COVER PAGE

National Institute of Justice
FY 2016 Research and Development in Forensic Science for Criminal Justice Purposes
Award 2016-DN-BX-0151
Forensic body fluid identification using microbiome signature attribution

Baneshwar Singh, PhD
Assistant Professor, Department of Forensic Science
Virginia Commonwealth University
(804) 828-9576
bsingh@vcu.edu

Andrea Publow, MFA, CRA
Director of Sponsored Programs – Gov't/Non-Profit Support
Virginia Commonwealth University
(804) 828-6772
ospgold@vcu.edu

Final Draft Summary Report
Submission Date: 09/30/2019

Virginia Commonwealth University
P.O Box 980568
800 East Leigh Street, Suite 3200
Richmond, VA 23298-0568

VCU FP No. 1617
VCU Index No. 5-44934; 5-44949
Project/Grant Period: 1/1/2017 – 12/31/2019 (extended, no-cost)

Abstract:

Detection and identification of body fluids plays a crucial role in criminal investigation, as it provides information on the source of the DNA as well as corroborative evidence regarding the crime committed, scene, and/or association with persons of interest. Historically, casework methods in serology have been chemical, immunological, catalytic, spectroscopic, and/or microscopic in nature. However, most of these methods are presumptive, with few absolutely robust confirmatory exceptions. More importantly, these methods can only be used to identify a single biological fluid, and thus elimination of the presence of other biological fluids and mixtures in forensic samples can be problematic. In recent years several new molecular methods (mRNA, miRNA, DNA methylation etc.) have been proposed; although promising, these methods require high quality human DNA or RNA. Additional steps are required for RNA extraction in RNA based methods. Additionally, RNA based methods cannot be used for old cases where only human DNA is available as evidence. In this study, a novel non-human DNA (microbiome) based method was developed for the identification of the major forensically relevant human biological fluids. Eleven hundred and sixty (n=1160) biological samples (semen, vaginal secretions, menstrual secretions, saliva, feces, urine, and venous blood) were collected and preserved using methods commonly used in forensic laboratories for evidence collection. Except urine, DNA from all other samples were extracted using the QIAamp DNA Investigator Kit standard forensic casework sample protocol on the QIAcube robotic workstation. DNA from urine was extracted using the QIAamp DNA Micro Kit. Variable region four (V4) of 16S ribosomal DNA (16S rDNA) was amplified using a dual-index strategy and then sequenced on the MiSeq FGx sequencing platform using the MiSeq Reagent Kit v2 (500 cycles) and following the manufacturer's protocol. Sequence data was analyzed using mothur version 1.39.1 and R version 3.4.0. A novel "*support vector machine*" technique was implemented in R to estimate the accuracy. Results show that all major body fluid samples can be distinguished and identified with an overall accuracy of 89.9% using this technique. Vaginal and menstrual secretions were indistinguishable from each other, and thus were classified together as "female intimate samples". Developmental validation of this method indicated that a reliable microbial profile could be obtained from sample that has bacterial DNA quantity of at least 5pg. The newly developed method is also robust to many common environmental contaminants. In conclusion, the proposed method is fast (no additional steps are needed and one test can identify all major body fluids), accurate, and can be easily added into a forensic high throughput sequencing (HTS) panel.

Introduction

The accurate confirmation of forensically relevant body fluids commonly found at crime scenes remains an important factor within the forensic science community, as it allows for crime scene reconstruction and the resolution of mixture samples from sexual assault cases[1,2]. The ability to differentiate between venous and menstrual blood, or semen and saliva within a stain can be extremely significant to the investigation and prosecution of a case[2,3].

A variety of conventional serological detection and testing methods are used in current casework; however, they are limited in that they typically exploit the catalytic properties of certain proteins within the biological fluid, such as P-TMB testing for hemoglobin or immunochromatographic tests for prostate specific antigens within semen[3]. Unfortunately, these protein-based assays suffer from several limitations that include sample consumption, cost, and labor intensiveness, but most of all, a lack of specificity[3]. Moreover, no confirmatory techniques are commonly used in forensic science casework for the presence of frequently encountered body fluids such as vaginal and menstrual secretions[3,4]. Therefore, one challenge that remains to be resolved is the establishment of reliable biomarkers for rapid and accurate identification of forensically relevant body fluids.

Over the past 15 years, more discriminatory methods for body fluid identification have been proposed and evaluated, including Raman spectroscopy, messenger RNA (mRNA), methylation and microRNA expression[5-19]. Of the methods evaluated thus far, mRNA analysis for the profiling and multiplexing of unique, tissue-specific gene transcripts within body fluids has gained the most traction through evaluation of reproducibility and robustness, sensitivity and specificity. This method has been the subject of collaborative exercises within the European community for implementation into casework[8], but unfortunately, the perception of the poor stability as well as the added steps of RNA-specific or co-extraction protocols, followed by reverse transcription and separate analysis procedures, have produced a lack of enthusiasm for implementation in the US forensic science community thus far.

From these 15 years of forensic research in molecular methods for body fluid identification, we have learned a significant lesson. For a novel molecular body fluid identification method to receive widespread acceptance and implementation in general casework, the method must meet several requirements. As would be expected of any forensic method, it must be robust, reliable, sensitive, and specific for the body fluid in question. It should also be easily added into the forensic workflow with little disruption or additional instrumentation and minimal additional procedures. With the coming advent of high-throughput sequencing (HTS) panels for human identity and phenotype characterization, a DNA-based method for the body fluid identification would be the logical solution.

Development of high-throughput sequencing method during last decade has revolutionized the field of metagenomics. As a result, now we know that 9 out of 10 cells in or on our body are of microbial origin. Traditionally, human milk, blood and urine has been considered as “sterile” in healthy individuals, however recent studies have indicated that these body fluids do contain bacteria, but in relatively low abundances[20-23]. The Human microbiome project and other similar projects have reported distinct microbiome signature in different body fluids, especially in saliva, vagina, semen, and fecal matters, where relative abundance of bacteria were much higher than the actual human cells, and microbiome signature associated with each body fluid was distinct, stable, and predictable[24-27].

These studies on microbiome associated with body fluids, although informative and indicative of microbiome signature associated with each body fluid, cannot be used for the

development of a new tool for forensic body fluid identification because 1) the purpose of the these studies were not forensic, and hence samples collected by these studies were not reflective of real forensic casework samples; 2) sample size in many studies were too small to account for the known variations that exist in each body fluid with ethnicity, sex, and age; and 3) only one or two body fluid samples were considered in any given single study, and hence a comparative study for all body fluids was not performed[4,22,23,26,27].

This study explored and identified groups of bacteria that are specific to each body fluids (saliva, blood, male urine, female urine, feces, semen, menstrual and vaginal secretions) irrespective of age, sex, and race, and utilized this information for the development of statistical models for prediction of each body fluid in a forensic setting. The technique was then validated using low quantity, mixed and compromised samples, and standard operating procedures for collection, analysis, and interpretation of results was generated. We demonstrated that bacterial communities present in each body fluid differ significantly from each other and can be utilized for the identification of each body fluid.

Objectives:

The overall objective of this project was to: 1) characterize bacterial community structure associated with seven human body fluids and develop a standard operating procedure for sampling, processing, and analyzing bacterial samples from body fluids, 2) perform developmental validation of the microbiome based method for body fluid identification, and 3) perform final validation of the protocol using a blind study, and estimation of error associated with this method. To achieve the goal noted above, the authors utilized 16S rDNA high-throughput sequencing (HTS) to characterize bacterial structure associated with more than 1160 samples belonging to seven forensically relevant biological fluids. Developmental validation was performed by characterizing bacteria associated with compromised samples (e.g., mixtures of two or more body fluid samples, environmentally exposed or chemically treated samples, and low quantity samples). These informations were used to develop statistical models for accurate identification of human biological samples in cross validation and blind test (n=175).

Method:

Sample Collection: More than 1160 samples belonging to venous blood, semen, saliva, vaginal and menstrual secretions, urine and feces were collected (**Table 1**) under a forensic sample registry with an approved human subjects protocol through the VCU Department of Forensic Science (VCU HM20002931). In order to protect the rights of all participants, no personal identifiable information was collected from volunteers and all body fluids collected were assigned a random identification number that was not traceable back to any of the participants. Detail sample collection procedure is available in standard operating procedure (SOP). Briefly, samples were collected using method very similar to what is commonly used in a crime lab, and is summarized below:

Blood: Using a finger lancet, donor's finger was pricked using a sterile collection technique and then blood samples were collected on a two sterile cotton swabs. Before finger pricking, donor's finger was wiped with a sterile wipe and the cotton swabs were collected from fingers of 20 donors as a "blood control". This was done to understand bacterial structure in blood that may have originated from skin. Swabs were dried and stored at room temperature.

Saliva: Two sterile cotton swabs were collected from donor's mouth. Donors were instructed to roll these swabs for 30-60 seconds inside their cheeks for salivary transfer. Swabs were dried and stored at room temperature.
Urine: Urine was collected in a sterile urine collection container by the donor and then 500 uL of collected urine was aliquoted in 2 mL microcentrifuge tubes, and were stored at -80^{0}C freezer.
Seminal Fluid: Semen was collected in a sterile container by the donor, and then 200uL of semen was added to sterile cotton swab. Swab was dried and stored at room temperature.
Feces: Donor's defecated on sterile cotton swab directly. Swabs were dried and stored at room temperature.
Vaginal Fluid and Menstrual Blood: Sterile cotton swabs were inserted only 2-3 inches into the vagina and the swabs were twisted while in the vagina for full coverage. Swabs were dried and stored at room temperature.
For developmental validation (objective 2), limit of detection was tested on two samples from each body fluids (n=14). For feces, saliva, menstrual secretion and vaginal secretions, input bacterial DNA ranged from 0.3 ng to 1pg, whereas for semen, venous blood, and urine, input bacterial DNA ranged from 0.02 ng to 1pg. Similarly, impact of physical and chemical treatments was tested on two samples from each sex per body fluid (n=22). In this developmental validation, samples were incubated at 37°C (24 hours and 96 hours) and 95°C (24 hours and 96 hours). Samples were also treated with bleach, soap and UV light, before DNA extraction step. For mixture test, two unique mixtures with seven mixture combinations (blood/saliva, vaginal secretions/semen, vaginal secretions/menstrual secretions/semen, vaginal secretions/venous blood, vaginal secretions/venous blood/semen, semen/saliva, and semen/feces) were created (n=14). Mixture samples were created by proportionally mixing different body fluids supplied by donors during sample collection. The tips of the swabs containing the different body fluids were cut into the same tube for the different mixtures except for venous blood and semen swabs where the whole swab or stain was cut into the tube.

DNA Extraction: DNA from feces, saliva, semen, vaginal fluid, menstrual blood and venous blood was isolated using QIAamp® DNA Investigator Kit (Qiagen™, Hilden, Germany) [16] on the QIAcube liquid handling robot (Qiagen™, Hilden, Germany) using the casework protocol [17]. For semen samples, 20 uL of DTT was added to tube before placement of sample. DNA from urine sample was isolated using QIAamp® DNA Micro Kit (Qiagen™, Hilden, Germany) following manufacturer's protocol. Except urine and venous blood, final elution volume for all samples was 30 uL. Urine and venous blood samples were eluted in 20 uL final volume. All DNA extracts were stored at -20°C. During each set of DNA extraction in QIAcube liquid handling robot, one reagent blank was used as a control for DNA extraction reagents. DNA from feces samples were also isolated with QIAamp® PowerFecal DNA Kit (Qiagen™, Hilden, Germany).

Bacterial and Human DNA Quantification:Bacterial DNA from all extracts were quantified using a quantitative PCR (qPCR) based method as described in Seashols-Williams et al. (2018)[28]. Briefly, a standard curve was generated by creating serial dilutions from a 9.6 ng/µL stock solution of ZymoBIOMICS Microbial Community DNA Standard (Zymo Research; Irvine, CA) down to 0.01 ng/ µL, in duplicate. The master mix containing 6.25µL of PerfeCTa SYBR Green SuperMix (2X) (Quantabio™, Beverly, MA, USA), 0.25µl of V4_515 forward primer sequence (10uM) (5'-

AATGATACGGCGACCACCGAGATCTACACGATCGTGTTATGGTAATTGTGTGYCAG CMGCCGCGGTAA), 0.25µL of V4_806 reverse primer sequence (10uM) (5'-CAAGCAGAAGACGGCATACGAGATAACTCTCGAGTCAGTCAGCCGGACTACNVGG GTWTCTAAT), and 3.75µL nuclease-free water was prepared for the number of reactions that was run. For every reaction, 10.5 µL of master mix along with 2 µL of DNA template (sample) was loaded into its respective well on a 96 -well plate. Samples were amplified using following qPCR protocol: Stage 1: 94°C for 3 minutes; Stage 2: 35 PCR cycles of: 94°C for 45 seconds, 50°C for 60 seconds, 72°C for 90 seconds; Stage 3: 72°C for 10 min. Data Analysis was performed using QuantStudio Real-Time software version 1.3. Negative results were classified as those with undetermined Ct values and positive results were classified as those where Ct value were present. Human DNA was quantified using Quantifiler™ Trio DNA Quantification Kit (Life Technologies) using half reactions and by following manufacturer's protocol. qPCR analysis was conducted using HID Real-Time PCR Analysis Software (Life Technologies), and data was analyzed at a threshold of 0.2 with baseline parameters between 3 and 15 cycles. Negative amplification controls were included on the plate.

16SrDNA MiSeq Sequencing-by-Synthesis: 16S rDNA was amplified and sequenced on MiSeq sequencing platform using primers and protocols as described by Kozich et al. (2013)[29] with some modifications (see below). Briefly, each reaction consisted of 10µL Promega 2X Master Mix (Promega; Madison, Wisconsin), 1.5µL of both the forward (V4_515F 5'AATGATACGGCGACCACCGAGATCTACACXXXXXXXXTATGGTAATTGTGTGYCA G CMGCCGCGGTAA- 3' and reverse primers (V4_806R 5'-CAAGCAGAAGACGGCATACGAGATXXXXXXXXAGTCAGTCAGCCGGACTACNVGG GTWTCTAAT-3') at a 10 µM concentration, Promega nuclease-free water, and DNA extract (sample) to equal a total volume of 20 µL. Use of barcode sequences (as indicated by XXXXXXXX) in each primer helped in multiplexing of several samples (up to 384) in a single lane of MiSeq flow cell, which ultimately reduce per sample sequencing cost. All samples except semen, venous blood and male urine, had a DNA input of 0.3 ng and were amplified on the Applied Biosystem Veriti™ 96-Well Thermal Cycler (Thermo Fisher Scientific, Waltham, MA USA. For semen, venous blood and male urine samples, 0.02 ng of bacterial DNA was used as input template DNA. For venous blood, final PCR volume and $MgCl_2$ concentration were increased from 20 uL to 25 uL and 1.5mM to 2.5mM, respectively. For each 96 well PCR plate, 1 well was used as a negative template control (NTC), 1 well used as a mock community control ZymoBIOMICS Microbial Community DNA Standard (Zymo Research; Irvine, CA) and the remaining 94 wells for samples. Use of the mock community control, which is pooled DNA of 8 known bacterial strains, helped us in assessing sequence error rate during sequence analysis steps[30]. After PCR amplification the samples were stored at -20°C until further use. Ten to twenty microliter of amplified PCR products were cleaned using the Agencourt AMPure XP PCR purification kit (Beckman Coulter, USA), and then purified amplicons were quantified using using Qubit® dsDNA HS Assay Kit on the Qubit® flourometer (Thermo Fisher Inc., USA). The purified amplicons will then be pooled in equimolar concentration.

