relative abundance of eukaryote taxa and an Analysis of Molecular Variance (AMOVA) indicated significant differences in the eukaryotic community structure between all body fluids, except male and female urine. At the family level, feces was characterized by the combined presence of *Saccharomycetaceae* (80.1%). *Malasseziaceae* (37.4% male, 36.6% female) and *Diptera* (22.6% male, 19.4% female) were indicators for urine, while *Poales* (43.3%) and *Asparagales* (10.3%) indicated saliva. The species richness was greater in urine but less in feces and saliva when compared to the species richness of the prokaryotic microbiome. Overall, these results indicate that each body fluid has a unique eukaryotic community and potential for use in body fluid identification, especially as a compliment to the bacterial signatures that have already been developed. The results of this study are a novel addition to previous work, and advance the use of microbial forensics as an alternative method in forensic serology.

Keywords: forensic science, microbiome, 18S rDNA, eukaryote, signature, body fluid identification, serology, next generation sequencing

Introduction

The human microbiome, a complex community of microbial organisms consisting of bacteria, archaea, and eukaryotes, has come to the forefront of medical research in recent years due to the wealth of information that can be gained from studying what thrives in and on our bodies (1, 2). The Human Microbiome Project (HMP), formed in 2007, has furthered the understanding of how the host and the microbes interact and has characterized at least a portion of the communities that make up the microbiome (3). However, the HMP focuses more on the biomedical and physiological applications, rather than how it can be utilized for forensic purposes. Most of the research performed so far has been done with bacteria, by far the most prevalent microorganism in the human host, by targeting the 16S ribosomal RNA (rRNA) gene. Because of the difference in the relative abundance of bacteria versus eukaryotes within the microbiome, there have been fewer studies in the characterization of the eukaryotic communities using the 18S rRNA gene (4).

Overview of Forensic Serology

Forensic serology is the identification, classification, and study of body fluids as they relate to a crime. Identifying the body fluid found on evidence is a crucial part of a forensic investigation, as it guides a forensic analyst to the proper course of action for the evidence sample. Early scientific methods of blood identification date back to the late 1800's, starting with an enzyme test that gave a reaction based on the peroxidase-like activity of hemoglobin (5). Body fluid identification usually begins with a screening to find areas of interest on larger pieces of evidence, typically with a chemical indicator or an alternate light source. A presumptive test is then performed that determines the likelihood that the substance is a particular fluid. Presumptive tests are subject to false positives because they rely on chemical reactions that occur in multiple

fluids or naturally in other substances (6). Finally, if one is available, a confirmatory test can be performed to positively identify a specific body fluid. True confirmatory tests are only available for blood, using either the Takayama crystal test or an antibody test, and semen, by microscopic identification or detection of the prostate specific antigen (PSA) (6). The current serological methods are rapid and low-cost, but are also non-specific and labor intensive. The current tests for saliva, urine and feces are all presumptive and employed in only about half of laboratories worldwide, according to a survey done by Desroches et. al. (2009) (6). As DNA analysis systems and technology have increased in sensitivity, the detection levels for serological techniques of most fluids have remained unchanged.

New Methodologies for Serological Identification

Due to the nature of current serological methods, more research is being conducted in alternative methods of body fluid identification that work in tandem with DNA analysis. Several studies have been done on the use of messenger RNA (mRNA) profiling, where body fluids are identified by the gene expression of the cell types that comprise the specific fluid (7, 8). Messenger RNA profiling employs the use of reverse-transcription polymerase chain reaction (RT-PCR) for detection and is designed to multiplex with capillary electrophoresis (CE) systems (7). Another, similar technique is the expression of microRNAs (miRNAs), which are small sections of non-coding RNA, to identify body fluids. In multiple studies, miRNAs have been shown to be differentially expressed in forensically relevant body fluids (9)(10)(11). Subsequent studies have developed methods to incorporate this technique using RT-PCR and CE systems to generate miRNA expression data specific to each body fluid (12). Both of these methods do have some disadvantages. Messenger RNAs are more prone to degradation because of their longer sequences compared to miRNAs and miRNA quantitation has been proven to vary depending on

the method of detection (13). Until recently, the use of the human microbiome for body fluid identification has remained unexplored.

Microbial Characterization of Body Fluids

Ribosomal RNA genes are sequences of DNA which code for the RNA strands that comprise sections of the subunits in ribosomes. Because ribosomes are needed to manufacture proteins and can be found in all cells, the rRNA genes can be sequenced to determine an organism's taxonomic group (14). Next-generation sequencing of rRNA genes, specifically the 16S subunit in bacteria and the 18S subunit in eukaryotes, is the current and most reliable method for phylogenetic classification of the microbiome (15). The samples are each tagged with a different barcoded primer, amplified through PCR, and parallel sequenced. The sequenced strands are separated by barcode and compared to reference sequences in SILVA, a database comprised of aligned rRNA sequences from all three domains: Bacteria, Archaea, and Eukarya (15). Since the need for further research into the eukaryotic taxa became apparent, the development of universal primers for 18S rRNA genes has been a priority. Studies done by both Hadziavdic et. al. (2014) and Wang et. al. (2014) identified the most suitable and effective primers for use in microbial classification among eukaryotic groups (16, 17). In a study by Amaral-Zettler et. al., the V9 hypervariable region of the eukaryotic rRNA gene was targeted for amplification in order to study Protista diversity in ecologically distinct bodies of water (18). The V9 hypervariable region is used as a target because it is highly conserved among eukaryotes and there are substantial resources for reference comparison (1).

Mammalian Blocking Primers

Specific primers that anneal to common eukaryotic sequences are used to amplify the V9 hypervariable region (1). In addition to these primers, a mammalian blocking primer is used to

obstruct the amplification of the host (i.e. vertebrate) DNA by blocking either the annealing or elongation step during the PCR process. Annealing blockers compete with the normal primers for binding space while elongation blockers bind between amplification sites and prevent elongation. The elongation arrest primers have an additional C3 spacer on the 3' end which inhibits elongation during amplification. The blocking primers have species-specific sequences to ensure that only the unwanted mammalian DNA is blocked (19). These blocking primers can greatly reduce the quantity of human DNA that gets amplified; however they cannot completely eliminate amplification of the host sequences.

Eukaryotic Taxonomy

Taxonomy, which is the classification of the diversity of organisms, is the foundation on which biological knowledge is structured. Traditionally, taxonomy has been based on morphological diversity, but given the complexity of the eukaryotic domain there has been a shift from the classic four kingdoms proposed by Robert Whittaker in 1969 to a new system of six supergroups (20). While the bacterial classification database still has its own issues, it continues to outperform the eukaryotic classification database in terms of accurately identifying taxa, simply because bacteria have been more extensively studied by the scientific community.

The most commonly used reference file for eukaryotic classification, SILVA NR v132, relies on a six-level system (Level 1 to Level 6), which corresponds to the classifications traditionally used at the Domain through Genus levels. This reference uses the UniEuk taxonomic framework for eukaryote sequence alignment and classification. UniEuk is an open-source, universal taxonomic framework that incorporates the collective current findings of the research community (21). As with most databases of its kind, it is not a complete and still produces unclassified sequences. Given the current classification structure of eukaryotic

taxonomy, it is important to note that analysis of sequences with classification at the genus level is not as successful as that of the higher taxonomic levels.

Summary of Previous Work

Because of the diversity and abundance of prokaryotes within the human microbiome and the well-established use of 16S rRNA genes for sequencing, the majority of researchers seeking to identify body fluids by their microbial signature have focused on the domain Bacteria. Specifically for forensic application, one study developed a bacterial signature for saliva by cross-validation against skin samples (22). Work done by Dobay et. al. (2019) determined that forensically relevant body fluids still retained their unique bacterial composition even after environmental exposure (23). Data analysis for 16S rDNA sequencing has been optimized to improve recognition of four different fluid types: oral, nasal, vaginal and fecal (24).