Sequence Data Analysis: Sequence data generated from these sequencing runs were analyzed following the MiSeq SOP[29] (http://www.mothur.org/wiki/MiSeq_SOP) on Mothur v 1.39.5[30]. First, each set of reads (forward and reverse) were combined using command *make.contig* and default options in Mothur v 1.39.5[30]. Contigs with ambiguous base or unusually long/short read

length were deleted from further analysis. All sequence will be checked for chimera formation using program Uchime[31] as implemented in Mothur v 1.39.5[30], and using the most abundant sequence as reference data. Suspected chimeras were deleted and rest all sequences were utilized for further analyses. Sequences from mock community samples were compared with reference mock community sequence file. This quality control step was performed to avoid spurious operational taxonomic unit (OTU) associated with each data set and also to make sure that data generated from each MiSeq run was consistent. Hierarchical classification of 16S rDNA sequences was performed using Naïve Bayesian rRNA classifier[32] as implemented in Mothur v 1.39.5[30]. The Mothur-formatted version of Greengene reference data (http://www.mothur.org/wiki/MiSeq_SOP) was used as a reference file for bacterial hierarchical classification. Only sequences that had ≥ 80% bootstrap support was considered classified at a particular hierarchical level. Operational taxonomic unit (OTU) at 5% genetic distance was used for calculation of α- (e.g., Shannon, inverse-Simpson) and β- (Bray Curtis) diversity indices in Mothur v 1.39.5[30]. To avoid spurious OTU count because of variation in sequence reads in each sample, sequences were subsampled at 5096 sequence reads from each sample, before α- and β-diversity estimations and hierarchical classification. Bray Curtis distance matrix was used for principle coordinate analysis (PCoA), analysis of molecular variance (AMOVA), and indicator species analyses (ISA). For body fluid prediction and to determine accuracy a novel "*support vector machine*" technique was implemented in R.

Result and Discussion:

Bacterial DNA Quantification and Bacterial Diversity: Before this study majority of microbiome research utilized total DNA associated with each sample for 16S rDNA amplification and HTS. To utilize this method in a crime lab, it was very important that the analyst know how much bacterial DNA is needed to get a profile that is consistent between runs and laboratories. So we developed a quantitative PCR based method for bacterial DNA and compared quantity of bacterial and human DNA in each sample type[28] **(Table 2)**. Fecal sample and semen sample had highest and lowest average bacterial DNA yield, respectively. Except feces and urine, all other samples had more human DNA than bacterial DNA. Impact of bacterial DNA input (0.3ng, 0.07ng, 0.02ng) on bacterial profile after HTS was also tested and it was observed that except three bacteria (e.g., *Corynebacterium, Prevotella,* and *Anaerococcus*, bacterial DNA input had little impact of major bacterial taxa[28] (**Figure 1**). As expected, feces, saliva, menstrual blood, vaginal secretion, and female urine had much higher bacterial DNA quantity than semen, venous blood, and male urine samples **(Figure 2)**. More than 50% of venous blood samples had bacterial DNA quantity of ≤ 0.4 ng (bacterial DNA quantity ranged from 0.6 ng to 0.02 ng). This is not unexpected result from this body fluid samples. Although semen and venous blood had very low bacterial DNA yields, 16S rDNA sequencing revealed a more diverse bacterial community in these samples than in many body fluid samples with a higher bacterial DNA yield **(Figure 3)**.

Bacterial Community Composition: Bacterial community structure differed significantly between body fluids, except for vaginal fluid, menstrual secretions and female urine samples **(Figure 4)**. Female intimate samples had relatively high abundance of *Lactobacillus* than all other body fluids. Bacterial structure associated with male and female urine samples differed significantly and hence can be utilized for gender determination. In general, *Lactobacillus* and

Streptococcus were the most dominant bacterial genera associated with female and male urine samples, respectively **(Figure 5)**. A novel *support vector machine* classification method was developed to identify biological fluid with known error rate. This newly developed method was able to accurately predict all major body fluid samples with an overall accuracy of 89.9%. Vaginal and menstrual secretions were indistinguishable from each other in this method and hence were classified together as "female intimate samples" (**Table 3**). In general, precision was higher at lower taxonomic level (i.e., genus) than at higher taxonomic level for all biological fluid. We are still refining this classification system and in the final report we will have well refined robust classification method for all biological samples with known error rate.

Limit of Detection: In general, all samples that had a DNA input of 0.005 ng (and higher) successfully amplified 16S rRNA gene (**Table 4; Figure 6**) and generated bacterial profile that was consistent with sample with other DNA inputs (**Figures 7 & 8**). The results demonstrated that the lowest DNA input that can be adequately used for sequencing across all body fluids was 0.005 ng, and the number of overlapping genera was relatively high across each body fluid at specified level of taxonomy (**Figures 9 & 10**). The ability to use smaller DNA input values into the HTS technologies would allow the forensic analysts in utilizing this method even in situations when samples are highly diluted/degraded.

Impact of Physical and Chemical Treatments: In general impact of various physical and chemical treatment was not consistent across all body fluids (**Figure 11**). Except menstrual blood and seminal fluid, bacterial DNA yield in bleach treated sample was negligible. However its impact on bacterial structure was only visible in saliva samples (**Figure 12)** at the phylum level. Urine and seminal fluid samples that were incubated at 95°C had negligible bacterial DNA yield, and same was true for soap treated fecal samples. However this didn't impact bacterial structure associated with these samples at the phylum level (**Figure 12**).

Impact of Mixture Samples: Although mixture sample had several bacteria similar to the original source samples, many bacteria associated with the original sample were missing as well. This was mainly because in the mixture, the biological fluid that have relative high abundance of bacteria contributed more and hence bacterial profile associated with these samples were more prominent in the mixture sample as well **(Figure 13)**. This data set is still under analysis. Final report will have more information on impact of mixtures on bacterial profile.

Conclusions & Future Implications: Overall this works provides a first extensive study on microbial structure associated with all major human body fluids in a forensic setting. Methods for sample collection and DNA extraction used in this project were almost similar to what is expected in a crime lab. A qPCR-based approach for quantification of microbial DNA from body fluid samples was developed as a part of this project. In future, this will have huge impact on criminal justice system, because this will allow addition of uniform microbial DNA in 16S rDNA sequencing, and hence will minimize any lab-to-lab variations in microbial structure determination from human body fluids and, in other forensic microbiology applications. A novel statistical classification method was developed for prediction of all seven body fluid samples with approximately 90% accuracy in a single test. It was not possible to differentiate vaginal fluid from menstrual fluid using this method (because bacterial structures associated with these two body fluid samples are very similar). However, it was possible to differentiate male and

female urine sample with very high precision. It was also demonstrated that the method will work greatly even when bacterial DNA input is just 5pg (diluted and degraded samples) and for samples that have been deteriorated by physical and chemical exposures (Bleach/UV/Soap/extreme temperature etc.). Mixture of two or more sample didn't result in profile similar to both source samples. Ongoing work on blind validation will further support and validate accuracy of newly developed method. We are in communication with Verogen Inc., USA to transfer our method for its commercial application but before it can be utilized commercially, we still need to perform several developmental validation of this method. Most importantly, we need to perform validation with respect to impact of substrates, storage temperatures, and storage period on bacterial profile. We also need to increase number of replication for mixed samples to account for variation in microbial structure associated with each sample. We also need to test how microbiome associated with human body fluid differ when we compared it with body fluid from other sources.

Tables and Figures:

Table 1: Donor's race, sex and age profile.

	Saliva	Feces	Urine	Semen	Blood	Menstrual Blood	Vaginal Secretions
Gender							
Female	138	137	140	0	146	114	168
Male	67	63	64	58	65	0	0
Age Group (years)							
<18	22	19	34	2	19	9	18
18-30	155	150	153	41	173	92	134
31-50	14	23	9	6	12	7	8
>50	9	4	4	5	3	0	3
Unreported	5	4	4	4	4	6	5
Ethnicity							
Caucasian	108	85	82	29	64	50	69
African American	50	51	60	16	54	22	37
Hispanic	12	19	17	3	15	11	13
Asian	21	30	19	5	40	22	28
Mixed	11	11	15	3	13	4	10
Other/Unreported	8	4	11	1	10	5	11
Total Samples	**205**	**200**	**204**	**58**	**211**	**114**	**168**

Table 2: Microbial:Human DNA Ratios from Forensically Relevant Body Fluid Samples

	Avg Microbial Yield (ng)	Avg Human Yield (ng)	Ratio (Microbial:Human)
Saliva	40.86 ± 0.87	1919.64 ± 35.81	1:47
Blood	0.63 ± 0.18	31.44 ± 0.80	1:50
Feces	340.28 ± 2.66	8.69 ± 0.17	40:1
Urine	18.15 ± 1.51	2.22 ± 0.18	8:1
Semen	0.33 ± 0.0078	1897.08 ± 36.21	1:5749
Vaginal Secretions	83.10 ± 1.62	6134.94 ± 124.17	1:74
Menstrual Secretions	48.30 ± 0.81	3164.28 ± 59.07	1:66

(n=5 saliva, blood, urine, semen, vaginal and menstrual samples. n=4 fecal samples)

Table 3. Average Precision and Recall (Sensitivity) over 100 cross validations by taxonomic levels. Bl=Venous blood, Fe=Feces, Sa= Saliva, Se= Semen, UF= Urine (Female), UM= Urine (Male), VF = Female intimate sample i.e., Vaginal Fluid/Menstrual Blood.

Precision

	Bl	Fe	Sa	Se	UF	UM	VF
Genus	0.867	0.988	0.994	0.883	0.809	0.957	0.864
Family	0.880	0.991	0.987	0.812	0.755	0.902	0.852
Order	0.884	0.992	0.988	0.833	0.681	0.699	0.835
Class	0.888	0.975	0.945	0.678	0.543	0.599	0.771
Phylum	0.872	0.849	0.901	0.322	0.481	0.201	0.739

Recall (Sensitivity)

	Bl	Fe	Sa	Se	UF	UM	VF
Genus	1.000	0.996	1.000	0.403	0.706	0.262	0.967
Family	0.993	0.980	1.000	0.432	0.703	0.320	0.931
Order	0.989	0.993	1.000	0.483	0.630	0.216	0.916
Class	0.981	0.989	0.970	0.409	0.464	0.124	0.876
Phylum	0.936	0.895	0.956	0.207	0.393	0.084	0.815

Table 4: Table showing minimum recommended bacterial DNA input successful and consistent result in microbial signature based body fluid identification method.

Body Fluid	**Control DNA Input (ng)**	**Minimum DNA Input (ng)**
Feces	0.3	0.005
Saliva	0.3	0.001
Blood	0.02	0.005
Urine	0.02	0.001
Semen	0.02	0.005
Vaginal Fluid	0.3	0.001
Menstrual Blood	0.3	0.001

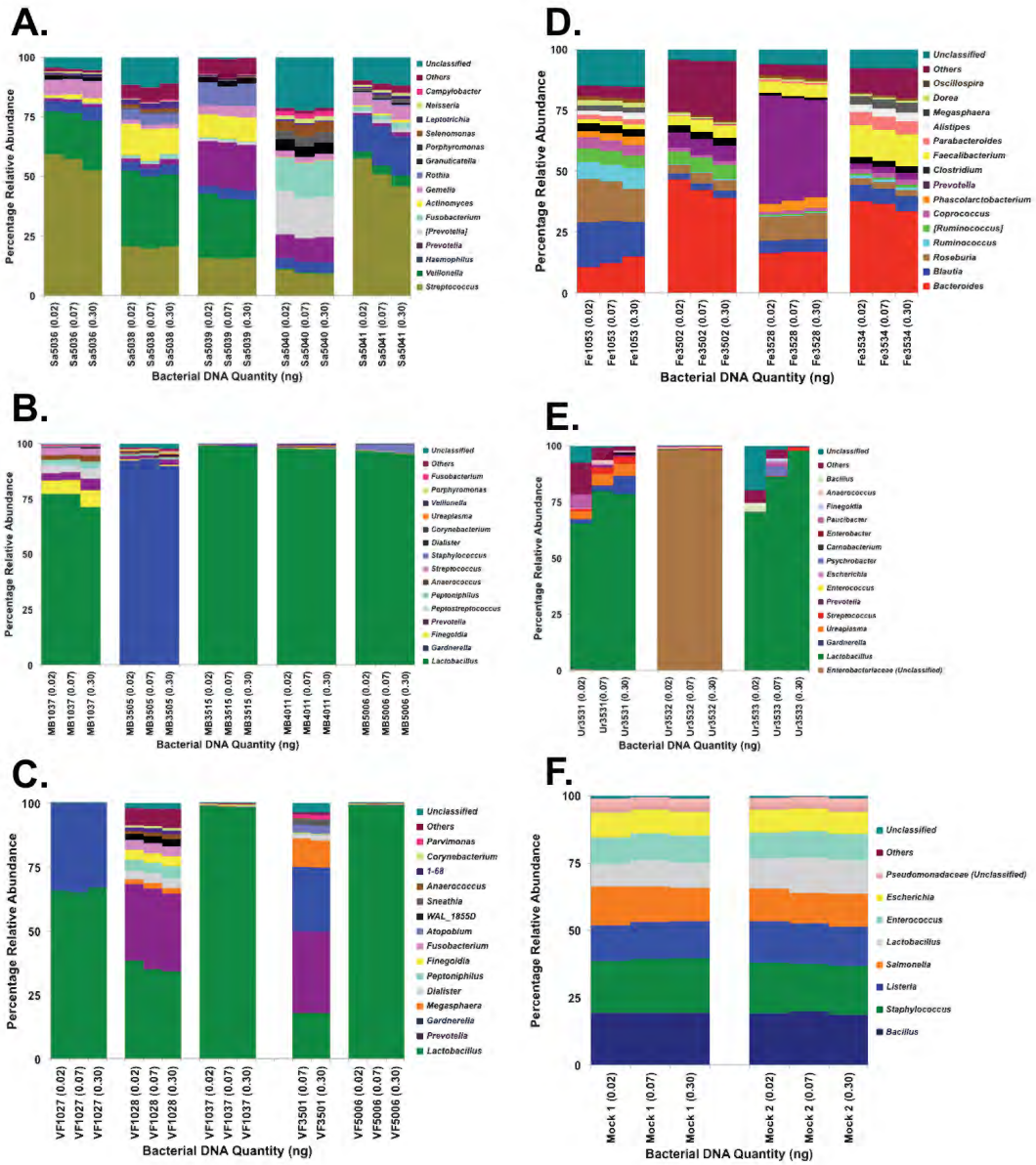


Figure 1: Relative abundance of bacterial genera associated with A.) saliva, B.) menstrual secretion, C.) vaginal secretion, D.) Feces, E.) urine, and F.) ZymoBIOMICS™ microbial community DNA standard at three bacterial DNA quantities. "Others" include those genera whose relative abundance didn't rank in top 15 (top 8 for microbial community DNA standard) for each body fluid[28].

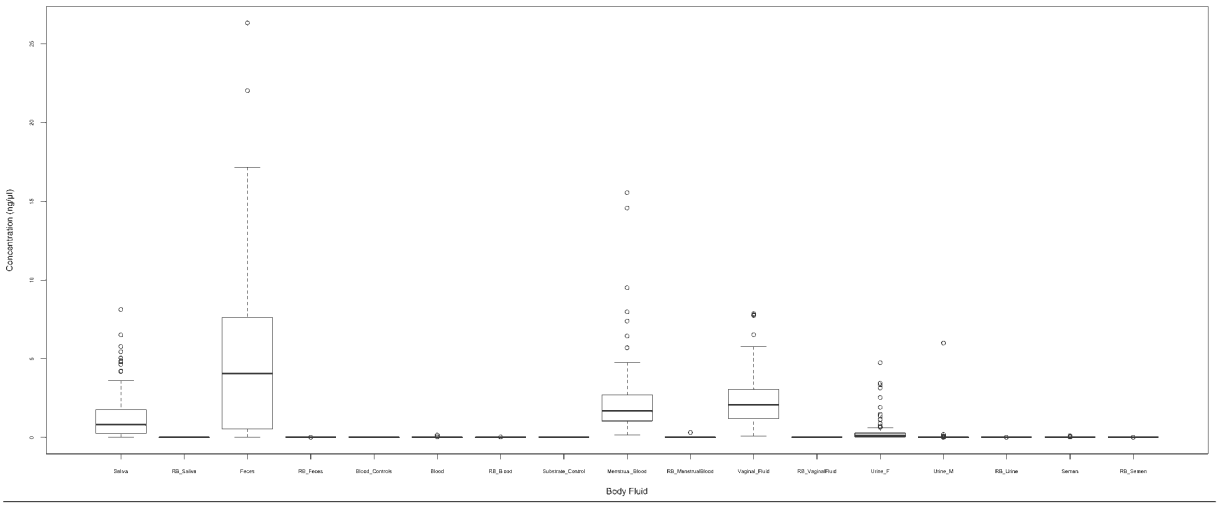


Figure 2: Box plot of bacterial DNA concentrations associated with various biological samples and reagent blanks.