The first phase of this study also focused on prokaryotic diversity; all of the forensically relevant body fluids were sampled, which included blood, saliva, urine, feces, semen, menstrual fluid, and vaginal fluid. Bacterial signatures were developed for all of the fluids, though the female intimate samples (vaginal fluid and menstrual blood) and semen could not be statistically differentiated from each other, as seen in the Dobay et. al. study (23). Additional analyses included limit of detection and mixture studies.

Project Goals

Now that signatures are being developed for the prokaryotic microbiome, the next step is the examination of the eukaryotic microbes of the human body. The eukaryotic signature could provide a greater specificity and statistical weight when discerning between body fluids. The goal of this study is to sequence the V9 hypervariable region of the 18S rRNA gene in saliva, feces and urine in order to identify the eukaryotic taxa present in each body fluid and compare

the microbial communities of each fluid for the development of a statistical model for body fluid identification.

Materials and Methods

Sample Collection and DNA Extraction

All samples were collected for previous research, following a VCU approved IRB protocol (HM20002931). Saliva and feces samples were collected on sterile cotton swabs and stored at room temperature in swab boxes. About half of each swab was used for extraction. Liquid urine samples were stored at -80°C and 1ml of sample was used for extraction. Both the fecal (n=72) and saliva (n=77) samples were extracted using Qiagen's QIAamp® DNA Investigator Kit (Hilden, Germany) on the Qiagen QIAcube® (Hilden, Germany). The samples were eluted in 50µl for feces and 30µl for saliva. The urine (n=100) samples were extracted using the QIAamp® DNA Micro Kit (Qiagen, Hilden, Germany) also on the QIAcube® and eluted into 20µl. Due to sequencing space and barcoding constraints, the number of analysed samples for each body fluid differed (**Table 1**). To ensure consistency and reliability throughout the process, associated extraction reagent blanks and mock communities were analysed and sequenced alongside the serological samples. The male and female urine samples were analysed separately to remain consistent with the 16S bacterial phase of this study.

Bacterial DNA Quantification

All samples were quantified using the Applied Biosystems QuantStudio™ Flex Real-Time PCR System and Software v1.3 (Thermo Fisher Scientific, Waltham, MA) using protocol as described in Seashols-Williams et al. (2018). The ZymoBIOMICS™ Microbial Community DNA Standard (Zymo Research Corp, Irvine, CA) was used as the mock community (positive control) for this study.

18S rDNA Amplification

The target region for eukaryote DNA identification was the hypervariable region V9 of the 18S rRNA gene. The samples were barcoded then amplified on the Applied Biosystems Veriti™ 96-Well Thermal Cycler (Thermo Fisher Scientific, Waltham, MA) using the primers and an adapted protocol described by the Earth Microbiome Project (http://press.igsb.anl.gov/earthmicrobiome/protocols-and-standards/18s/) as well as the dual-index strategy described by Kozich et al. (25). Each primer set consists of an Illumina® adapter sequence so the DNA can attach to the MiSeq® flow cell, an eight nucleotide index sequence and a ten nucleotide pad/linker sequence to increase the melting temperature, a two nucleotide non-complementary linker sequence (CG for forward primer and CA for reverse primer), and a 18S rRNA gene specific forward primer (V8_1391f: 5'- GTACACACCGCCCGTC -3') and reverse primer (V9_EukBr: 5'- TGATCCTTCTGCAGGTTCACCTAC -3') (18, 26). To prevent the host's DNA from being amplified, mammalian blocking primer (Mammal_block_I-short_1391f) was used in an 8:1 ratio of blocking primer to forward/reverse primer for the master mix. The master mix included 12.5 µl of Promega 2X PCR master mix (Promega, Madison, WI), 1 µl (5 pmol/µl) each of forward and reverse primers, and 4 µl (10 pmol/µl) of mammalian blocking primer. The amount of DNA added to each well in 1 µl was dependant on body fluid; 0.3 ng was added for saliva and feces, while 0.02 ng was added for both male and female urine. These inputs were based on the initial 16S bacterial phase of this study using the bacterial qPCR quantification results (27). Nuclease free water was added to give a final total volume of 25 µl.

Gel Extraction and PCR Purification

After amplification, the samples were run on a 1.6% agarose gel using 3µl of PCR product to make sure the amplified DNA was of the expected size. If a sample yielded more than

one PCR product, the target band of 260bp was excised and the DNA purified from the gel using the QIAquick® Gel Extraction Kit (Qiagen, Hilden, Germany) according to the manufacturer's protocol, with the remaining 22µl of PCR product and eluted at 50µl (28). All saliva and feces samples required gel extraction while the urine samples did not. All amplified PCR products were cleaned using the Agencourt® AMPure® PCR Purification Kit (Beckman Coulter, Brea, CA) according to the manufacturer's protocol using 10µl of PCR product for the urine samples or gel extracted product for the saliva and feces samples and eluted in 35µl of elution buffer (29). After purification, the samples were quantified using the Qubit™ dsDNA HS Assay Kit on the Qubit® Fluorometer.

MiSeq® Sequencing

All clean, amplified 18S rRNA gene products were pooled in an equimolar concentration of 1ng/µl. Average inputs were used for reagent blanks and negatives; i.e. if the average input for the feces samples was 2µl, the associated reagent blanks and negative controls would be added to the pool at that volume. The completed library pool was then sequenced on the MiSeq FGX sequencing platform (Illumina Inc., USA) using the MiSeq 2X250 Reagent Kit v2 (Illumina Inc., USA) following the manufacturer's protocol. There were two sequencing runs for this study, including 41 reagent blanks, 2 mock communities, 5 PCR negatives, and a total of 249 body fluid samples.

Data Analysis

The sequence data was analysed in Mothur (v1.39.5), an open source data analysis program (30). A contig file was created to pair the forward and reverse reads and the sequences were then aligned using the SILVA NR v132 reference database. The UCHIME program in Mothur was used to isolate and remove chimera and singleton sequences. A read threshold of

200 reads was set as a baseline; if a sample contained 200 reads or less, it was removed from analysis. The samples were classified and any unwanted taxa were removed to leave only the eukaryotes. This removal included *Archaea*, *Bacteria*, *Vertebrata*, which includes *Mammalia* (the host), and unknown or unclassified sequences. The family *Aspergillaceae* (Level 5), was found to be common in reagent blanks across all body fluids and was also removed. It was concluded that the reagent blank sequences were due to *Aspergillaceae* and not contamination. The taxonomy file was used to construct a relative abundance chart at the family level (Level 5). Taxa that contained less than 100 reads across all body fluids were grouped together as "Rare" taxa. The Log10 of the relative abundance values was used to construct a heat map. Values of 0 were changed to 0.001 for logarithmic transformation. The distribution of highly abundant taxa was used to determine possible predictor taxa for each body fluid. Alpha (α) and beta (β) diversity measurements were assessed in Mothur using operational taxonomic units (OTUs) and phylogenetic approaches. Alpha diversity was evaluated using OTUs at a 5% genetic distance. Beta diversity was tested using analysis of molecular variance (AMOVA), to determine significant differences between fluids, and visualized using principal coordinate analysis (PCoA) using weighted UniFrac distances. PCoA and heat map data were plotted using the ggplot2 package in RStudio v3.6.3 (31).

Results

After lineage removal of the host sequences (i.e. vertebrates) and unclassified sequences and after classification, the total number of samples dropped from 249 to 148 samples with a read count above the threshold. This accounted for a 21% loss in feces samples, a 61% overall loss in urine samples, and a 32.5% loss in saliva samples. The total number of reads for all samples after sequence clean-up was 841,875 reads. There was a large variation in the read count

across samples, ranging from 41,113 to 201 reads, with an overall average of 5,688 reads. Saliva and urine experienced an 87% and 84% loss in sequence reads, respectively, while feces had a 49% loss in reads after sequence clean-up (**Table 2**).