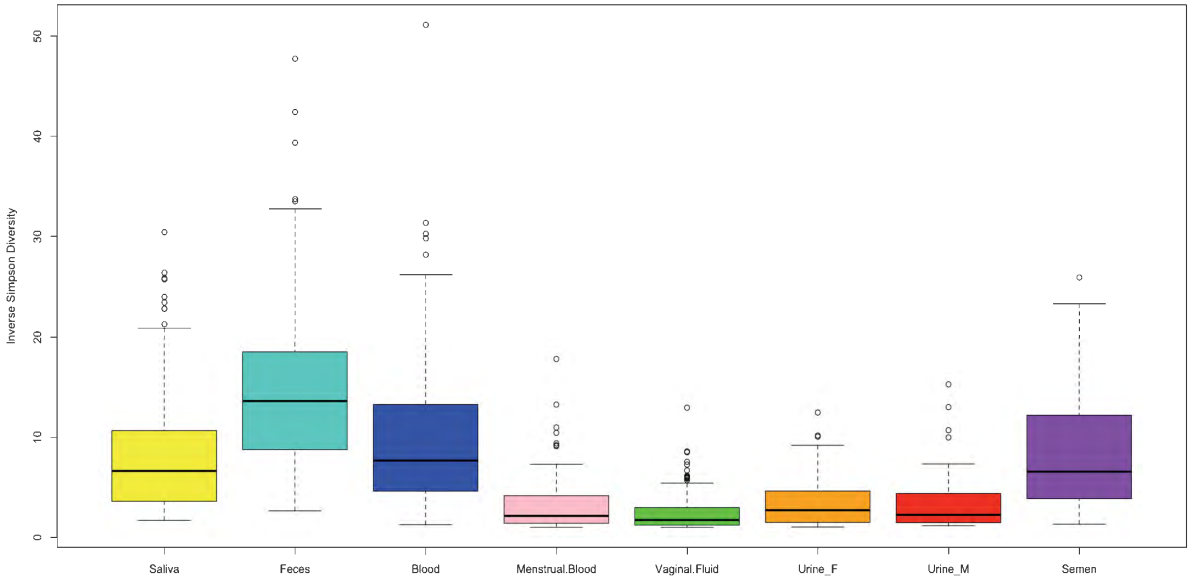
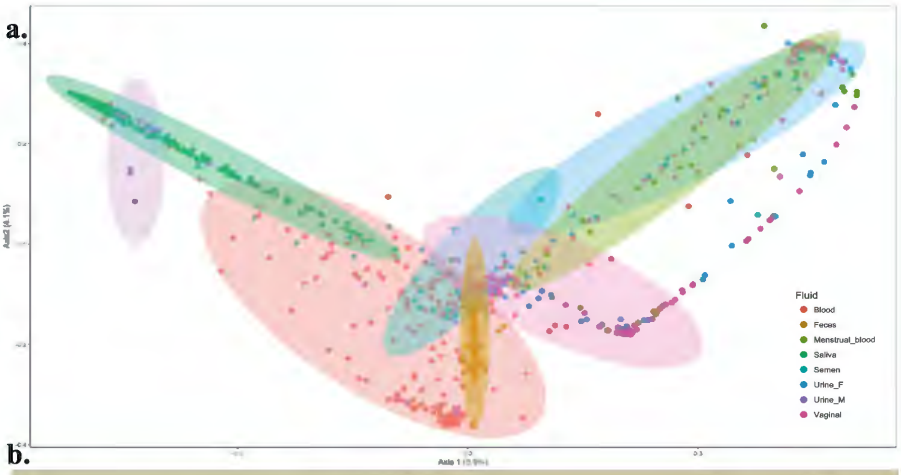


Figure 3. Box plot of bacterial inverse Simpson diversity indices associated with seven biological fluid sample types. Feces, semen, saliva and venous blood demonstrated high bacterial diversity.

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

a.



b.

	Blood	Feces	Menstrual Blood	Saliva	Semen	Urine Female	Urine Male
Blood							
Feces	<0.0002*						
Menstrual Blood	<0.0002*	<0.0002*					
Saliva	<0.0002*	<0.0002*	<0.0002*				
Semen	<0.0002*	<0.0002*	<0.0002*	<0.0002*			
Urine Female	<0.0002*	<0.0002*	0.3276	<0.0002*	<0.0002*		
Urine Male	<0.0002*	<0.0002*	<0.0002*	<0.0002*	<0.0002*	<0.0002*	
Vaginal Fluid	<0.0002*	<0.0002*	0.0676	<0.0002*	<0.0002*	0.0048	<0.0002*

*Indicates significance; Bonferroni pair-wise error rate: 0.0018

Figure 4: a) Principal coordinate analysis of bacterial community structure associated with seven body fluid sample types. b) Analysis of molecular variance (AMOVA) tests showing significant difference in bacterial structure associated with all biological samples except between female urogenital tract samples (i.e., female urine, menstrual secretion, and vaginal secretions).

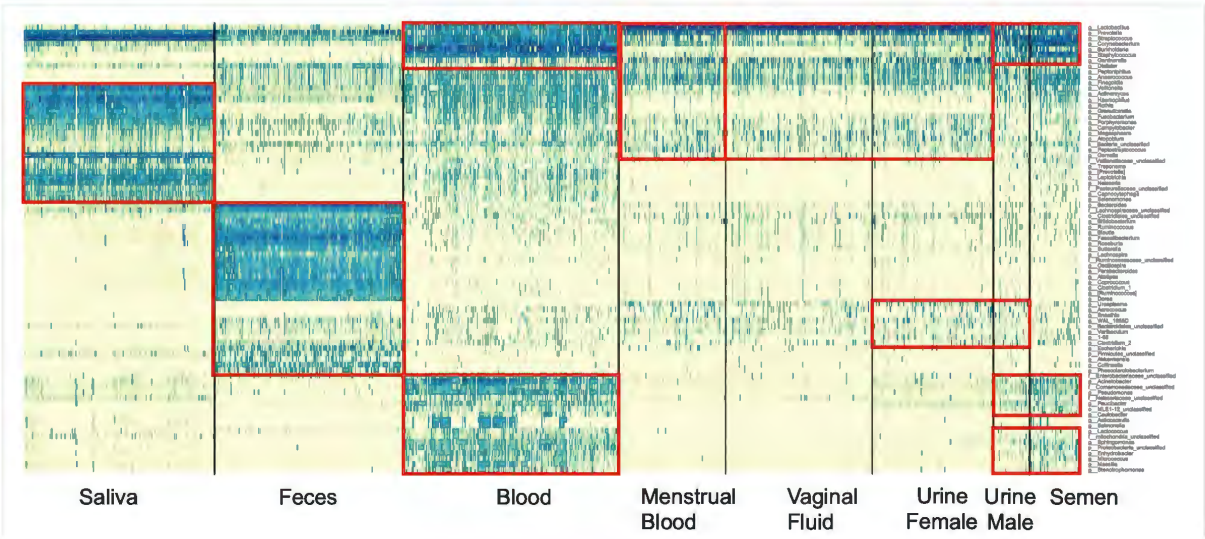


Figure 5. Heatmap of bacterial genera (relative abundance >0.1%) associated with saliva, feces, venous blood, menstrual secretion (MF), vaginal fluid (VF), urine, and semen. Deep blue color indicates higher relative abundance of bacteria.

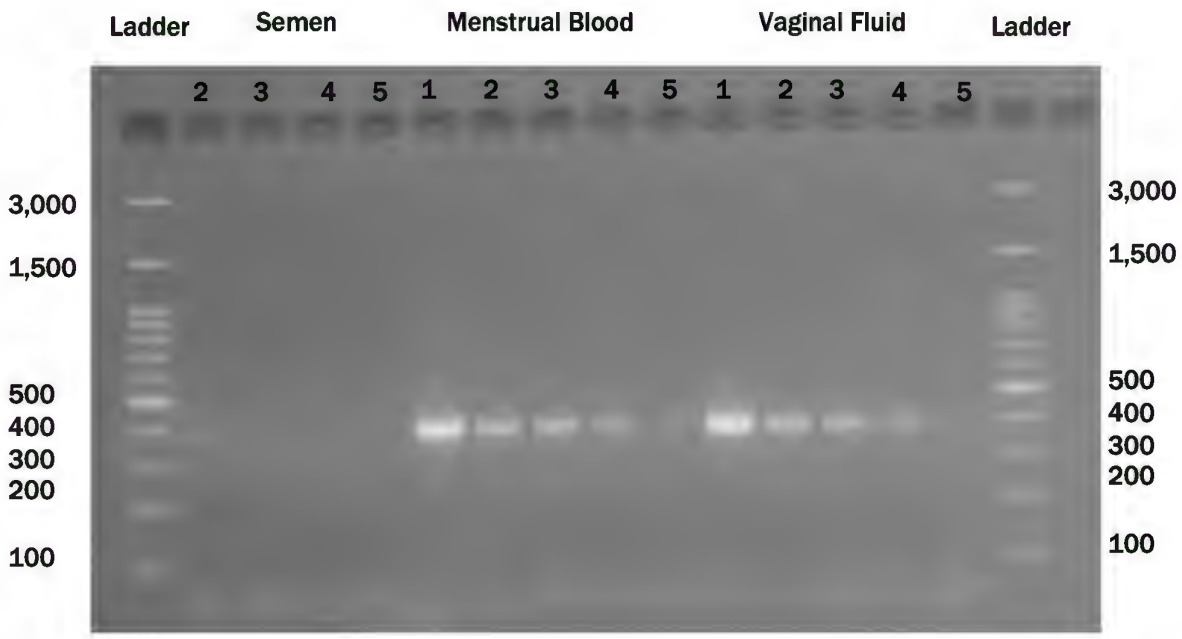


Figure 6: Impact of bacterial DNA input on PCR amplification success. (1) 0.3 ng, (2), 0.02 ng, (3) 0.01 ng, (4) 0.005 ng, (5) 0.001 ng.

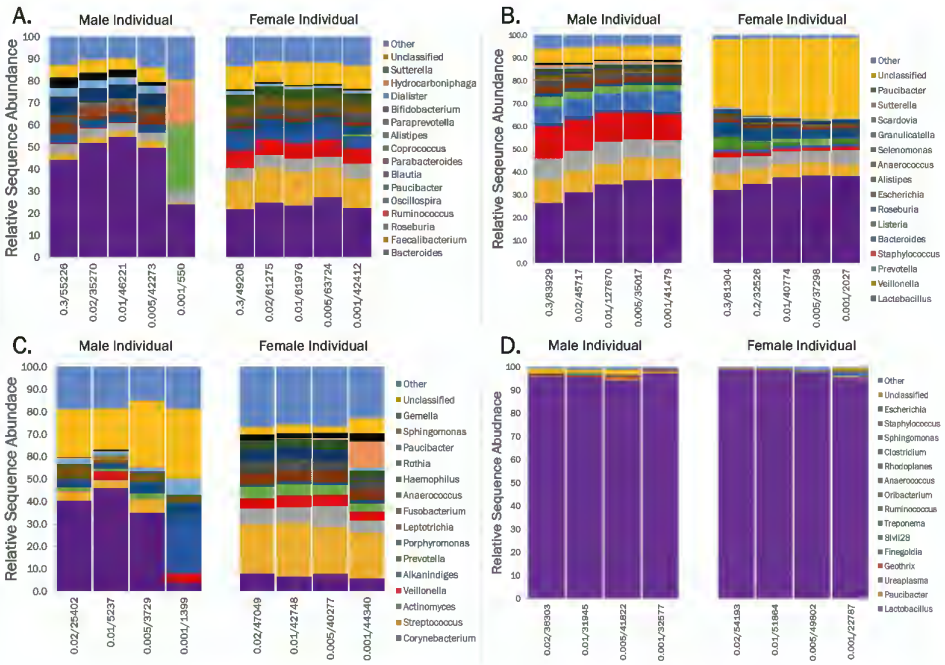


Figure 7: Relative abundance of top 15 genera associated with A) feces, B) saliva, C) blood, and D) urine at various bacterial DNA inputs. Other genera include all those taxonomies whose total relative abundance was under 5%. Unclassified includes all bacteria that could not be classified at the genus level.

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

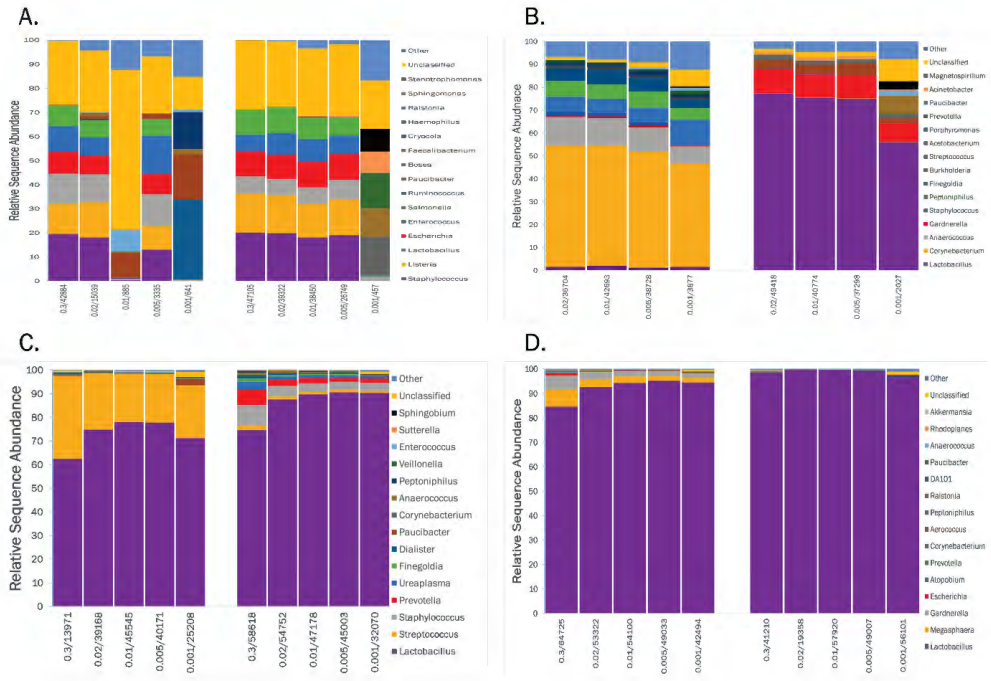


Figure 8: Relative abundance of top 15 genera associated with (A) mock community, (B) semen, (C) menstrual blood, and (D) vaginal fluid at various bacterial DNA inputs. Other genera include all those taxonomies whose total relative abundance was under 5%. Unclassified includes all bacteria that could not be classified at the genus level.

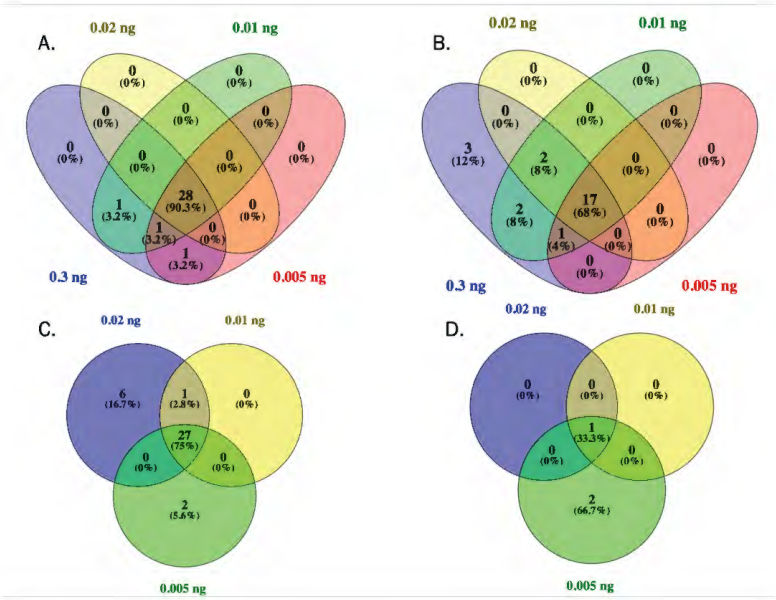


Figure 9: Shared bacteria genera associated with (A) feces (maintained 90.3% of taxa between the different DNA input), (B) saliva (maintained 68% of taxa between the different DNA input), (C) blood (maintained 75% of taxa between the different DNA input), and (D) urine (maintained 33.3% of taxa between the different DNA inputs but was dominated by *Lactobacillus*).

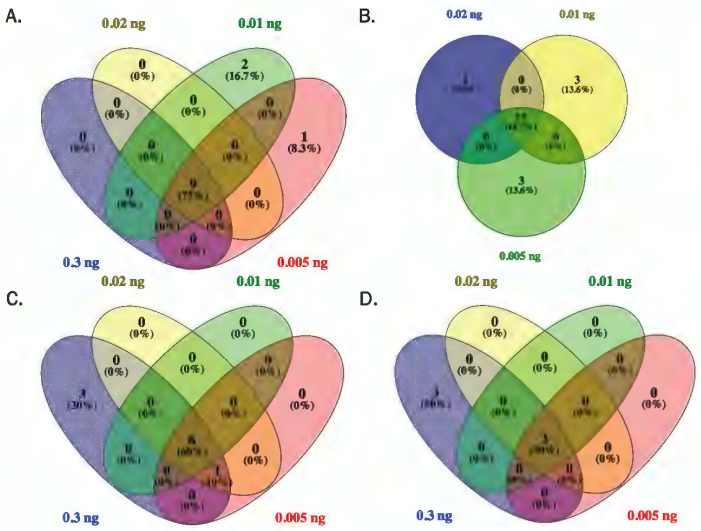


Figure 10: Shared bacteria genera associated with (A) mock community (maintained 75% of taxa between DNA input), (B) semen (maintained 68.2% of taxa between DNA input), (C) menstrual blood (maintained 60% of taxa between DNA input), (D) vaginal fluid (maintained 50% of taxa between DNA input).

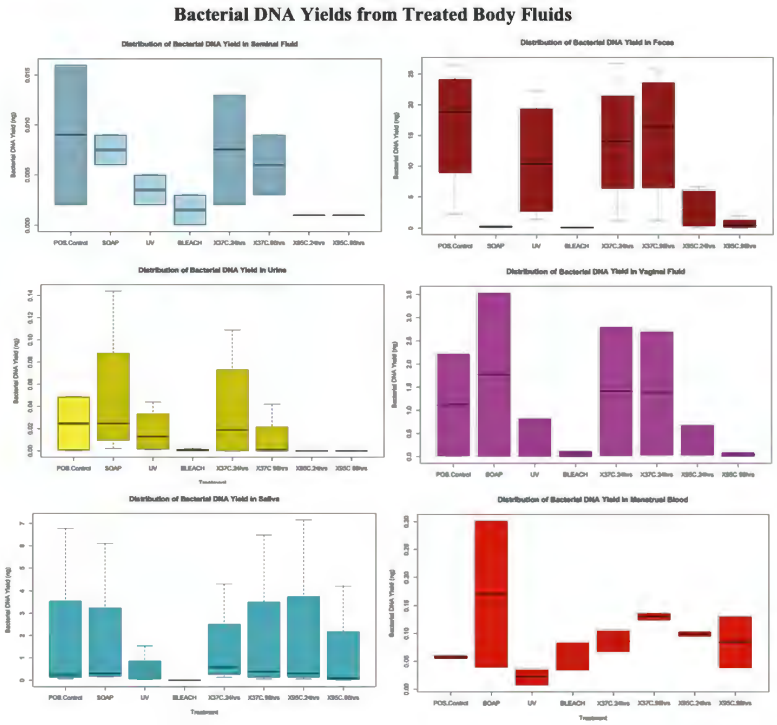


Figure 11. Bacterial DNA yield in soap, Ultraviolet (UV) light, bleach, and temperature treated samples. POS Control=Positive control (i.e., no treatment).

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

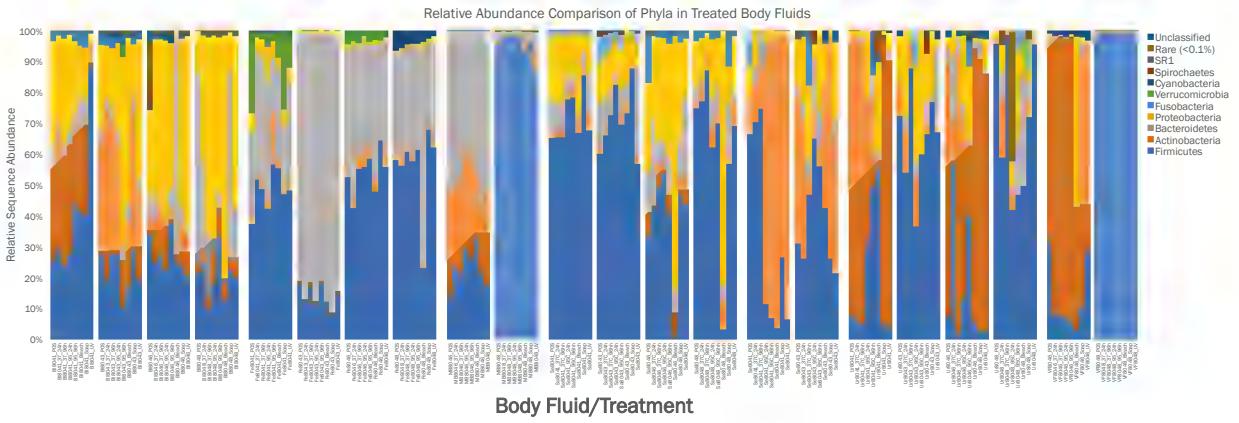


Figure 12. Impact of various physical and chemical treatments on microbial profile generated from forensically relevant biological samples. Bl=venous blood, Fe=Fecal samples, MB=menstrual blood, Sa=saliva, Se=Semen, Ur=Urine, VF=Vaginal fluid, POS=Positive control (i.e., no treatment), UV=ultraviolet light treatment.

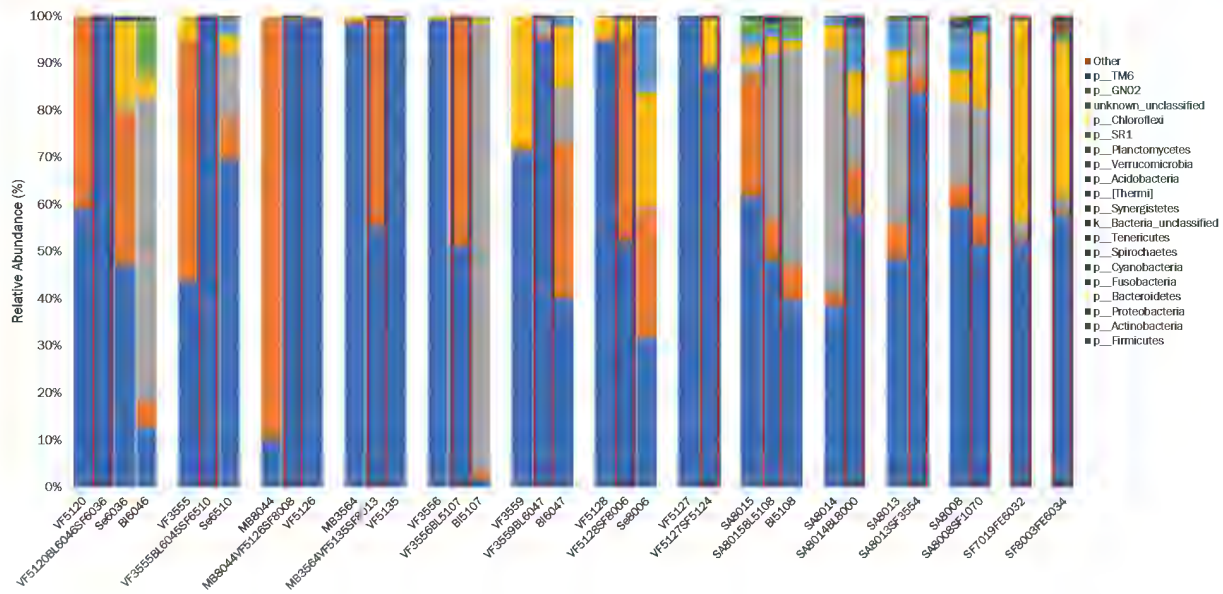


Figure 13: Relative abundance of the top 20 bacterial phyla associated with the fourteen mixture samples, with mixtures outlined in red. Other phyla include all phyla whose relative abundance was not represented in the top 20. VF= Vaginal Fluid, BL=Blood, SF/Se=Seminal Fluid, MB=Menstrual blood, SA=Saliva, FE=Feces.

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

References:

1 Williams, G., Uchimoto, M. L., Coult, N., World, D. & Beasley, E. Body fluid mixtures; resolution using forensic microRNA analysis. *Forensic Science International: Genetics Supplement Series* **4**, 292-293, doi:10.1016/j.fsigss.2013.10.149 (2013).

2 Torres, Y. *et al.* DNA mixtures in forensic casework: a 4-year retrospective study. *Forensic Sci Int* **134**, 180-186 (2003).

3 Butler, J. M. *Forensic DNA Typing, Second Edition: Biology, Technology, and Genetics of STR Markers*. Second edn, 688 (Academic Press, 2005).

4 Jakubowska, J., Maciejewska, A., Pawlowski, R. & Bielawski, K. P. mRNA profiling for vaginal fluid and menstrual blood identification. *Forensic Sci Int Genet* **7**, 272-278, doi:10.1016/j.fsigen.2012.11.005 (2013).

5 Juusola, J. & Ballantyne, J. mRNA profiling for body fluid identification by multiplex quantitative RT-PCR. *J Forensic Sci* **52**, 1252-1262, doi:10.1111/j.1556-4029.2007.00550.x (2007).

6 Day, J. S., Edwards, H. G. M., Dobrowski, S. A. & Voice, A. M. The detection of drugs of abuse in fingerprints using Raman spectroscopy I: latent fingerprints. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **60**, 563-568, doi:http://dx.doi.org/10.1016/S1386-1425(03)00263-4 (2004).

7 Virkler, K. & Lednev, I. K. Raman spectroscopy offers great potential for the nondestructive confirmatory identification of body fluids. *Forensic Sci Int* **181**, e1-5, doi:10.1016/j.forsciint.2008.08.004 (2008).

8 Haas, C. *et al.* RNA/DNA co-analysis from human saliva and semen stains--results of a third collaborative EDNAP exercise. *Forensic Sci Int Genet* **7**, 230-239, doi:10.1016/j.fsigen.2012.10.011 (2013).

9 Nussbaumer, C., Gharehbaghi-Schnell, E. & Korschineck, I. Messenger RNA profiling: a novel method for body fluid identification by real-time PCR. *Forensic Sci Int* **157**, 181-186, doi:10.1016/j.forsciint.2005.10.009 (2006).

10 Juusola, J. & Ballantyne, J. Messenger RNA profiling: a prototype method to supplant conventional methods for body fluid identification. *Forensic Sci Int* **135**, 85-96 (2003).

11 Bauer, M. & Patzelt, D. Evaluation of mRNA markers for the identification of menstrual blood. *J Forensic Sci* **47**, 1278-1282 (2002).

12 Setzer, M., Juusola, J. & Ballantyne, J. Recovery and stability of RNA in vaginal swabs and blood, semen, and saliva stains. *J Forensic Sci* **53**, 296-305, doi:10.1111/j.1556-4029.2007.00652.x (2008).

13 Antunes, J. *et al.* High-resolution melt analysis of DNA methylation to discriminate semen in biological stains. *Anal Biochem* **494**, 40-45, doi:10.1016/j.ab.2015.10.002 (2016).

14 Antunes, J., Balamurugan, K., Duncan, G. & McCord, B. Tissue-Specific DNA Methylation Patterns in Forensic Samples Detected by Pyrosequencing(R). *Methods Mol Biol* **1315**, 397-409, doi:10.1007/978-1-4939-2715-9_27 (2015).

15 Courts, C. & Madea, B. Micro-RNA - A potential for forensic science? *Forensic Sci Int* **203**, 106-111, doi:10.1016/j.forsciint.2010.07.002 (2010).

16 Bai, P. *et al.* Micro RNA profiling for the detection and differentiation of body fluids in forensic stain analysis. *Forensic Science International: Genetics Supplement Series* **4**, e216-e217, doi:http://dx.doi.org/10.1016/j.fsigss.2013.10.111 (2013).

17 Hanson, E. K., Lubenow, H. & Ballantyne, J. Identification of forensically relevant body fluids using a panel of differentially expressed microRNAs. *Analytical Biochemistry* **387**, 303-314, doi:http://dx.doi.org/10.1016/j.ab.2009.01.037 (2009).

18 Wang, Z. *et al.* Screening and confirmation of microRNA markers for forensic body fluid identification. *Forensic Sci Int Genet* **7**, 116-123, doi:10.1016/j.fsigen.2012.07.006 (2013).
19 Zubakov, D. *et al.* MicroRNA markers for forensic body fluid identification obtained from microarray screening and quantitative RT-PCR confirmation. *International journal of legal medicine* **124**, 217-226, doi:10.1007/s00414-009-0402-3 (2010).
20 Fernández, L. *et al.* The human milk microbiota: Origin and potential roles in health and disease. *Pharmacological Research* **69**, 1-10, doi:http://dx.doi.org/10.1016/j.phrs.2012.09.001 (2013).
21 Amar, J. *et al.* Involvement of tissue bacteria in the onset of diabetes in humans: evidence for a concept. *Diabetologia* **54**, 3055-3061, doi:10.1007/s00125-011-2329-8 (2011).
22 Fouts, D. E. *et al.* Integrated next-generation sequencing of 16S rDNA and metaproteomics differentiate the healthy urine microbiome from asymptomatic bacteriuria in neuropathic bladder associated with spinal cord injury. *J Transl Med* **10**, 174, doi:10.1186/1479-5876-10-174 (2012).
23 Pearce, M. M. *et al.* The female urinary microbiome: a comparison of women with and without urgency urinary incontinence. *MBio* **5**, e01283-01214, doi:10.1128/mBio.01283-14 (2014).
24 Human Microbiome Project, C. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207-214, doi:10.1038/nature11234 (2012).
25 Hou, D. *et al.* Microbiota of the seminal fluid from healthy and infertile men. *Fertil Steril* **100**, 1261-1269, doi:10.1016/j.fertnstert.2013.07.1991 (2013).
26 Ravel, J. *et al.* Vaginal microbiome of reproductive-age women. *Proc Natl Acad Sci U S A* **108 Suppl 1**, 4680-4687, doi:10.1073/pnas.1002611107 (2011).
27 Chaban, B. *et al.* Characterization of the vaginal microbiota of healthy Canadian women through the menstrual cycle. *Microbiome* **2**, 23, doi:10.1186/2049-2618-2-23 (2014).
28 Seashols-Williams, S. *et al.* An accurate bacterial DNA quantification assay for HTS library preparation of human biological samples. *Electrophoresis* **39**, 2824-2832, doi:10.1002/elps.201800127 (2018).
29 Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K. & Schloss, P. D. Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the MiSeq Illumina Sequencing Platform. *Appl Environ Microbiol* **79**, 5112-5120, doi:Doi 10.1128/Aem.01043-13 (2013).
30 Schloss, P. D. *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**, 7537-7541, doi:10.1128/AEM.01541-09AEM.01541-09 [pii] (2009).

31 Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C. & Knight, R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**, 2194-2200, doi:10.1093/bioinformatics/btr381btr381 [pii] (2011).

32 Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73**, 5261-5267, doi:10.1128/AEM.00062-07 (2007).

Opportunities for training and professional development & results dissemination

Mentorship:

Currently, four graduate students (1 PhD, 3 MS) from the Department of Forensic Science and one PhD student from the Department of Statistical Sciences and Operations Research are working on their graduate thesis/directed research utilizing samples collected from this study. Two undergraduate students are also involved in work associated with this proposal. Last year three graduate students from the Department of Forensic Science worked on objective two of this proposal and two of them graduated and currently working in the crime laboratories. A year before that, two graduate students from the Department of Forensic Science worked on objective one of this proposal, and graduated. Both students joined crime labs after their graduation. Miss. Francy Nogales an hourly worker associated with this project also joined state crime lab in Fall 2018. This year two more master students from the department of forensic science have joined our lab and has a plan to work on research related with body fluid ID using eukaryotic approach.