The two mock communities, with 0.3ng and 0.02ng of DNA input, respectively, were expected to contain the yeast species *Saccharomyces cerevisiae* and *Cryptococcus neoformans*. Both mock communities contained the expected sequences and classified at the family level (**Figure 1**). After removal of *Aspergilaceae* and classification, five out of forty-one reagent blanks had read counts above threshold, all belonging to urine (**Figure 2**). The variety and relative abundance of taxa within the reagent blanks was similar to the profiles of the urine samples, indicating possible contamination. Thirteen associated urine samples were removed from analysis, contributing 11% to the overall percent loss of urine samples.

General Sequence Characteristics

The total 841,875 sequences were classified into 27 phyla (Level 2), 53 classes (Level 3), 105 orders (Level 4), 138 families (Level 5), and 205 genera. At all levels, 1.86% of all reads were unclassified and could only be identified as eukaryote. It is important to note here that 2% of all reads classified as *Incertae Sedis*, which is Latin for "uncertain placement". It is a taxonomic group that is essentially a placeholder for taxa whose broader relationships are unknown or not yet defined (32).

Taxonomic Distribution of Sequences

At the family level (Level 5), the most abundant classified reads were different in taxa and relative abundance for each body fluid (**Figures 3 and 4**). For feces, the most abundant taxa were *Saccharomycetaceae* (80.1%), *Malasseziaceae* (3.2%), *Incertae Sedis* (3.1%), *Poales* (2.6%), and *Caryophyllales* (2.0%). The male and female urine samples shared the top two most

abundant taxa but differed afterwards. The male urine was comprised of *Malasseziaceae* (37.4%), *Diptera* (22.6%), *Poales* (9.7%), *Saccharomycetaceae* (4.1%), and *Incertae Sedis* (3.7%), while the most abundant taxa in female urine were *Malasseziaceae* (36.6%), *Diptera* (19.4%), *Debaryomycetaceae* (12.1%), *Trichosporonaceae* (7.0%), and *Mucoraceae* (5.4%). For saliva, *Poales* (43.3%), *Asparagales* (10.3%), *Saccharomycetaceae* (8.9%), *Chromulinales* (7.8%), and *Debaryomycetaceae* (7.6%) had the highest relative abundance. For further visualization purposes, a heat map was created to indicate possible predictor taxa for each body fluid at the family level (Level 5). Relative abundance was Log10 transformed and all taxa with 0% relative abundance were changed to 0.001 for logarithmic transformation. Possible indicator taxa include *Saccharomycetaceae* for feces, *Malasseziaceae* and *Diptera* for urine, and *Poales* and *Asparagales* for saliva.

Eukaryotic Diversity Measurements

The Inverse Simpson Index was used as a measure for α-diversity associated with each body fluid (**Figure 5**). The index values begin at 1, indicating only a single species, and range to a value that would indicate perfect evenness and be equal to the number of present taxa. The average index values for each body fluid were 1.37 for feces, 2.42 for male urine, 1.54 for female urine, and 1.70 for saliva. Male urine showed the greatest amount of species richness and feces the least.

Analysis of molecular variance (AMOVA) based on weighted UniFrac distances (β-diversity) indicated significant differences in eukaryotic community structure between all body fluid types, except male and female urine samples (**Table 2**).

A two-dimensional principal coordinate analysis (PCoA) plot based on weighted UniFrac distances explained 3.18% of variation associated with this data set and showed some clustering of samples for each body fluid group; however distinct clustering and separation between different fluid samples was not observed (**Figure 6**).

Discussion

This study utilized MiSeq sequencing of 18S rDNA to analyse the eukaryotic community structure of feces, urine and saliva. A high quantity of mammalian blocking primer, which is a species-specific oligonucleotide with a 3' modification to prevent elongation of the host DNA during PCR amplification, was used to prevent host DNA amplification (33). However, it did not inhibit all host DNA amplification and in most samples a majority of the sequences were classified as class *Mammalia* which resulted in a 41% sample loss. This was demonstrated particularly in the urine samples, where there was an overall 61% loss in those samples after sequence clean-up, indicating that there may be more human DNA from potential sources such as skin cells than other eukaryotic DNA in that fluid. Percent sample loss in the remaining fluids was about half or less than half that of the urine samples, suggesting that there was less host DNA amplified in saliva and feces.

Another potential contributor to the greater loss of urine samples was the original DNA input of 0.02ng for both male and female samples. This input was carried over from the 16S bacterial phase of this study and based on bacterial quantitation (27). However, due to the relatively reduced presence of eukaryotes versus bacteria in the human microbiome, the low input of 0.02ng compared to the 0.3ng input used for saliva and feces could have prevented amplification of eukaryotic DNA other than that of the host.

Additionally, as the samples were processed, there was a continual dilution of the PCR product from amplification to gel extraction, purification and sequencing. If the sample had a reduced eukaryotic community from the start, some of those sequences may have been lost with each step as the sample was diluted.

Taxonomic Diversity of Body Fluids

The alpha diversity, or species richness, of feces and female urine were less than that of saliva and male urine, though male urine was the only fluid to have a significantly higher average species richness (**Figure 5**). When compared to the alpha diversity of the prokaryotic microbiome in these fluids, there was more species richness of bacteria in feces and saliva than eukaryotes, but less richness of bacteria than eukaryotes in urine (27). This indicates that possible eukaryotic indicator taxa for urine may be of more assistance in distinguishing body fluids in conjunction with bacterial signatures than those of feces and saliva.

Both saliva and urine had more variation in eukaryotic taxa than feces, though each fluid shared at least one taxon in the top most abundant eukaryotes. As all of the studied fluids are a part of the digestive process, it stands to reason that they would all contain some common taxa between them, as demonstrated by the overlap of different body fluid samples in the PCoA plot (**Figure 6**). *Saccharomycetaceae*, a yeast family, was the most abundant in feces, but was also present at lower levels in both saliva and urine. Also known as “sugar fungus” or “brewer’s yeast”, these yeasts are very important for food production and fermentation and have been previously identified as part of the human gut microbiome (34). They are likely abundant in feces to assist with the decomposition of waste. *Poales*, another shared taxon, was the most prevalent in saliva and is a large order of flowering plants that includes grass families, such as rice, wheat and grain, and other plants such as flowering fruits (35). In addition, the second most

abundant taxa in saliva and more unique to the fluid is also an order of plants: *Asparagales*, whose taxa are more related by genetic sequence than by morphological structure. It is unsurprising that plant material, which makes up the majority of the food that humans consume, would be found in high abundance in the fluid that aids in breaking down food for digestion. In both male and female urine, *Malasseziaceae*, a fungal family, was the most abundant taxa, though it is shared between body fluids. *Malassezia* are yeast-like fungi known to be common on human skin that are sustained by fatty acids and can cause skin disease (36). The second most prevalent taxon in urine was *Diptera*, which is an order of flies. This was an unusual result and there is nothing in the current literature to suggest why this occurred so abundantly across urine samples or even in smaller amounts the other fluids.

All three fluids were found to be statistically significantly different from one another by AMOVA pair-wise comparison, with the exception of male and female urine. Unlike their bacterial 16S rDNA counterparts, male and female urine cannot be distinguished from each other by their eukaryotic communities, even though the variety of taxa within male urine was significantly higher than that of the female urine.