Professional development & results dissemination to the community:

PI, CO-PI's and student's have disseminated results obtained from this research to scientific community either through invited lectures or through conference or workshop presentations (see product section for more detail). PI Dr. Baneshwar Singh delivered an invited 45 minutes lecture on new microbiome-based method for the identification body fluid to more than 150 forensic scientists associated with Virginia department of forensic science (state crime labs) during Potomac Regional Symposium on Forensic DNA Analysis at Roanoke, VA in April 2018. Co-PI Dr. Sarah Seashols-William presented results from this study as a poster presentation at Gordon Research Conference on Forensic Analysis of Human DNA in Maine. Another CO-PI, Dr. J Paul Brooks was invited to present statistical results from this study during International Symposium on Mathematical Programming, Bordeaux, France, during INFORMS Annual Meeting, Houston, TX and during INFORMS Healthcare, Rotterdam, The Netherlands. Many student's presented results at 70th and 71st annual meeting of the American Academy of Forensic Sciences (AAFS) at Seattle and Baltimore, respectively. Students also presented finding from this proposal to Mid-Atlantic Association of Forensic Scientists (MAAFS) 2018 and 2019 Annual Meetings at Hunt Valley, MD and Morgantown, WV, respectively. One of the students presenting these findings also received student travel award to attend this meeting. One graduate student presented findings from this proposal during Pittcon 2019 in Philadelphia, PA. This year three graduate students are scheduled to present findings obtained from this proposal to the forensic community during 72rd annual meeting of the American Academy of Forensic Sciences (AAFS) in Anaheim, CA (Date: 17-22 February 2020) and during MAAFS 2020 annual meetings.

Products:

Publication:

S Seashols-Williams, R Green, D Wohlfahrt, A Brand, A Lumjuco Tan-Torres, F Nogales, J P Brooks, B Singh (2018). An accurate bacterial DNA quantification assay for HTS library preparation of human biological samples. Electrophoresis, 39(21), 2824-2832. All raw 16S rDNA sequences associated with this manuscript were submitted to European Nucleotide Archive (ENA) as part of the study with accession # # PRJEB23956.

R Package (new version released):
"pcaL1: L1-Norm PCA Methods," R package, S Jot, JP Brooks, A Visentin, YW Park, Y-H Zhou, http://cran.r-project.org/web/packages/pcaL1, Version 1.5.2 released July 2017.

Conference Proceeding:

JP Brooks, JH Dul´a (2017) "Characterizing L1-Norm Best-Fit Subspaces", Proceedings of SPIE Compressive Sensing Conference, Defense and Commercial Sensing, 10211-2, Anaheim, CA, April 2017.

Conference/workshop Presentations:

1. **JP Brooks** (Invited). Sufficient Conditions for L1-Norm Best-Fit Lines," International Symposium on Mathematical Programming, Bordeaux, France, July 2018.

2. **S Seashols-Williams**, Denise Wohlfart, Eric Abshier, Raquel Green, Francy Nogales, Elena Martinez Planes, Antonio Limjuco Tan-Torres, Angela Brand, Baneshwar Singh, and J Paul Brooks. Forensic body fluid identification using microbiome signature attribution. Poster presented at the Gordon Research Conference on Forensic Analysis of Human DNA, Sunday River, Maine, June 2018.

3. **Angela Brand**, S Seashols-Williams, R Green, D Wohlfahrt, A Lumjuco Tan-Torres, F Nogales, J P Brooks, B Singh. Forensic Body Fluid Identification Utilizing Microbiome Signature Attribution from 16S rDNA High Throughput Sequencing. Talk, presented at the Mid-Atlantic Association of Forensic Scientists Annual Meeting, May 2018.

4. **Baneshwar Singh (Invited)**, Denise Wohlfart, S Seashols-Williams, Angela Brand, Eric Abshier, Raquel Green, Francy Nogales, Elena Martinez Planes, Antonio Limjuco Tan-Torres, and J Paul Brooks. A Novel Metagenomic Approach for Human Body Fluid Identification. Oral presentation (45 minutes) at Potomac Regional Symposium on Forensic DNA Analysis, Roanoke, VA. April 27, 2018.

5. **Raquel Green**, Elena Martinez Planes, Denise Wohlfahrt, Baneshwar Singh, J. Brooks, Sarah J. Seashols-Williams. A Quantitative Assay for Accurate 16S DNA Quantification for High-Throughput Sequencing (HTS) Library Preparation of Microbial Samples. Poster presented during 70th American Academy of Forensic Sciences Annual Meeting, in Seattle, February 2018.

6. **JP Brooks (Invited).** "L1-Norm Subspace Estimation for Microbiome Data Analysis," INFORMS Annual Meeting, Houston, TX, October 2017.

7. **JP Brooks (Invited).** "Analyzing Longitudinal and Multi-omic Microbiome Data," INFORMS Healthcare, Rotterdam, The Netherlands, July 2017.

8. **JP Brooks (Invited).** "Characterizing L1-Norm Best-Fit Subspaces," SPIE Compressive Sensing VI: From Diverse Modalities to Big Data Analytics, Anaheim, CA, April 2017.

9. **Denise Wohlfahrt**, Raquel Green, Angela L Brand, Eric A. Abshier, Elena Martinez Planes, Antonio Limjuco Tan-Torres, J Paul Brooks, Sarah J. Seashols-Williams, and Baneshwar Singh. Forensic Body Fluid Identification Using Microbiome Signature Attribution Through16S rDNA High Throughput Sequencing. Oral presentation during 71st annual scientific meeting of the American Academy of Forensic Sciences (AAFS) on Thursday, February 21, 2019 (4.15pm-4.30pm) at Baltimore, MD.

10. **Francy Nogales**, Sarah J. Seashols-Williams, Baneshwar Singh, J Paul Brooks, Denise Wohlfahrt, Raquel Green, Antonio Limjuco Tan-Torres, Kathleen Brim, and Najai Bradley. An Evaluation of the Stability of High Throughput Sequencing of Microbial DNA from Compromised Human Samples. Poster presentation during 71st annual scientific meeting of the American Academy of Forensic Sciences (AAFS) on Friday, February 22, 2019 (11.30am-1.00pm) at Baltimore, MD.

11. **Kathleen Brim**, Denise Wohlfahrt, Raquel Green, Angela L Brand, Eric A. Abshier, Najai Bradley, Francy Nogales, J Paul Brooks, Baneshwar Singh, Sarah J. Seashols-Williams, Antonio Limjuco Tan-Torres, Elena Martinez Planes. Forensic Body Fluid Identification Using Microbiome Signature Attribution Through16S rDNA High Throughput Sequencing. Oral presentation during PITTCON Conference & Expo 2019 on Monday, March 18, 2019 (1.50-2.10PM) at Philadelphia, PA.

12. **Green, R.**, Wohlfahrt, D., Brand, A. L., Abshier, E. A., Brim, K., Bradley, N., Nogales, F., Martinez, E. P., Limjuco Tan-Torres, A., Brooks, J. P., Seashols-Williams, S. J., and **Singh, B.** Forensic Body Fluid Identification Utilizing Microbiome Signature Attribution through 16S rDNA High Throughput Sequencing. Oral presentation at the Mid-Atlantic Association of Forensic Scientists (MAAFS) Annual Meeting, Morgantown, May 2019.

13. *MAAFS STUDENT TRAVEL AWARD WINNER* **Bradley, N.** Wohlfahrt, D., Brand, A. L., Abshier, E. A., Brim, K., Green, R., Nogales, F., Martinez, E. P., Limjuco Tan-Torres, A., Brooks, J. P., Seashols-Williams, S. J., and **Singh, B.** Developmental validation of a novel method for body fluid identification using microbial signatures. Oral presentation at the Mid-Atlantic Association of Forensic Scientists (MAAFS) Annual Meeting, Morgantown, May 2019

Website(s) and other Internet site(s): R Package "pcaL1: L1-Norm PCA Methods http://cran.r-project.org/web/packages/pcaL1

Technologies or Techniques: A new method for quantitation of microbial DNA from human biological samples was developed. This will help forensic scientists in standardization of microbial sequencing experimental design, which will ultimately lead to generation of consistent results across forensic laboratories. A new method for identification of human body fluid is also developed and will be validated soon. A new version of **R Package** "pcaL1: L1-Norm PCA Methods," was released in July 2017 for the scientific community.

Other products: Physical collection and 16S rDNA sequence data generation of of 1160 body fluid samples from seven body fluids (58 semen, 168 vaginal secretions, 114 menstrual secretions, 205 saliva, 200 feces, 204 urine, and 211 venous blood). All samples preserved and catalogued according to approved human subjects protection protocol. Currently, we are in a process to generate eukaryotic community profile from these samples for improvement of body fluid ID method based on microbial signature associated with these samples.

1) Rationale for Protocol

The purpose of this protocol is to define the steps and materials needed for the preparation of 16S rDNA sequencing and data analysis for the microbiome-based body fluid identification.

2) Consumables

Reagent/Kit	Catalog #	Price
MiSeq® Reagent Kit v2 (500 cycle)	MS-102-2003	$1,182.75
EBT Elution Buffer with Tris	15025394	$13.00
PhiX Control Kit v3	FC-110-3001	$152.00
16S rDNA Primers (V4515F) – 16 unique indices		$511.36
16S rDNA Primers (V4806R) – 24 unique indices		$721.92
Sequencing Primers (Read 1 V4)		$59.57
Sequencing Primers (Read 2 V4)		$60.04
Sequencing Primers (Index V4)		$60.04
QIAamp DNA Investigator Kit	56504	$263.00
QIAcube Elution Tubes, 1.5 mL (50)	1050875	$16.30
QIAcube Filter-Tips, 200 µL (8 x 128)	990332	$104.00
QIAcube Filter-Tips, 1000 µL (8 x 128)	990352	$105.00
QIAcube Rotor Adapters (240)	990394	$45.60
QIAcube Sample Tubes, 2 mL (1000)	990381	$12.00
Rainin Tips RT-LTS-A-10µL-/F/L-960/10	30389226	$64.50
Rainin Tips RT-LTS-A-200µL-/F/L-960/10	30389240	$83.85
Rainin Tips RT-LTS-A-1000µL-/F/L-768/8	30389213	$83.85
Promega 2X PCR Master Mix (1000 Reactions)	M7505	$697.50

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Nuclease free PCR Water, 500 mL (Promega Corp, USA)	PAP1195	$79.00
25 mM $MgCl_2$, 4x 1.25 mL	R0971	$18.63
Qubit dsDNA HS Assay kit (500 Reactions)	Q32854	$269.00
Qubit Tubes (500)	Q32856	$78.50
Agarose Powder, 1 kg	D0012.SIZE.1kg	$307.20
6X Loading Dye, 5 mL	102877-610	$58.55
VWR 96 Wells PCR Plate (100)	82006-636	$234.37
qPCR Thermal Seal RT2RR films (100)		$127.43
Applied Biosystems™ MicroAmp™ Clear Adhesive Film	4306311	$132.00
qPCR Plates (25)		$54.07
PerfeCTa SYBR Green SuperMix (500 reactions), 10 x 1.25ml	95054-500	$341.70
ZymoBIOMICS ™ microbial community DNA standard D6306	D6305	$208.00
Agencourt AMPure XP, 60 mL	A63881	$1,104.00
Biohazard Sample Bags (100)		$12.40
Swab Boxes (1000)		$150.00
Cotton Swabs, sterile (200)		$199.99
Specimen Cups (100)		$110.00
Lancet (200)		$31.17
Alcohol wipes (100)		$2.50
Band-aids (100)		$6.30

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

3) Average reagents cost and hours for processing of 384 samples

Steps	Sample Collection	DNA Extraction/ Quantification	PCR/Barcoding /Gel	CleanUp & Normalization	Pooling and Library QC	Sequencing	Total
Reagents Cost	1008.62	2813.28	438.59	684.5	9.14	1392.8	6346.93
Man Hours	55	120	20	24	4	4 (~45 hours run time)	227

4) Library Preparation Workflow

Sample Collection

Blood

1. Using an alcohol wipe, clean the whole surface of a finger on the non-dominant hand. Using a single use lancet, prick the donor's finger. Collect blood samples onto a minimum of two sterile cotton swabs and allow to dry at room temperature before storing.
   For control samples (optional): Using a sterile, dry cotton swab, swab the donor's finger prior to pricking the finger as a skin control sample.

Saliva

1. Direct the donor to place two sterile swabs into their mouth and roll it along the inside of their cheeks for 15-20 seconds for salivary transfer. Place the swabs in a swab box and allow to dry at room temperature before storing.

Urine

1. Supply the donor with the appropriate collection materials including a sterile liquid collection cup, ice box as well as required ice pack, and the instructions for individual collection to be returned within 24 hours.

2. Upon return, aliquot 500 µL into three microcentrifuge tubes and fill a 15mL conical tube up to 13 mL to allow for expansion upon freezing. Discard the remaining urine. Store all aliquoted samples at -80°C.

Semen

1. Supply the donor with the appropriate collection materials including a sterile liquid collection cup, ice box as well as required ice pack, and the instructions for individual collection to be returned within 24 hours.

2. Upon return, flick-mix the semen sample and aliquot all semen in 250-500 µL aliquots, depending on volume, into microcentrifuge tubes. Store all aliquoted samples at -80°C.

Feces

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

1. Provide the donor with sterile cotton swabs, swab boxes, and instructions for individual collection. In the privacy of their own home, direct the donor to swab as they are defecating. Here, skin contact or contact with the toilet bowl/water should be avoided. Direct them to place the swabs in the provided swab boxes and allow them to dry at room temperature.

Vaginal Fluid and Menstrual Blood

1. Provide the donor with sterile cotton swabs, swab boxes, and instructions for individual collection. Direct the donor to insert the swabs only 2-3 inches into the vagina and twist the swabs while in the vagina for full coverage of the swab. Direct the donor that they should not force the swab or push it past the cervical opening, nor should they donate these samples if they are or think they may be pregnant.

Sample Storage

Swab Samples

Body fluids collected onto cotton swabs will be dried and stored at room temperature.

Liquid Samples

Store body fluids collected into sterile collection cups (i.e. urine, semen) after aliquot at -80°C. For semen, deposit 150-200 µL of semen onto a sterile cotton swab and dry at room temperature before DNA extraction.

Body Fluid Extraction

Blood, menstrual blood, saliva, semen, feces, and vaginal fluid

1. Using sterilized scissors, cut the ~1/4-1/2 cotton swab with the sample into a clean, labeled 2 mL tube. If extracting semen, pipette 20 µL of 1M DTT into the bottom of the tube before placing the sample cutting into it. Scissors should be heat- or blue flame sterilized between cuttings.
2. Using QIAamp DNA Investigator Kit (Qiagen™ N.V., Venlo, The Netherlands) manufacturer's protocol on the QIAcube robotic platform (Qiagen), prepare the liquid handling robot accordingly for the Lysis Step by properly placing p200 and p1000 tips, samples, as well as the ATL reagent.
3. Following the lysis step, remove the sample tubes from the QIAcube and discard the solid swab from the 2 mL tube. Place the 2 mL tubes back into the shaker rack in the same order.
4. Follow the manufacturer's protocol for setting up the rotor adapters using a labeled elution tube and a spin column.
5. Place the rotor adapters in their corresponding spot on the on-board centrifuge.
6. Once the QIAcube is prepared with reagents, rotor adapters, and samples, follow the manufacturer protocols until the purification step is highlighted on the screen.

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

7. Press edit to change the elution volume to 20 µL (e.g. urine), 30 µL (e.g. blood, menstrual blood, saliva, semen, and vaginal fluid) or 50 µL (e.g. feces), and continue with the purification step.
8. After the protocol has finished, store the eluted DNA at -20°C.

Urine

Notes

1. Equilibrate samples to room temperature (15-25 °C).
2. Set a thermomixer or heated orbital incubator to 56 °C for use in step 5. If thermomixer or heated orbital incubator is not available, a heating block or water bath can be used instead.
3. If Buffer AL or Buffer ATL contains precipitates, dissolve by heating to 70 °C with gentle agitation.