Possible Predictor Taxa

After studying the relative abundance chart of each body fluid at the family level (Level 5), it appeared as though each fluid has its own eukaryotic microbial signature (**Figures 3 and 4**). When choosing taxa that might distinguish one fluid from another, it was important to choose taxa that were the most abundant and unique to the particular body fluid. *Saccharomycetaceae* was selected as the indicator taxa for feces, as it was notably more abundant than any other taxon in that fluid. *Malasseziaceae* and *Diptera* were chosen for urine because their combined

abundance was unique to that fluid. *Poales*, for its abundance, and *Asparagales*, for its uniqueness, were chosen as indicators for saliva.

Conclusion

Overall, the significant differences found between feces, urine and saliva indicate that their eukaryotic characterization could be a useful complement to the previously developed bacterial signatures for body fluid identification. These results can be further applied by studying the effects of diluted, compromised or mixed fluid forensic samples on the species diversity found in each fluid. This is a novel step forward in understanding the human eukaryotic microbiome, but the general knowledge in this area is lacking in the scientific community. While the bacterial microbiome has been thoroughly reviewed, more research is needed to better understand and classify eukaryotes so that results produced from studies such as this are refined with greater accuracy. The ultimate purpose of this research is to develop a confirmatory assay combining both the bacterial and eukaryotic microbiomes for body fluid identification that works in tandem with other DNA based methods.

References

1. Dollive S, Peterfreund GL, Sherrill-Mix S, Bittinger K, Sinha R, Hoffmann C, et al. A tool kit for quantifying eukaryotic rRNA gene sequences from human microbiome samples. Genome Biol 2012; 13(7):R60.
2. Wampach L, Heintz-Buschart A, Hogan A, Muller EEL, Narayanasamy S, Laczny CC, et al. Colonization and Succession within the Human Gut Microbiome by Archaea, Bacteria, and Microeukaryotes during the First Year of Life. Front Microbiol 2017; 8:738.
3. Human microbiome. Britannica Online Academic Edition, Encyclopædia Britannica, Inc 2018
4. Popovic A, Parkinson J. Characterization of Eukaryotic Microbiome Using 18S Amplicon Sequencing. In: Beiko RG, Hsiao W, Parkinson J, editors. Microbiome Analysis: Methods and Protocols. New York, NY: Springer New York, 2018; 29–48.
5. Gaensslen RE. Sourcebook in forensic serology, immunology, and biochemistry. Washington, D.C.: U.S. Dept. of Justice, National Institute of Justice, 1983
6. Desroches AN, Buckle JL, Fourney RM. Forensic Biology Evidence Screening Past and Present. Can Soc Forensic Sci J 2009; 42(2):101–20.
7. Juusola J, Ballantyne J. Multiplex mRNA profiling for the identification of body fluids. Forensic Sci Int 2005; 152(1):1–12.
8. Xu Y, Xie J, Cao Y, Zhou H, Ping Y, Chen L, et al. Development of highly sensitive and specific mRNA multiplex system (XCYR1) for forensic human body fluids and tissues identification. PLoS One 2014; 9(7):e100123.
9. Hanson EK, Lubenow H, Ballantyne J. Identification of forensically relevant body fluids using a panel of differentially expressed microRNAs. Anal Biochem 2009; 387(2):303–14.

10. Kulstein G, et al. Automation of DNA and MiRNA Co-extraction for MiRNA-based Identification of Human Body Fluids and Tissues. Electrophoresis 2016; 37(21): 2742–50.

11. Sauer E, et al. Validation of Forensic Body Fluid Identification Based on Empirically Normalized MiRNA Expression Data. Forensic Sci Int Genet 2015; 5(C): e462-e464.

12. van der Meer D, Uchimoto ML, Williams G. Simultaneous Analysis of Micro-RNA and DNA for Determining the Body Fluid Origin of DNA Profiles. J Forensic Sci 2013; 58(4):967–71.

13. Zubakov D, et al. MicroRNA Markers for Forensic Body Fluid Identification Obtained from Microarray Screening and Quantitative RT-PCR Confirmation. Int J of Legal Medicine 2010; 124(3): 217–26.

14. Meyer A, Todt C, Mikkelsen NT, Lieb B. Fast evolving 18S rRNA sequences from Solenogastres (Mollusca) resist standard PCR amplification and give new insights into mollusk substitution rate heterogeneity. BMC Evol Biol 2010; 10(1):70.

15. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res 2007; 35(21).

16. Hadziavdic K, Lekang K, Lanzen A, Jonassen I, Thompson EM, Troedsson C. Characterization of the 18S rRNA gene for designing universal eukaryote specific primers. PLoS One 2014; 9(2): e87624.

17. Wang Y, Tian RM, Gao ZM, Bougouffa S, Qian P-Y. Optimal Eukaryotic 18S and Universal 16S/18S Ribosomal RNA Primers and Their Application in a Study of Symbiosis. PLoS One 2014; 9(3): e90053.

18. Amaral-Zettler LA, McCliment EA, Ducklow HW, Huse SM. A Method for Studying Protistan Diversity Using Massively Parallel Sequencing of V9 Hypervariable Regions of Small-Subunit Ribosomal RNA Genes. PLoS One 2009; 4(7):1–9.

19. Boessenkool, S., Epp, L.S., Haile, J., Bellemain, E., Edwards, M., Coissac, E., et al. Blocking human contaminant DNA during PCR allows amplification of rare mammal species from sedimentary ancient DNA. Molecular Ecology, 2012; 21: 1806-1815.

20. Parfrey, LW, Barbero, E, Lasser E, et al. Evaluating Support for the Current Classification of Eukaryotic Diversity (Evaluating Eukaryotic Classification). PLoS Genet 2006; 2(12): e220.

21. Berney, C, Ciuprina, A, Bender, S, Brodie, J et al. UniEuk: Time to Speak a Common Language in Protistology. J Eukaryotic Microbiol 2017; 3(64): 407–11.

22. Hanssen EN, Avershina E, Rudi K, Gill P, Snipen L. Body fluid prediction from microbial patterns for forensic application. Forensic Sci Int Genet 2017; 30:10–7.

23. Dobay A, Haas C, Fucile G, Downey N, Morrison HG, Kratzer A, et al. Microbiome-based body fluid identification of samples exposed to indoor conditions. Forensic Sci Int Genet 2019; 40:105–13.

24. Hanssen EN, Liland KH, Gill P, Snipen L. Optimizing body fluid recognition from microbial taxonomic profiles. Forensic Sci Int Genet 2018; 37:13–20.

25. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Appl Environ Microbiol 2013; 79(17).

26. Bradley IM, Pinto AJ, Guest JS. Design and Evaluation of Illumina MiSeq-Compatible, 18S rRNA Gene-Specific Primers for Improved Characterization of Mixed Phototrophic Communities. Appl Environ Microbiol 2016; 82(19): 5878–91.

27. Seashols-Williams S, Green R, Wohlfahrt D, et al. An Accurate Bacterial DNA Quantification Assay for HTS Library Preparation of Human Biological Samples. Electrophoresis 2018; 39(21): 2824–32.

28. The QIAquick Gel Extraction Kit and the QIAquick PCR & Gel Cleanup Kit Quick-Start Protocol. 2018 Available at: https://www.qiagen.com/us/resources/resourcedetail?id=a72e2c07-7816-436f-b920-98a0ede5159a&lang=en

29. Agencourt AMPure PCR Purification Instructions for Use. 2013 Available at: https://www.beckmancoulter.com/wsrportal/techdocs?docname=B37419AA.

30. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. Appl Environ Microbiol 2009; 75(23):7537–41.

31. R, D.C.T. R: A Language and Environment for Statistical Computing. (R Foundation for Statistical Computing, http://www.R-project.org, Vienna, Austria. 2011).

32. Incertae Sedis. Bergey's Manual of Systematics of Archaea and Bacteria. American Cancer Society 2015; p. 1

33. Vestheim, H, Simon NJ. Blocking Primers to Enhance PCR Amplification of Rare Sequences in Mixed Samples – a Case Study on Prey DNA in Antarctic Krill Stomachs. Front Zool 2008; 5(1): 12.

34. Cui L, Morris A, et al. The Human Mycobiome in Health and Disease. Genome Medicine 2013; 5(7)

35. Leandro TD, Rodrigues T, et al. Fusoid Cells in the Grass Family Poaceae (Poales): A Developmental Study Reveals Homologies and Suggests New Insights into Their Functional Role in Young Leaves. Annals of Botany 2018; 122(5): 833–48.

36. Thayikkannu, AB, Kindo, AJ, Veeraraghavan, M. Malassezia-Can It Be Ignored? Ind J Dermatol 2015; 60(4): 332–39

Acknowledgements

I would like to thank the VCU Forensic Department and specifically the members of my committee, Dr. Baneshwar Singh, Dr. Sarah Seashols-Williams, and Dr. Paul Brooks for their guidance and assistance with the project and also their patience when we ran into problems. To the members of the Singh Lab, special thanks go to Denise Wohlfahrt, for her continuous assistance in all matters concerning this research and her endurance of our antics in the lab, and also Claire Cartozzo, for answering questions and being an all-around amazing human being. Additional thanks to the graduate students who came before us and paved the way for the completion of this study. I could not have accomplished any of this without the constant, rock-steady support of my family, who encouraged me every step of the way, or my friends that gave me a refuge away from work when I needed it. My sincerest thanks go to my research partner and best friend, Ines Benaissa. We would not have made it through this without each other.

Figures and Tables

Table 1. Summary of analysed samples with demographic data.

	Saliva	Urine	Feces
Gender			
Female	47	57	47
Male	30	43	15
Unreported	0	0	10
Age Group (Years)			
<18	1	0	0
18-30	64	82	58
31-50	8	9	2
>50	2	5	2
Unreported	2	4	10
Ethnicity			
Caucasian	36	49	23
African American	24	28	14
Hispanic	3	8	4
Asian	4	5	10
Mixed	6	4	5
Other/Unreported	4	6	16
Total Samples	77	100	72

Table 2. Summary of the ranges and averages sequence reads for each body fluid before and after quality control measures, i.e., after removal of sequences belonging to host (vertebrate).

Body Fluid	Before		After	
	Range	Average	Range	Average
Feces	3 – 49,141	**16,367**	212 - 29,677	**8,347**
Urine	14 – 80,050	**25,642**	256 - 37,695	**4,094**
Saliva	3 – 75,901	**29,084**	201 - 41,113	**3,971**

Table 3. P values of pair-wise AMOVA comparisons of each body fluid based on weighted UniFrac distances. Experiment-wise error rate = 0.05 and pair-wise error rate (Bonferroni) = 0.00833333. Yellow indicates there is a significant difference between body fluids and blue indicates no significant difference.

	Feces	Saliva	Urine Female
Saliva	0.00002*		
Urine Female	0.00002*	0.00002*	
Urine Male	0.00002*	0.00002*	0.05074
*Indicates significance; Bonferroni pair-wise error rate: 0.00833333			

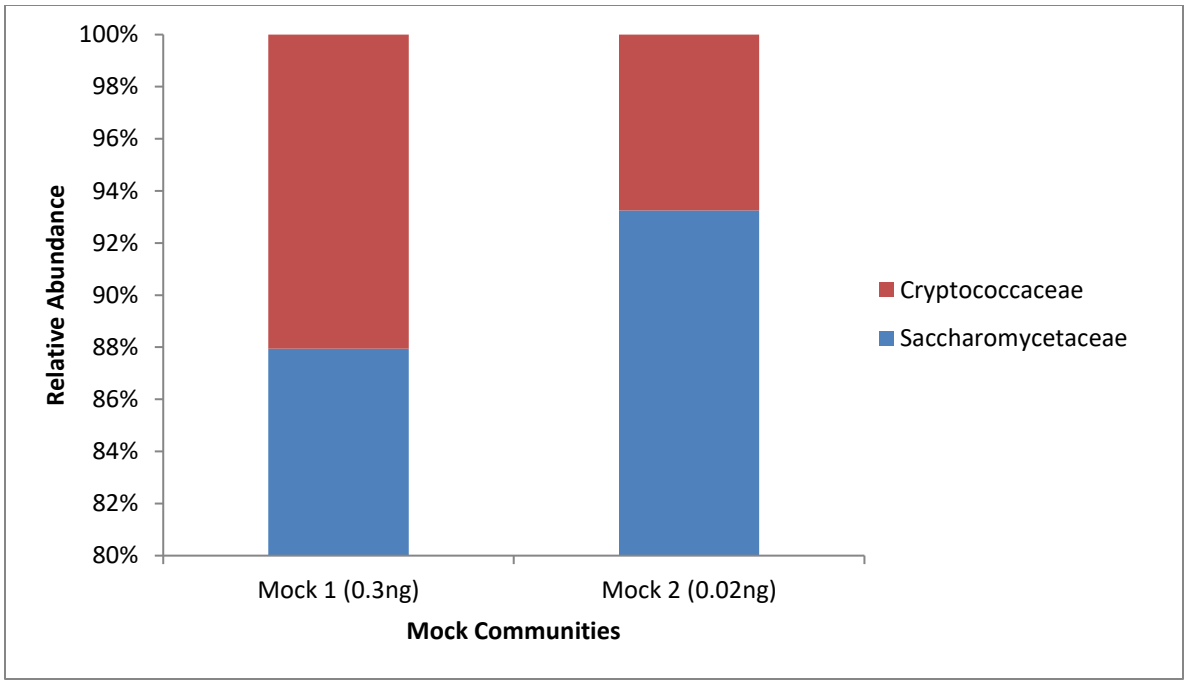

Figure 1. *Relative abundance of taxa at the family level in two mock communities (positive controls).*
Each mock community represented the two different sample DNA input amounts of 0.3ng and 0.02ng respectively. Both controls contained the expected yeast taxa.