Protocol

1. DNA from urine samples will be extracted using the QIAamp® DNA Micro Kit (QiagenTM, Hilden, Germany) following the manufacturer's protocol.
2. Transfer 1 mL of urine to a 1.5 mL microcentrifuge tube (not provided) and centrifuge at 6,000 x g for 2 minutes.
3. Discard the supernatant, add 500 µL of Buffer AE, and vortex for 5 seconds.
4. Centrifuge at 6,000 x g for 2 minutes.
5. Discard the supernatant, add 300 µL of Buffer ATL and 20 µL of proteinase K to the pellet, and mix by pulse-vortexing for 10 seconds.
6. Place the 1.5 mL tube in a thermomixer and incubate at 56 °C while shaking at 76 x g for 1 hour. If using a heating block or water bath, vortex the tube for 10 seconds every 15 minutes to improve lysis.
7. Briefly centrifuge the 1.5 mL tube to remove drops from inside the lid.
8. Add 300 µL of Buffer AL and 50 µL of 100% ethanol. Close the lid and mix by pulse-vortexing for 10 seconds.
9. Briefly centrifuge the 1.5 mL tube to remove drops from inside the lid. (Note: the white precipitate that may appear in step 7 does not need to be pelleted; it can be transferred to the QIAamp MinElute column with the lysate and does not interfere with the QIAamp procedure.)
10. Carefully transfer the supernatant from step 8 to the QIAamp MinElute column in a 2 mL collection tube without wetting the rim. Close the lid and centrifuge at 6,000 x g for 1 minute. Place the QIAamp MinElute column in a clean 2 mL collection tube and discard the collection tube containing the flow-through. If lysate has not completely passed through the membrane after centrifugation, centrifuge again at a higher speed until the QIAamp MinElute column is empty.
11. Carefully open the QIAamp MinElute column and add 500 µL of Buffer AW1 without wetting the rim. Close the lid and centrifuge at 6,000 x g for 1 minute. Place the QIAamp MinElute column in a clean 2mL collection tube and discard the collection tube containing the flow-through.

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

12. Carefully open the QIAamp MinElute column and add 500 µL of Buffer AW2 without wetting the rim. Close the lid and centrifuge at 6,000 x g for 1 minute. Place the QIAamp MinElute column in a clean 2mL collection tube and discard the collection tube containing the flow-through.
13. Centrifuge at full speed (20,000 x g) for 3 minutes to dry the membrane completely.
14. Place the QIAamp MinElute column in a clean 1.5mL microcentrifuge tube (not provided) and discard the collection tube containing the flow-through. Carefully open the lid of the QIAamp MinElute column and apply 20 µL of Buffer AE to the center of the membrane.
15. Close the lid and incubate at room temperature (15-25°C) for 5 minutes. Centrifuge at full speed (20,000 x g) for 1 minute.

Quantitative PCR (qPCR) Analysis (For detailed information on qPCR method and analysis parameters, see Seashols-Williams et al. (2018) An accurate bacterial DNA quantification assay for HTS library preparation of human biological samples, Electrophoresis, 39(21), 2824-2834).

1. Using ZymoBIOMICS ™ microbial community DNA standard D6306 (Zymo Research, USA), create a serial dilution from 9.6 ng/µL (stock) to 0.001 ng/µL, making enough volume for a standard curve in duplicate.
   Note: for vaginal fluid, menstrual blood, feces, and saliva, a standard curve from 9.6 ng/µL to 0.02 ng/µL is sufficient.

Working Stock Concentration (ng/µL)	Working Stock Volume (µL)	ddH_2O Volume (µL)
9.6	10	0
5	5	5
2.5	5	5
1.25	5	5
0.63	5	5
0.31	5	5
0.16	5	5
0.08	5	5
0.04	5	5
0.02	5	5
0.01	5	5
0.005	5	5
0.0025	5	5
0.001	5	5

For each qPCR reaction, add the following:

PerfeCTa SYBR Green SuperMix (2X)	6.25 µL
V4_515f primer (10 µM)	0.25 µL
V4_806r primer (10 µM)	0.25 µL
Nuclease-Free water	3.75 µL
DNA template	2 µL

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Total Reaction Volume	12.5 µL

2. According to sample number, prepare a master mix containing all components except the DNA, plus 5% volume for pipetting overage.
3. Load 10.5 µL of Master Mix into each well of a 96-well plate.
4. Add 2 µL of sample DNA to its respective well.
5. Seal the plate with a cover and centrifuge plate to get rid of any bubbles.
6. 3-Step Cycling Protocol using the Applied Biosystems QuantStudio 6 Flex Real-Time PCR System
    a. Stage 1: 94°C for 3 minutes.
    b. Stage 2: 35 PCR cycles of:
        i. 94°C for 45 seconds.
        ii. 50°C for 60 seconds.
        iii. 72°C for 90 seconds.
    c. Stage 3: 72°C for 10 minutes.
    d. Stage 4: Melt curve:
        i. 94°C for 15 seconds.
        ii. 60°C for 60 seconds.
        iii. 95°C for 15 seconds.
7. Data Analysis using the integrated QuantStudio Real-Time PCR Software v1.3
    a. Threshold value: 0.07.
    b. Baseline parameters: 1 – 35 cycles.
    c. Positive results: Ct value present.
    d. Negative results (Undetected sample): Undetermined Ct value.

Detailed 16S Protocol

Barcoding using PCR

1. Prepare PCR Barcoding plate map using:
    a. SA701-SA712 with SA501-SA508.
    b. SA701-SA712 with SB501-SB508.
    c. SB701-SB712 with SB501-SB508.
    d. SB701-SB712 with SA501-SA508.

2. For 20 µL reaction volume: (Note: sample and water volumes may be adjusted as needed, such as excluding PCR water from reaction if extract is low in DNA yield.)
    a. 10 µL of Promega 2X Master Mix.
    b. 1.5 µL of 10pM Forward Primer (SA/SB501-508).
    c. 1.5 µL of 10pM Reverse Primer (SA/SB701-712).
    d. 1 µL of 0.02 ng/ µL (e.g. blood, urine, semen) or 0.3 ng/ µL (e.g. saliva, menstrual blood, vaginal fluid, feces) DNA extract
    e. 6 µL of Promega Nuclease-Free Water.
    f. For low DNA samples (e.g. urine, semen, blood), substitute 0.8 µL nuclease-free water with 0.8 µL (1 mM) of with $MgCl_2$ for a total of 2.5 mM $MgCl_2$.
3. Add appropriate amount of master mix to each well.

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

4. Add primers according to PCR plate map, ensuring no duplicates were used, and add appropriate sample to each.
5. Place samples into Thermocycler (Thermo Fisher Scientific Veriti Thermal Cycler) and run the "Kozich_16S_v3_v4_MiSeq" program as follows:
    a. 95°C for 2 minutes.
    b. 30 PCR cycles of:
        i. 95°C for 20 seconds.
        ii. 55°C for 15 seconds.
        iii. 72°C for 5 minutes.
    c. 72°C for 10 minutes.
    d. 4°C hold.
6. Freeze amplified plate in post-PCR -20 °C freezer until further use.

Gel Electrophoresis

1. On a sheet of parafilm, mix 2 µL of 6X loading dye, 3 µL of amplified product, and 7 µL of 1X TAE buffer and run on a 1.6% agarose gel to test for successful amplification.
2. Run at 120 V for 45 minutes for single lane or 20 minutes for double lane, checking the migration status regularly.

Post-PCR (Agencourt AMPure XP)

1. Obtain Agencourt AMPure XP bottle stored in a 4 °C refrigerator and shake to re-suspend magnetic particles settled in the bottle.
2. Add 18 µL of AMPure XP to 10 µL of PCR product to bind DNA fragments to magnetic beads. Pipette mixture 10 times to mix and let samples incubate at room temperature for 5 minutes.
3. Place reaction mixture on Agencourt SPRIPlate 96 Super Magnet Plate for 2 minutes to separate the beads from the solution.
4. While mixture is still on plate, remove the cleared solution from the plate or tubes and discard. Leave approximately 5 µL of supernatant behind as not to disturb the ring of magnetic beads.
5. While plate or tubes are still on the magnetic stand, add 200 µL of 70% ethanol to each well and incubate at room temperature for 30 seconds. Aspirate ethanol and discard.
6. Repeat ethanol step for a total of two washes.
7. Remove the plate or tubes from the magnetic plate and add 40 µL of PCR water to each well and mix by pipetting 10 times. Incubate mixture at room temperature for 2 minutes. (Note: if liquid is less than 40 µL, reaction may need additional mixing to ensure that PCR product is eluted off of the beads.)
8. Place reaction mixture onto magnetic plate for 1 minute to separate the beads and the transfer eluate, now containing the DNA, to a new plate or set of tubes. (Note: if reaction volume does not reach the magnetic ring, slowly move the reaction plates to allow beads to make contact with the ring.) (Note: eluate should be kept on the magnetic stand to avoid magnetic bead residue carry over into downstream reactions; this residue may affect the Qubit quantification values.)

Qubit Quantification

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

1. Set up an appropriate number of 0.5mL Qubit assay tubes. This number should include the number of samples plus the required number of standards.
2. Label the tube lids. Do not write on the tube sides as this may interfere with the sample quantification.
3. Prepare the Qubit working solution.
    a. For Qubit dsDNA HS kit, mix 199 µL of Qubit dsDNA HS Buffer with 1 µL of Qubit dsDNA HS Reagent dye for each sample.
    Note: do not mix working solution in a glass container
        i. Ex. 8 samples + 2 standards = 10 total = 1990 µL Buffer + 10 µL Reagent.
4. Add 190 µL of working solution to each prepared **standard** tube.
5. Add 10 µL of each **standard** to the appropriate tube and vortex for 2-3 seconds, ensuring that no bubbles are created.
6. Add 180-199 µL of working solution to each prepared **assay** tube. Add the appropriate amount of sample to each assay tube, ensuring that the total volume is 200 µL. Vortex for 2-3 seconds. (Note: 1 – 2 µL of sample should be used in assay tubes involving samples that are expected to have a high DNA yield; samples that are expected to have a low yield DNA yield can be adjusted upward as needed.)
7. Allow samples to incubate at room temperature for 2 minutes.
8. On Qubit instrument:
    a. Select assay (DNA)
    b. Select Qubit Kit (dsDNA HS)
9. The instrument will ask if new standards will be read. Select "Yes."
10. Follow onscreen instructions.
    a. Load appropriate standard(s).
    b. Load assay tubes.

MiSeq Library Preparation

1. While on magnetic stand, quantify all AMPure purified samples using the Qubit. (Note: all quantifications should be done using the same Qubit standards.) (Note: ensure quantifications and dilutions are done on the same day to ensure evenness of samples distribution in the final pool).
2. Dilute all samples to 1 ng/µL. Pool 1 µL of each diluted 1 ng/µL sample.
3. For samples less than 1 ng/µL, add enough of the purified sample to equal 1ng of DNA.
    a. Ex. Sample A = 0.5 ng/µL. 2 µL of Sample A will be added to the pool.
4. Add 1 µL of all Reagent Blanks or Negative Controls with no quantifiable DNA. Add up to 4 µL of Reagent blanks and negative controls with quantifiable amounts of DNA.
5. If necessary, dry down pool using a vacuum centrifuge with **NO** heat.

a. Ex. If 300 samples were added to the pool, the total volume should equal ~300 µL.

6. Quantify final pool to ensure total quant is 1.0 ng/µL. Keep pool at -20 ^{o}C until sequencing.

5) Sequencing

MiSeq Protocol

1. Prepare a container of de-ionized water for the thawing reagent box prior to preparing the library and PhiX. (Note: thawing time is approximately 1 hour.) (Note: the cartridge is

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

only good for 2 – 6 hours once thawed. Ensure that the library is ready for sequencing prior to thawing the reagent cartridge.)

2. Prepare fresh NaOH.
    a. 10N to 1 N NaOH:
        i. 100 µL of 10N NaOH + 900 µL of PCR H_2O = 1000 µL of 1N NaOH
    b. 1N to 200mN NaOH:
        i. 200 µL of 1N NaOH + 800 µL of PCR H_2O = 1000 µL of 200mN NaOH
3. Prepare Library: 4nM Library in 100 µL volume:
    a. Using a calculation excel sheet, use Qubit readings to make 4nM Library.
        i. Ex. 1.01 ng/µL Qubit reading: 98.851 µL of pooled DNA + 1.1485 µL of EBT = 100 µL of 4nM Library.
    b. Denaturing Library:
        i. Combine 5 µL of 4nM Library with 5 µL of 200mM NaOH.
        ii. Mix and incubate at room temperature for 5 minutes.
        iii. Add 990 µL of HT1 (Hybridization buffer) to make 1000 µL of 20pM denatured Library.
4. Prepare PhiX:
    a. Prepare 4nM PhiX
        i. Combine 2 µL of 10nM PhiX with 3 µL of EBT to make 5 µL of 4nM PhiX.
        ii. Combine 5 µL of 4nM PhiX and 5 µL of 200mM NaOH.
        iii. Mix and incubate at room temperature for 5 minutes.
        iv. Add 990 µL of HT1 to make 1000 µL of 20pM denatured PhiX.
5. Dilute Denatured Library:
    a. For 10pM:
        i. Combine 300 µL of denatured 20pM Library and 300 µL of HT1 to make 600 µL of 10pM Library.
6. Dilute Denatured PhiX:
    a. For 10pM:
        i. Combine 300 µL of denatured 20pM PhiX and 300 µL of HT1 to make 600 µL of 10pM PhiX.
7. To create 10% PhiX/90% Library:
    a. Combine 60 µL 10pM PhiX with 540 µL 10pM Library.

8. Once Cartridge is thawed:
    a. Invert Cartridge several times.
    b. Using a 1mL pipette tip, puncture well #12.
        i. Load 3 µL (100 µM) of Read 1 and pipette 10 times to mix.
    c. Using a 1mL pipette tip, puncture well #13.
        i. Load 3 µL (100 µM) of Index and pipette 10 times to mix.
    d. Using a 1mL pipette tip, puncture well #14.
        i. Load 3 µL (100 µM) of Read 2 and pipette 10 times to mix.
    e. Using a 1 mL pipette tip, puncture well #17.
        i. Load 600 µL of 10pM Library/PhiX Mixture.
9. Clean Flow Cell.
    a. Using forceps or hand, carefully pull cell out of liquid.

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

b. Rinse cell with continuous flow of ddH_2O for some time with special emphasis on the areas where glass and frame connect to ensure that all salts are rinsed away.
c. Dry flow cell carefully and use 100% EtOH on a KimWipe to dry all water while removing lint and debris. (Note: avoid black area by glass.) Here, canned air may be used to gently move liquid out from under the frame, but black area must be avoided.
d. Place flow cell in MiSeqFGx.

10. Once Reagents, Cartridge, and Flow cell are loaded, perform a pre-run check.
11. Prepare Sample Sheet according to V2 kit or V3 kit template. (Note: do **NOT** include hyphens, periods, or spaces in the sample names; this will result in analysis or indexing error.)
12. Add prepared Sample Sheet to MiSeq via flash drive.
13. Select RUO mode, select sample sheet, and follow the on-screen instructions.

6) Data Analysis

Opening Mothur

Navigate to the mothur github and download the mothur version appropriate for your operating system (https://github.com/mothur/mothur/releases/tag/1.44.0).

- To do this on Windows:
  - Double click the zipped folder, and click "Extract all" icon under Compressed Folder Tools in the Extract tab of the File Explorer.
  - Select a destination, then click Extract.
  - OR right click -> "Extract all"
  - Select a destination, the click Extract.
- To do this on Mac OSX:
  - Double click on file.
- Access the Mothur command prompt:

Mac OS:

1. Open terminal in the applications folder.
2. Change the working directory to the pathway where you plan to run your analysis.
   - cd /your/location/here will set the working directory.

Note: the mothur executable and reference files need to be in the same folder as the MiSeq read1 and read 2 .fastq files to be analyzed.

3. All sequencing files will be available for download on Illumina's BaseSpace. Download fastq.gz files from BaseSpace and extract .fastq files into analysis location.
4. Once all required files (e.g. mothur, reference files, fastq) are in place and working directory has been set, open mothur using the command line.
   - ./mothur = opens mothur

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

IMPORTANT For compatibility and formatting necessities, a text editor such as BBedit or TextWrangler is needed (https://www.barebones.com/products/bbedit/).