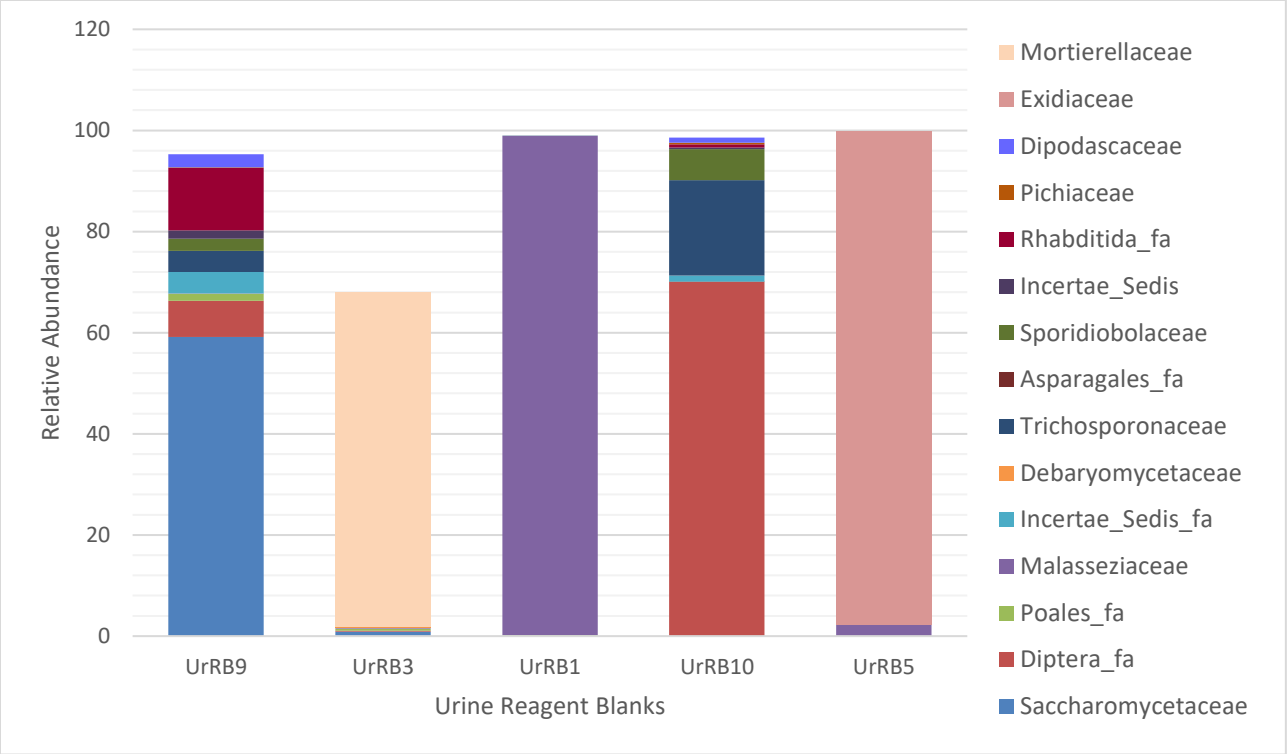

Figure 2. *Relative abundance of taxa at the family level in urine reagent blanks.*
These five reagent blanks contained sequences after quality control measures (i.e. host sequence and *Aspergillaceae* removal). Read counts ranged from 561 to 35,046 reads. The wide variety and relative abundance of taxa in each were similar to the profile of the urine samples. Only the top 15 most abundant taxa were included in this depiction.

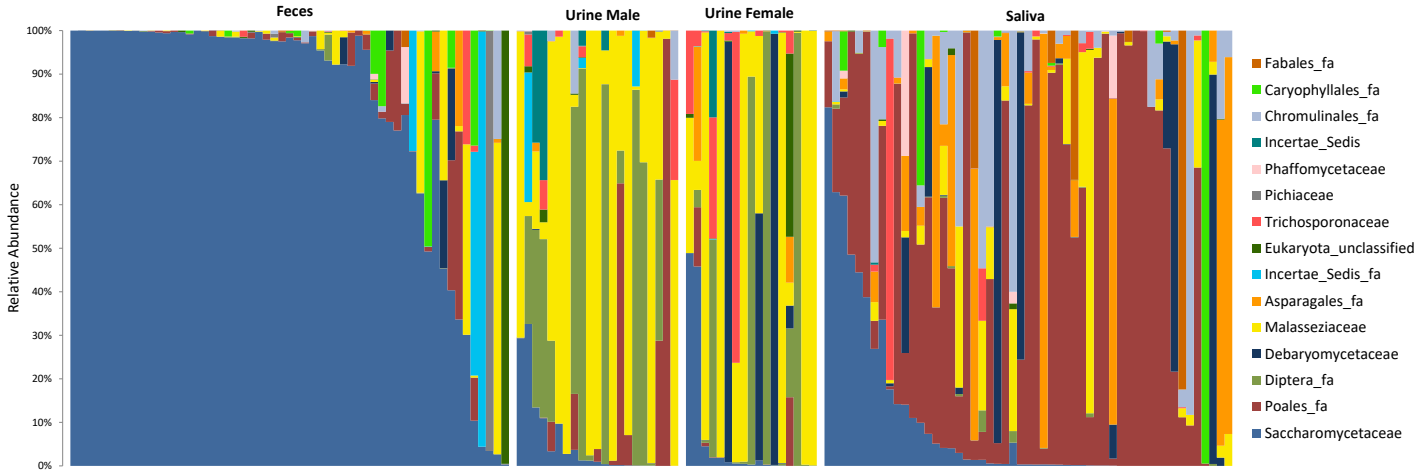

Figure 3. *Relative abundance of classified eukaryotes at family level (level 5).*
Only the top fifteen most abundant taxa for all body fluids were included in this depiction.

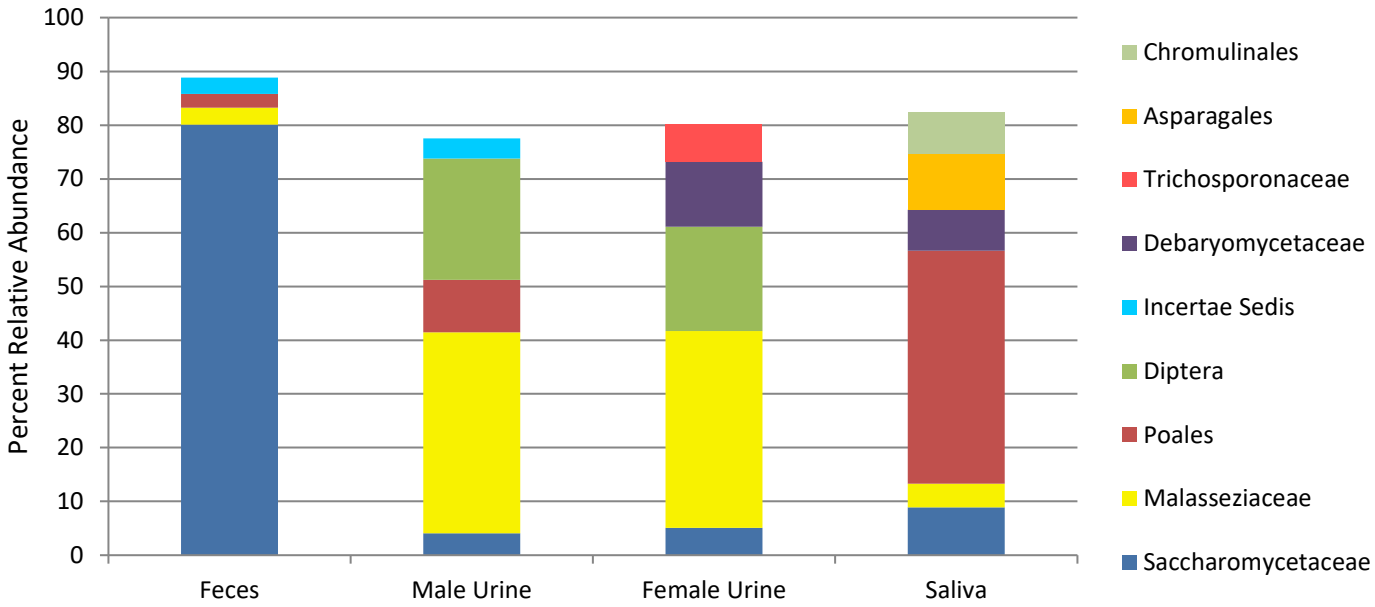

Figure 4. *Relative abundance of the most abundant classified eukaryotes at family level (level 5).* For feces, the most abundant taxa is *Saccharomycetaceae* (80.1%). The male and female urine samples share the top two most abundant taxa, *Malasseziaceae* (37.4% male, 36.6% female) and *Diptera* (22.6% male, 19.4% female). For saliva, *Poales* (43.3%) and *Asparagales* (10.3%) have the highest relative abundance.