Making the Names/Identifier Text File

5. Using excel, create table with sample name, read 1 fastq, and read 2 fastq. No headers should be used for this file.

Sample	Read1.fastq	Read2.fastq
Sample1	Sample1_R1_001.fastq	Sample1_R2_001.fastq
Sample2	Sample2_R1_001.fastq	Sample2_R2_001.fastq

6. Save table as .txt file for downstream analysis.
   - Ex. make.contigs (file=Sample.txt)

Making contiguous sequences and Pre-filter screening

7. To make contiguous sequences and screen out "bad" sequences, use the commands given below. If computing power allows it, processors may be increased using the processors= command.
   - Please note, due to computing power, the initial steps of the analysis are done on a body fluid type basis.
   - The silva bacterial reference should be in the same location as the .fastq files.
   - For full commands (by fluid), please see Mothur_commands_All_byFluid.txt file.

```
//BLOOD
make.contigs(file=Blood.txt, trimoverlap=f, insert=20, processors=8)
summary.seqs(fasta=Blood.trim.contigs.fasta)

screen.seqs(fasta=Blood.trim.contigs.fasta, group=Blood.contigs.groups,
    maxambig=0, minlength=200, maxlength=275)
summary.seqs(fasta=Blood.trim.contigs.good.fasta, processors=8)
unique.seqs(fasta=Blood.trim.contigs.good.fasta)
summary.seqs(fasta=Blood.trim.contigs.good.unique.fasta,
    name=Blood.trim.contigs.good.names)
```


This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

```
align.seqs(fasta=Blood.trim.contigs.good.unique.fasta, reference=silva.bacteria.fasta,
    flip=TRUE)
summary.seqs(fasta=Blood.trim.contigs.good.unique.align,
    name=Blood.trim.contigs.good.names)
screen.seqs(fasta=Blood.trim.contigs.good.unique.align,
    name=Blood.trim.contigs.good.names, group=Blood.contigs.good.groups,
    summary=Blood.trim.contigs.good.unique.summary, start=13862, end=23444, maxhomop=8)
summary.seqs(fasta=Blood.trim.contigs.good.unique.good.align,
    name=Blood.trim.contigs.good.good.names)
```

Combined Filtering

8. To avoid downstream problems during DNA analysis, all files are filtered (alignment column removal) together. This ensures all samples are filtered under same criteria but will output separate, in this case by body fluid, filter.fasta files.

```
filter.seqs(fasta=Blood.trim.contigs.good.unique.good.align-
    Feces.trim.contigs.good.unique.good.align-Saliva.trim.contigs.good.unique.good.align-
    Semen.trim.contigs.good.unique.good.align-
    MenstrualBlood.trim.contigs.good.unique.good.align-
    VaginalFluid.trim.contigs.good.unique.good.align-
    Urine.trim.contigs.good.unique.good.align, vertical=T, trump=., processors=8)
```

Unique sequence filter, Pre-clustering and Chimeric sequence removal

9. To decrease potential processing time and the processing of identical sequences, fasta-formatted sequences are filtered unique sequences.
    - Please note that this section of commands are processed on a fluid by fluid basis.
10. For initial linkage and removal of sequences derived from sequencing errors, fasta-formatted sequences are pre-clustered.
11. Chimeric sequences (sequences combining two or more biological sequences during PCR steps) are removed to avoid analyzing these “fake” sequence.
    - Some OSX versions may require the use of VSEARCH instead of UCHIME. The commands and executable for this should be adjusted as such (chimera.vsearch()).

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

```
//BLOOD
unique.seqs(fasta=Blood.trim.contigs.good.unique.good.filter.fasta,
    name=Blood.trim.contigs.good.good.names)
summary.seqs(fasta=Blood.trim.contigs.good.unique.good.filter.unique.fasta,
    name=Blood.trim.contigs.good.unique.good.filter.names)

pre.cluster(fasta=Blood.trim.contigs.good.unique.good.filter.unique.fasta,
    name=Blood.trim.contigs.good.unique.good.filter.names,
    group=Blood.contigs.good.good.groups, diffs=2, processors=8)
summary.seqs(fasta=Blood.trim.contigs.good.unique.good.filter.unique.precluster.fasta,
    name=Blood.trim.contigs.good.unique.good.filter.unique.precluster.names)

chimera.uchime(fasta=Blood.trim.contigs.good.unique.good.filter.unique.precluster.fasta,
    name=Blood.trim.contigs.good.unique.good.filter.unique.precluster.names,
        group=Blood.contigs.good.good.groups, dereplicate=t, processors=8)
remove.seqs(fasta=Blood.trim.contigs.good.unique.good.filter.unique.precluster.fasta,
    accnos=Blood.trim.contigs.good.unique.good.filter.unique.precluster.denovo.uchime.accnos,
    name=Blood.trim.contigs.good.unique.good.filter.unique.precluster.names,
    group=Blood.contigs.good.good.groups, dups=T)
summary.seqs(fasta=Blood.trim.contigs.good.unique.good.filter.unique.precluster.pick.fasta,
    name=Blood.trim.contigs.good.unique.good.filter.unique.precluster.pick.names, processors=8)
```

Merge body fluid, .fasta, .names, and .groups files for combined Analysis

12. Remaining analysis will use merged files consisting of data from all body fluids.

```
merge.files(input=Blood.trim.contigs.good.unique.good.filter.unique.precluster.pick.fasta
    -Feces.trim.contigs.good.unique.good.filter.unique.precluster.pick.fasta
    -Saliva.trim.contigs.good.unique.good.filter.unique.precluster.pick.fasta
    -Semen.trim.contigs.good.unique.good.filter.unique.precluster.pick.fasta
    -VaginalFluid.trim.contigs.good.unique.good.filter.unique.precluster.pick.fasta
    -MenstrualBlood.trim.contigs.good.unique.good.filter.unique.precluster.pick.fasta
    -Urine.trim.contigs.good.unique.good.filter.unique.precluster.pick.fasta,
    output=Body_fluids.fasta)
merge.files(input=Blood.trim.contigs.good.unique.good.filter.unique.precluster.pick.names
    -Feces.trim.contigs.good.unique.good.filter.unique.precluster.pick.names
    -Saliva.trim.contigs.good.unique.good.filter.unique.precluster.pick.names
    -Semen.trim.contigs.good.unique.good.filter.unique.precluster.pick.names
    -VaginalFluid.trim.contigs.good.unique.good.filter.unique.precluster.pick.names
    -MenstrualBlood.trim.contigs.good.unique.good.filter.unique.precluster.pick.names
    -Urine.trim.contigs.good.unique.good.filter.unique.precluster.pick.names,
    output=Body_fluids.names)
merge.files(input=Blood.contigs.good.good.groups-Feces.contigs.good.good.groups
    -Saliva.contigs.good.good.groups-Semen.contigs.good.good.groups
    -MenstrualBlood.contigs.good.good.groups-VaginalFluid.contigs.good.good.groups
    -Urine.contigs.good.good.groups, output=Body_fluids.groups)
```

Classify sequences, remove unwanted lineages, and normalizing the data

13. Sequences will be classified using the GreenGenes database.
    - Ensure that the required reference files (GreenGenes fasta and taxonomy files) are available in the same location as mothur and the fastq files.

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

```
classify.seqs(fasta=Body_fluids.fasta, name=Body_fluids.names, group=Body_fluids.groups,
    template=gg_13_8_99.fasta, taxonomy=gg_13_8_99.gg.tax, cutoff=70, processors=8)
```

14. Non-bacterial kingdoms such as chloroplast, mitochondria, archea, and unknowns should be removed before continuing analysis. Additionally, *Acinetobacter* and *Burkholderia* should be removed as these are mainly reagent blank specific.

```
remove.lineage(fasta=Body_fluids.fasta, name=Body_fluids.names, group=Body_fluids.groups,
    taxonomy=Body_fluids.gg.wang.taxonomy, taxon=Chloroplast-Mitochondria-unknown
    -Archaea-Eukaryota-g__Burkholderia-g__Acinetobacter)
```

15. To compare data of varying sequence numbers, data should be normalized (sub-sampled). Sub-sampling number may vary but should not be below 6000. Minimal sample loss should be considered when selecting this value, while still being representative of the bacterial communities.
    - Open the count.groups file using excel for easy sorting. Sort smallest to largest to determine the sub-sampling number.

      ```
      count.groups()
      ```

    - Sub-sample sequences by sample (persample=TRUE) at the selected size.

      ```
      sub.sample(fasta=Body_fluids.pick.fasta, name=Body_fluids.pick.names,
          group=Body_fluids.pick.groups, persample=true, size=6046)
      ```

16. Once unwanted kingdoms and taxa are removed and normalized, sequences should be reclassified to reflect actual sequence taxa.

```
classify.seqs(fasta=Body_fluids.pick.subsample.fasta, name=Body_fluids.pick.subsample.names,
    group=Body_fluids.pick.subsample.groups, template=gg_13_8_99.fasta,
    taxonomy=gg_13_8_99.gg.tax, cutoff=80, processors=8)
```

17. *Optional.* Samples may be renamed but this step is not necessary for downstream analysis success. Here, cp original name followed by the new name may be used.

```
system(cp Body_fluids.pick.subsample.gg.wang.tax.summary body_fluid_subsample_gg.summary)
system(cp Body_fluids.pick.subsample.gg.wang.taxonomy body_fluid_subsample_gg.taxonomy)
system(cp Body_fluids.pick.subsample.groups body_fluid_subsample_gg.groups)
system(cp Body_fluids.pick.subsample.names body_fluid_subsample_gg.names)
system(cp Body_fluids.pick.subsample.fasta body_fluid_subsample_gg.fasta)
```

Sequence similarity, clustering, and OTU Classification

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

18. Uncorrected pairwise distances between DNA sequences are calculated using a fasta-formatted alignment file. Here, no distances larger than 0.20 will be saved.

```
dist.seqs(fasta=body_fluid_subsample_gg.fasta, cutoff=0.20, countends=F, processors=8)
```

19. Sequences are assigned to Operational Taxonomic Units (OTU) using the average neighbor algorithm and an increased cutoff of 0.20 to optimize computing time.

```
cluster(column=body_fluid_subsample_gg.dist, name=body_fluid_subsample_gg.names,
    method=average, cutoff=0.20)
```

20. Next OTUs should be organized into a table and rabund files (can be plotted as rank-abundance plots) created, only listing lines labeled at 0.05.

```
make.shared(list=body_fluid_subsample_gg.an.list, group=body_fluid_subsample_gg.groups,
    label=0.05)
```

21. Now the taxonomy of each OTU may be identified.

```
classify.otu(list=body_fluid_subsample_gg.an.list, name=body_fluid_subsample_gg.names,
    taxonomy=body_fluid_subsample_gg.taxonomy, label=0.05)
```

Alpha-, Beta-Diversity, and Principal Coordinate Analysis

22. To determine the within body fluid diversity, the inverse Simpson index is used and calculated values are summarized.

```
collect.single(shared=body_fluid_subsample_gg.an.shared, calc=invsimpson, freq=10)
summary.single(shared=body_fluid_subsample_gg.an.shared, calc=nseqs-coverage-sobs
    -invsimpson, iters=1000)
```

23. To compare species richness observed across all samples, a rarefaction curve is generated. As the curve plateaus (becomes parallel to the x-axis), the analyst can be confident in the sampling size.

```
rarefaction.single(shared=body_fluid_subsample_gg.an.shared, calc=sobs, freq=100)
```

24. Using the shared file, dissimilarities among multiple groups are calculated using the jaccard (unweighted) and Bray-Curtis (weighted) calculators.
Note that the shared file was created using 0.05 label, but this may be adjusted if needed. To visualize these distance matrices, multidimensional Principal Coordinate Analysis (PCoA) is used. For plotting information, please see R-commands below.

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

```
dist.shared(shared=body_fluid_subsample_gg.an.shared, calc=jclass-braycurtis)
pcoa(phylip=body_fluid_subsample_gg.an.braycurtis.0.05.lt.dist)
pcoa(phylip=body_fluid_subsample_gg.an.jclass.0.05.lt.dist)
```

AMOVA and Indicator Taxa

25. To determine whether spatial differences between two or more groups are significant, we run an Analysis of Molecular Variance (AMOVA). First, a design file specifying which body fluid each sample belongs to must be created.

    Making the .design file

    - First open the .axes file created by pcoa() in excel. Save as .design. The error message about file formatting may be ignored at this time. Using the axes file will ensure that this design file is also compatible with PCoA plotting, etc.
    - Insert a new second column. This column will identify the respective body fluid design:

group	design	axis1	axis2	axis3	axis4	axis5	axis6
Bl1	Blood	-0.068566	0.228107	0.015316	-0.119483	-0.099091	0.013926
Bl2	Blood	-0.02231	0.240888	0.01711	0.085778	-0.077985	0.023894
Bl5	Blood	-0.061714	0.256791	0.098877	0.019836	-0.142417	0.013925
Fe1	Feces	-0.095146	0.26592	0.086189	0.0491	-0.125378	-0.02231
Fe6	Feces	-0.001626	0.317993	-0.022188	0.083148	-0.130276	-0.064927
VF1	VaginalFluid	-0.005805	0.314193	-0.054804	0.099611	-0.094116	-0.013467
MB3	MenstrualBlood	-0.195313	0.242982	0.079676	0.051568	-0.015129	-0.065641

    - Once completed for all samples, re-open the .design file using a text editor (BBedit) and save as .design for correct formatting.

    Now the AMOVA may be run as follows:

```
amova(phylip=body_fluid_subsample_gg.an.braycurtis.0.05.lt.dist,
    design=body_fluid_clean.design, sets=all, iters=5000)
```

26. Using the above created .design file, indicative taxa may be identified as follows:

```
indicator(shared=body_fluid_subsample_gg.an.shared, design=body_fluid_clean.design)
```

Loading R Studio v3.6.1

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Download R using the regionally appropriate CRAN.

- https://cran.r-project.org/mirrors.html

Once installed, download and install R Studio.

- https://rstudio.com/products/rstudio/download/

Run required libraries for plots and figures.

- library(rgl)
- library(ggplot2)
- library(RColorBrewer)

Heatmap of Top Genera

1. For the heatmap of the top genera in all samples, first navigate to the "body_fluid_subsample_gg.summary" file.
   - Save file as excel.
   - Create new worksheet and label appropriately (Genus).

2. Label the new sheet header using Rows 1 (headers) and Rows 3 (k_Bacteria).
   - k_Bacteria is used for the total bacteria at each sample for relative abundance calculations.

3. Copy all rows of taxlevel 6 (Genus) to the new sheet and calculate the relative abundance of each taxon at each sample.
   - Unclassified taxa (taxa no longer classified at lower levels) may be grouped together.
   - The taxa may now be sorted by their relative abundance.

4. Using the top ~50 taxa, create a new worksheet. The taxa number may be increased or decreased to determine an appropriate percent cutoff.

5. Transform relative abundance of all taxa using Log10. This will enable better visualization of the relative abundance. Save Log10 transformed taxa worksheet in new workbook as .csv.

6. Proceed to R Studio.

7. Set working directory to same folder as the mothur analysis using setwd(/your/path/here).

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

8. For this example, the csv file was named “Indicator” as reflected; run code as follows:

```
#load csv file into R and save as "indicator"
indicator = read.csv("body_fluid_indicator_log10.csv",header=TRUE)
#select Indicator column
row.names (indicator) = indicator$Indicator
#counts number of columns
ncol(indicator)
#saves a matrix from column 2 to the last column
indicator1= indicator [,2:778]
#creates a data matrix and saves it as indmatrix
indmatrix = data.matrix (indicator1)
#runs the heatmap of the prepared data matrix
heatmap (indmatrix, Rowv=NA, Colv = NA, scale="none",cexRow=0.80,cexCol=0.5,margins=c(5,5),
          key=TRUE, keysize= 1.0, symkey=TRUE, trace="none", density.info="none",
         col=brewer.pal(9,"YlGnBu"))
```

PCoA Plot

9. Save PCoA .axes file as .csv.
10. Insert design column into column 2 of new PCoA .csv file and title “Origin”.
11. Load PCoA file and plot as follows:

```
#loading pcoa and saving as "pcoa_fem" ----
pcoa_fem<-read.csv("body_fluid_braycurtis_pcoa_female_nocontrols.csv", header = TRUE)
#using ggplot, create point graph
ggplot()+
  geom_point(aes(x=axis1, y=axis2, col=Origin), size=2, data = pcoa_fem)+
  xlab("Axis 1 (3.5%)")+
  ylab("Axis 2 (2.9%)") +
  theme_bw()+
  theme(axis.line = element_line(colour = "black"),
        panel.grid.minor = element_blank(),
        panel.grid.major = element_blank())+
  stat_ellipse(aes(x=pcoa_fem$axis1, y=pcoa_fem$axis2,color=pcoa_fem$Origin,
                   group=pcoa_fem$Origin),type = "norm")
```

- X- and Y-axis labels should be adjusted according to the PCoA loadings file.