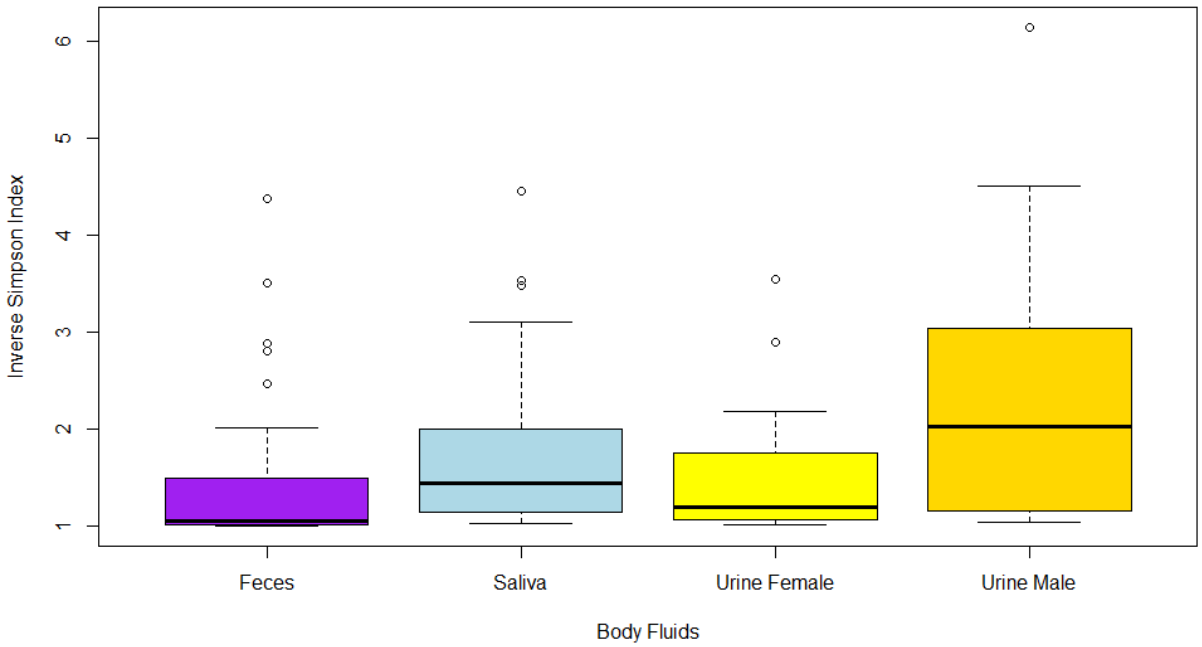

Figure 5. *Inverse Simpson Index body fluid comparisons.*
Alpha diversity of the eukaryotic communities within each body fluid. The black bar within each box denotes the mean (average) index value for that fluid.

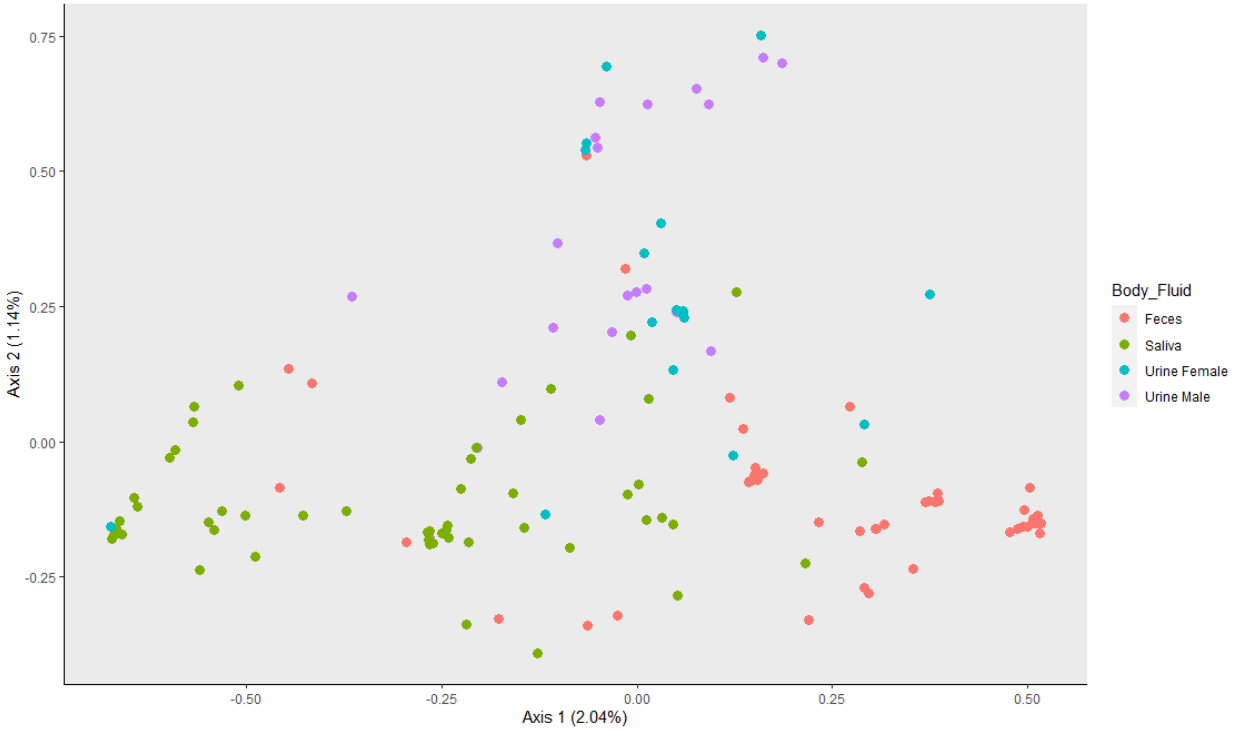

Figure 6. *Principle coordinate analysis using weighted Unifrac distances*
Beta diversity of the eukaryote communities for each body fluid. The first two axes explain 2.04% and 1.14% of the variation in the original dissimilarity matrix, respectively. There is slight clustering of each body fluid group; however they do not make distinct clusters.

APPENDIX I
Mothur Commands & Options

Curation & Classification:

```
make.contigs(file=18S_BfID.txt,trimoverlap=t,insert=30, processors=4)
summary.seqs(fasta=18S_BfID.trim.contigs.fasta)
screen.seqs(fasta=18S_BfID.trim.contigs.fasta,group=18S_BfID.contigs.g
roups, maxambig=0, minlength=50, maxlength=200)
summary.seqs(fasta=18S_BfID.trim.contigs.good.fasta)
unique.seqs(fasta=18S_BfID.trim.contigs.good.fasta)
summary.seqs(fasta=18S_BfID.trim.contigs.good.unique.fasta,
name=18S_BfID.trim.contigs.good.names)
align.seqs(fasta=18S_BfID.trim.contigs.good.unique.fasta,
reference=silva.nr_v132.align, processors=4)
summary.seqs(fasta=18S_BfID.trim.contigs.good.unique.align,
name=18S_BfID.trim.contigs.good.names)
screen.seqs(fasta=18S_BfID.trim.contigs.good.unique.align,
name=18S_BfID.trim.contigs.good.names,
group=18S_BfID.contigs.good.groups,
summary=18S_BfID.trim.contigs.good.unique.summary,start=42554,end=4311
6, maxhomop=8)
summary.seqs(fasta=18S_BfID.trim.contigs.good.unique.good.align,
name=18S_BfID.trim.contigs.good.good.names)
filter.seqs(fasta=18S_BfID.trim.contigs.good.unique.good.align,
vertical=T, trump=.)
unique.seqs(fasta=18S_BfID.trim.contigs.good.unique.good.filter.fasta,
name=18S_BfID.trim.contigs.good.good.names)
summary.seqs(fasta=18S_BfID.trim.contigs.good.unique.good.filter.uniqu
e.fasta, name=18S_BfID.trim.contigs.good.unique.good.filter.names)

pre.cluster(fasta=18S_BfID.trim.contigs.good.unique.good.filter.unique
.fasta, name=18S_BfID.trim.contigs.good.unique.good.filter.names,
group=18S_BfID.contigs.good.good.groups, diffs=1)
summary.seqs(fasta=18S_BfID.trim.contigs.good.unique.good.filter.uniqu
e.precluster.fasta,
name=18S_BfID.trim.contigs.good.unique.good.filter.unique.precluster.n
ames)
chimera.uchime(fasta=18S_BfID.trim.contigs.good.unique.good.filter.uni
que.precluster.fasta,
name=18S_BfID.trim.contigs.good.unique.good.filter.unique.precluster.n
ames, group=18S_BfID.contigs.good.good.groups, dereplicate=t)
remove.seqs(fasta=18S_BfID.trim.contigs.good.unique.good.filter.unique
.precluster.fasta,
accnos=18S_BfID.trim.contigs.good.unique.good.filter.unique.precluster
.denovo.uchime.accnos,
name=18S_BfID.trim.contigs.good.unique.good.filter.unique.precluster.n
ames, group=18S_BfID.contigs.good.good.groups, dups=T)
summary.seqs(fasta=18S_BfID.trim.contigs.good.unique.good.filter.uniqu
e.precluster.pick.fasta,
```

name=18S_BfID.trim.contigs.good.unique.good.filter.unique.precluster.pick.names)

*File names were shortened from this point forward
count.groups(group=18S_BfID.pick.groups)
get.groups(fasta=18S_BfID.pick.fasta, name=18S_BfID.pick.names, group=18S_BfID.pick.groups, accnos=reducedfluids)
count.groups()
classify.seqs(fasta=18S_BfID.pick.pick.fasta,name=18S_BfID.pick.pick.names,group=18S_BfID.pick.pick.groups,template=silva.nr_v132.align, taxonomy=silva.nr_v132.tax, cutoff=80, processors=4)
remove.lineage(fasta=18S_BfID.pick.pick.fasta,name=18S_BfID.pick.pick.names,group=18S_BfID.pick.pick.groups,taxonomy=18S_BfID.pick.pick.nr_v132.wang.taxonomy,taxon=Chloroplast-Mitochondria-unknown-Archaea-Bacteria-unclassified-uncultured-Vertebrata-Aspergillaceae)
count.groups(group=18S_BfID.pick.pick.pick.groups)
get.groups(fasta=18S_BfID.pick.pick.pick.fasta,name=18S_BfID.pick.pick.pick.names,group=18S_BfID.pick.pick.pick.groups,taxonomy=18S_BfID.pick.pick.nr_v132.wang.pick.taxonomy, accnos=reducedfluids2.txt)
classify.seqs(fasta=18S_BfID.pick.pick.pick.pick.fasta,name=18S_BfID.pick.pick.pick.pick.names,group=18S_BfID.pick.pick.pick.pick.groups, template=silva.nr_v132.align,taxonomy=silva.nr_v132.tax,cutoff=80, processors=6)
count.groups(group=18S_BfID.pick.pick.pick.pick.groups)

dist.seqs(fasta=18S_BfID.pick.pick.pick.pick.fasta,cutoff=0.20, countends=F)
cluster(column=18S_BfID.pick.pick.pick.pick.dist, name=18S_BfID.pick.pick.pick.pick.names, method=opti, cutoff=0.20)
make.shared(list=18S_BfID.pick.pick.pick.pick.opti_mcc.list, group=18S_BfID.pick.pick.pick.pick.groups, label=0.05)
classify.otu(list=18S_BfID.pick.pick.pick.pick.opti_mcc.list, name=18S_BfID.pick.pick.pick.pick.names,taxonomy=18S_BfID.pick.pick.pick.pick.nr_v132.wang.taxonomy,label=0.05)

α-diversity measurements:

collect.single(shared=18S_BfID.pick.pick.pick.pick.opti_mcc.shared, calc=chao-ace-invsimpson-shannon-shannoneven, freq=10)
rarefaction.single(shared=18S_BfID.pick.pick.pick.pick.opti_mcc.shared, calc=sobs, freq=10)
summary.single(calc=nseqs-coverage-sobs-invsimpson-shannon-shannoneven, iters=1000)

Construct Phylogenetic Trees (β-diversity):

dist.seqs(fasta=18S_BfID.pick.pick.pick.pick.fasta, output=phylip)
clearcut(phylip=18S_BfID.pick.pick.pick.pick.phylip.dist)
phylo.diversity(tree=18S_BfID.pick.pick.pick.pick.phylip.tre, name=18S_BfID.pick.pick.pick.pick.names,group=18S_BfID.pick.pick.pick.pick.groups, rarefy=T, freq=100)

//UniFrac unweighted//

unifrac.unweighted(tree=18S_BfID.pick.pick.pick.pick.phylip.tre,name=18S_BfID.pick.pick.pick.pick.names,group=18S_BfID.pick.pick.pick.pick.groups, distance=lt, random=F)
tree.shared(phylip=18S_BfID.pick.pick.pick.pick.phylip.tre1.unweighted.phylip.dist)

//UniFrac weighted//

unifrac.weighted(tree=18S_BfID.pick.pick.pick.pick.phylip.tre, name=18S_BfID.pick.pick.pick.pick.names,group=18S_BfID.pick.pick.pick.pick.groups, distance=lt, random=F)
tree.shared(phylip=18S_BfID.pick.pick.pick.pick.phylip.tre1.weighted.phylip.dist)

Tests for (β-diversity):

amova(phylip=18S_BfID.pick.pick.pick.pick.phylip.tre1.weighted.phylip.dist, design=AD_design_file.design, sets=all, iters=50000)
pcoa(phylip=18S_BfID.pick.pick.pick.pick.phylip.tre1.weighted.phylip.dist)
nmds(phylip=18S_BfID.pick.pick.pick.pick.phylip.tre1.weighted.phylip.dist, mindim=3, maxdim=3, iters=5000)

APPENDIX II
R Commands & Options

2-Dimensional PCOA:

```
pcoaw<-read.csv(file.choose(),header=TRUE)
require(ggplot2)
ggplot()+geom_point(aes(x=axis1,y=axis2,col=Body_Fluids),size=3,data=p
coaw) + xlab("Axis 1 (2.04%)") + ylab("Axis 2 (1.14%)") + theme() +
theme(axis.line = element_line(colour = "black"),panel.grid.minor =
element_blank(),panel.grid.major = element_blank())
```

Heat Map:

```
dlevel5<-read.csv(file.choose(),header=TRUE)
library(gplots)
library("RColorBrewer")
row.names (dlevel5) = dlevel5$taxon
dlevel5= dlevel5 [2:149]
dlevel5_matrix = data.matrix (dlevel5)
heatmap.2 (dlevel5_matrix,Rowv=FALSE, Colv=FALSE,
     scale="none",cexRow=0.5,cexCol=0.3,margins=c(8,12),  key=TRUE,
     keysize= 1.0, symkey=FALSE, trace="none",, density.info="none",
     col=brewer.pal(9,"YlGnBu"))
heatmap.2 (dlevel5_matrix,Rowv=TRUE, Colv=FALSE,
     scale="none",cexRow=0.5,cexCol=0.3,margins=c(8,12),  key=TRUE,
     keysize= 1.0, symkey=FALSE, trace="none",, density.info="none",
     col=brewer.pal(9,"YlGnBu"))
```

Inverse Simpson Boxplot:

```
alpha<-read.csv(file.choose(),header=TRUE)
head(alpha)
bf<-(alpha$Body_Fluid)
is<-(alpha$invsimpson)
boxplot(is~bf, main="Alpha Diversity",xlab="Body Fluids",ylab="Inverse
Simpson Index",col=c("purple","lightblue","yellow","gold"))
```

Vita

Alyssa earned her Bachelor of Science degree in Forensic Science with honors from Virginia Commonwealth University in 2015. This was followed by a semester internship at the Center for Forensic Science Research and Education in partnership with Arcadia University. She returned to VCU a few years later and earned a Master of Science in Forensic Science in 2020. During her graduate program, Alyssa attended the Future Trends in Forensic DNA Technology Seminar Series in 2019. She also worked as a graduate research associate in the VCU Forensic Science Department, researching eukaryotic signatures for body fluid identification. She presented this research at the American Academy of Forensic Science 72nd Annual Meeting in February of 2020.