Support Vector Machine (SVM) Body Fluid Classification Model

Set up Support Vector Machine (SVM) classification model to predict fluid types of new specimens.

Model Set-Up

12. Load required R libraries for Support Vector Machine (SVM) Linear classifier.

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

13. Read in training data (e.g. 20200317Final_Full.csv below); ensure correct layout (rows contain specimens, columns contain taxa count) and adjust data accordingly.
14. Read in new data to classify (e.g. ValidationFull.csv and ValidationFluidTypes.csv below); ensure correct layout (same as above) and adjust data accordingly.
15. Zero out taxa counts < 7.

```
library(kernlab)
options(warn=-1)

MyData <- read.csv(file="20200317Final_Full.csv", header=TRUE, sep=",")
ValData <- read.csv(file="ValidationFull.csv", header=TRUE, sep=",")
ValData2 <- read.csv(file="ValidationFluidTypes.csv", header=TRUE, sep=",")

# C parameter
copt <- 12.
# zero out reads < 7;
BactReads <- MyData[2:dim(MyData)[2]]
BactReads[BactReads<7] <- 0.

ValReads <- ValData[2:dim(MyData)[2]]
ValReads[ValReads<7] <- 0.

Subjects <- unlist( MyData[1] )
BodyFluidType <- substr(Subjects, 1, 2)
TaxaNames <- colnames(BactReads)

ValSubj <- unlist( ValData[1] )
ValFluidType <- unlist( ValData2[2] )
```

16. Adjust taxa counts for suspected reagent blank taxa, listed below; if relative abundance < 3%, zero out. (used absolute counts of 182 = 6046 * 3%).
17. Perform check for training data and new specimen data.

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

```
# Adjust for reagent blank taxa
# 182 <- 6046 * 3%  reagentblank taxa zeroed out if < 3%
for (try in 1:dim(BactReads)[1]){
  k <- 1 #g__Lactobacillus - 1
  if( BactReads[try,k] < 182 ){BactReads[try,k] <- 0}
  k <- 2 #g__Streptococcus - 2
  if( BactReads[try,k] < 182 ){BactReads[try,k] <- 0}
  k <- 5 #g__Gardnerella - 5
  if( BactReads[try,k] < 182 ){BactReads[try,k] <- 0}
  k <- 32 #g__.Prevotella. - 32
  if( BactReads[try,k] < 182 ){BactReads[try,k] <- 0}
  k <- 44 #g__Caulobacter - 44
  if( BactReads[try,k] < 182 ){BactReads[try,k] <- 0}
  k <- 61 #g__Hydrocarboniphaga - 61
  if( BactReads[try,k] < 182 ){BactReads[try,k] <- 0}
  }
for (try in 1:dim(ValReads)[1]){
  k <- 1
  if( ValReads[try,k] < 182 ){ValReads[try,k] <- 0}
  k <- 2
  if( ValReads[try,k] < 182 ){ValReads[try,k] <- 0}
  k <- 5
  if( ValReads[try,k] < 182 ){ValReads[try,k] <- 0}
  k <- 32
  if( ValReads[try,k] < 182 ){ValReads[try,k] <- 0}
  k <- 44
  if( ValReads[try,k] < 182 ){ValReads[try,k] <- 0}
  k <- 61
  if( ValReads[try,k] < 182 ){ValReads[try,k] <- 0}
  }
```

Set up Body Fluid Classes for Training Data

18. Separated Male vs Female Urine using specimen label.
19. Labeled Menstrual Blood (MB) and Female Urine (UF) specimens as Vaginal Fluid (VF).
20. Body Fluid Classification Types: "Bl","Fe","Sa","Se","UM","VF" correspond to Blood, Fecal Matter, Saliva, Semen, Male Urine and Vaginal Fluid (combined Menstrual Blood, Female Urine and Vaginal Fluid), respectively.

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

```
BodyFluidType2 <- BodyFluidType
BodyFluidType2[which(BodyFluidType =='Ur' & substr( Subjects, 7, 7 ) == "M" )] <- 'UM'
BodyFluidType2[which(BodyFluidType =='Ur' & substr( Subjects, 7, 7 ) == "F" )] <- 'UF'

print("Original Table of Body Fluid Counts")
print(addmargins(table(BodyFluidType2), FUN = list(Total = sum), quiet = TRUE))
print(' ')

BodyFluidType <- BodyFluidType2

# statement below lumps menstrual blood into vaginal fluid
BodyFluidType[which(BodyFluidType =='MB')] <- 'VF'
BodyFluidType[which(BodyFluidType =='UF')] <- 'VF'

# convert to factor
BodyFluidType<-factor(BodyFluidType, levels = c("Bl","Fe","Sa","Se","UM","VF"))
ValFluidType<-factor(ValFluidType, levels = c("Bl","Fe","Sa","Se","UM","VF"))

BodyFluidNames <- levels(BodyFluidType)
nfluids <- length(BodyFluidNames)

# BodyFluidNames
print("Table of Body Fluid Counts")
print(addmargins(table(BodyFluidType), FUN = list(Total = sum), quiet = TRUE))
print(' ')
```

Set up SVM Training Model, then Predict Fluid Type of New Specimens.

21. Convert absolute taxa counts to relative abundances.
22. Take square root of relative abundances for input to SVM classifier.
23. Set up model training data (e.g. mydf below) and new specimen data (e.g. valdf below).
24. Train SVM model using function ksvm from R library kernlab.
25. Predict body fluid type of new specimens using function predict.

```
# 6046 is normalized absolute count
BactProp <- BactReads / 6046.
ValProp <- ValReads / 6046.
# TAKE SQUARE ROOT OF ABUNDANCES
BactProp <- BactProp ^ 0.5
ValProp <- ValProp ^ 0.5
# set up datasets
mydf <- data.frame(BodyFluidType, BactProp )
valdf <- data.frame(BodyFluidType[1:141], ValProp )

mysvm <- ksvm(BodyFluidType ~ ., data=mydf,
              type = 'C-svc',
              kernel = 'vanilladot',
              C = copt,  kpar=list())

# predict blind validation specimens
mypred <- predict(mysvm, newdata = valdf, type="response")
```


This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Print Confusion Matrix and Metrics (Accuracy, Precision, Recall/Sensitivity, F-1 Score).

```
# CONFUSION TABLE
myconfusion <- table(factor( ValFluidType, levels=BodyFluidNames ),
                     factor( mypred, levels=BodyFluidNames ))
# metrics
cm = as.matrix( myconfusion )
accuracy <- sum( diag(cm) ) / sum( cm )
precision <- diag(cm) / colSums(cm)
recall <- diag(cm) / rowSums(cm)
f1score <-  ifelse(precision + recall == 0, 0,
                   2 * precision * recall / (precision + recall))
# print confusion matrix and metrics
print(' Blind Validation Confusion Table')
print(addmargins(myconfusion, FUN = list(Total = sum), quiet = TRUE))
accuracy
precision
recall
f1score
```

Results from Predicting Fluid Type of Blind Validation Specimens:

```
         Bl  Fe  Sa  Se  UM  VF Total
   Bl    12   0   0   1   0   0    13
   Fe     0  21   0   0   0   1    22
   Sa     0   0  25   0   0   0    25
   Se     0   0   0   3   5   4    12
   UM     0   0   0   0   4   4     8
   VF     0   0   0   0   2  59    61
   Total 12  21  25   4  11  68   141
 > accuracy
 [1] 0.8794326
 > precision
       Bl        Fe        Sa        Se        UM        VF
 1.0000000 1.0000000 1.0000000 0.7500000 0.3636364 0.8676471
 > recall
       Bl        Fe        Sa        Se        UM        VF
 0.9230769 0.9545455 1.0000000 0.2500000 0.5000000 0.9672131
 > f1score
       Bl        Fe        Sa        Se        UM        VF
 0.9600000 0.9767442 1.0000000 0.3750000 0.4210526 0.9147287
 `
```

Print Predictions for All Samples.

```
# print all predictions
print(" Actual vs. Predicted")
for (indx in 1:length(mypred)){
  print( paste( indx, ValSubj[indx], ValFluidType[indx] , mypred[indx]) )
}
```


This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

```
[1] " Actual vs. Predicted"
[1] "1 VAL002 VF VF"
[1] "2 VAL003 Sa Sa"
[1] "3 VAL004 VF VF"
[1] "4 VAL005 VF VF"
[1] "5 VAL007 VF VF"
[1] "6 VAL008 Fe Fe"
[1] "7 VAL009 Sa Sa"
[1] "8 VAL011 Se VF"
[1] "9 VAL012 VF VF"
[1] "10 VAL013 VF VF"
[1] "11 VAL014 Sa Sa"
[1] "12 VAL016 VF VF"
[1] "13 VAL017 Sa Sa"
[1] "14 VAL018 Bl Bl"
[1] "15 VAL019 VF VF"
[1] "16 VAL020 VF VF"
[1] "17 VAL022 Fe Fe"
[1] "18 VAL023 Se VF"
[1] "19 VAL024 Bl Bl"
[1] "20 VAL025 VF VF"
[1] "21 VAL026 Fe VF"
[1] "22 VAL028 Bl Bl"
[1] "23 VAL029 Fe Fe"
[1] "24 VAL030 VF VF"
[1] "25 VAL031 VF VF"
[1] "26 VAL032 VF VF"
[1] "27 VAL034 Fe Fe"
[1] "28 VAL035 VF VF"
[1] "29 VAL036 VF VF"
[1] "30 VAL038 UM VF"
[1] "31 VAL039 Fe Fe"
[1] "32 VAL040 VF VF"
[1] "33 VAL041 Sa Sa"
[1] "34 VAL042 Bl Bl"
[1] "35 VAL043 VF VF"
[1] "36 VAL045 Sa Sa"
[1] "37 VAL046 UM UM"
[1] "38 VAL047 VF VF"
[1] "39 VAL048 Bl Bl"
[1] "40 VAL049 Sa Sa"
[1] "41 VAL050 Bl Bl"
[1] "42 VAL051 VF VF"
[1] "43 VAL053 Fe Fe"
[1] "44 VAL054 VF VF"
[1] "45 VAL055 Bl Bl"
[1] "46 VAL057 VF VF"
[1] "47 VAL058 Fe Fe"
[1] "48 VAL059 Se UM"
[1] "49 VAL060 Bl Bl"
[1] "50 VAL061 VF VF"
[1] "51 VAL064 Sa Sa"
[1] "52 VAL065 VF VF"
[1] "53 VAL066 Fe Fe"
[1] "54 VAL067 Sa Sa"
[1] "55 VAL070 Fe Fe"
[1] "56 VAL071 Bl Bl"
[1] "57 VAL072 UM VF"
[1] "58 VAL073 Sa Sa"
[1] "59 VAL074 VF VF"
[1] "60 VAL075 Sa Sa"
[1] "61 VAL076 VF VF"
[1] "62 VAL077 Sa Sa"
[1] "63 VAL078 Fe Fe"
[1] "64 VAL079 Bl Bl"
[1] "65 VAL081 UM VF"
[1] "66 VAL083 VF VF"
[1] "67 VAL085 UM UM"
[1] "68 VAL086 Fe Fe"
[1] "69 VAL089 VF VF"
[1] "70 VAL090 Sa Sa"
[1] "71 VAL093 VF VF"
[1] "72 VAL094 VF VF"
[1] "73 VAL095 Fe Fe"
[1] "74 VAL097 VF VF"
[1] "75 VAL099 UM UM"
[1] "76 VAL100 VF VF"
[1] "77 VAL101 UM UM"
[1] "78 VAL102 Fe Fe"
[1] "79 VAL105 VF VF"
[1] "80 VAL106 Fe Fe"
[1] "81 VAL108 VF VF"
[1] "82 VAL109 Sa Sa"
[1] "83 VAL113 VF VF"
[1] "84 VAL114 Se UM"
[1] "85 VAL115 VF VF"
[1] "86 VAL117 Se VF"
[1] "87 VAL118 VF VF"
[1] "88 VAL119 VF VF"
[1] "89 VAL120 VF VF"
[1] "90 VAL122 Fe Fe"
[1] "91 VAL123 VF VF"
[1] "92 VAL124 VF VF"
[1] "93 VAL125 Se UM"
[1] "94 VAL129 Sa Sa"
[1] "95 VAL130 Se UM"
[1] "96 VAL131 Bl Bl"
[1] "97 VAL132 VF VF"
[1] "98 VAL133 VF VF"
[1] "99 VAL134 VF VF"
[1] "100 VAL135 Fe Fe"
[1] "101 VAL136 Sa Sa"
[1] "102 VAL139 VF VF"
[1] "103 VAL142 VF UM"
[1] "104 VAL146 VF VF"
[1] "105 VAL148 VF VF"
[1] "106 VAL149 Sa Sa"
[1] "107 VAL150 Bl Se"
[1] "108 VAL151 Se UM"
[1] "109 VAL152 Sa Sa"
[1] "110 VAL154 VF VF"
[1] "111 VAL155 VF VF"
```

```
[1] "112 VAL157 Sa Sa"
[1] "113 VAL158 VF VF"
[1] "114 VAL159 Se Se"
[1] "115 VAL160 Fe Fe"
[1] "116 VAL161 VF VF"
[1] "117 VAL162 VF VF"
[1] "118 VAL163 Sa Sa"
[1] "119 VAL164 Se Se"
[1] "120 VAL165 VF VF"
[1] "121 VAL166 Sa Sa"
[1] "122 VAL167 VF VF"
[1] "123 VAL168 VF VF"
[1] "124 VAL169 Sa Sa"
[1] "125 VAL171 Fe Fe"
[1] "126 VAL172 VF VF"
[1] "127 VAL173 Sa Sa"
[1] "128 VAL174 Fe Fe"
[1] "129 VAL175 Se Se"
[1] "130 VAL179 Fe Fe"
[1] "131 VAL180 Bl Bl"
[1] "132 VAL181 VF VF"
[1] "133 VAL183 VF VF"
[1] "134 VAL184 Fe Fe"
[1] "135 VAL186 Sa Sa"
[1] "136 VAL188 VF UM"
[1] "137 VAL190 Sa Sa"
[1] "138 VAL191 VF VF"
[1] "139 VAL193 UM VF"
[1] "140 VAL194 Se VF"
[1] "141 VAL196 VF VF"
```

SVM Linear, c-Parameter Tuning Results

26. The c-parameter = 12 for the SVM Linear Classifier was selected by a one-time process of testing several values of c, then selecting the c-value that generated the highest over-all average accuracy rate.
27. The average accuracy rate was determined from 100 repetitions of 5-fold cross validation for each tested c-parameter.
28. Th value c = 12 was selected with the highest over-all accuracy (92.02%) over all candidate c-values.
29. The resulting confusion table and corresponding metrics are shown below.

Tuning Confusion Table

	Bl	Fe	Sa	Se	UM	VF	Total
Bl	14,879	-	136	366	111	408	15,900
Fe	-	3,796	-	-	-	4	3,800
Sa	109	-	14,391	-	-	-	14,500
Se	732	-	-	2,791	435	342	4,300
UM	252	-	124	557	1,370	1,197	3,500
VF	129	-	93	209	424	27,645	28,500
Total	16,101	3,796	14,744	3,923	2,340	29,596	70,500

Tuning Model Accuracy, Precision, Sensitivity/Recall and F-1 Scores by Body Fluid

TUNING MODEL	
Over-all Accuracy	92.02%

	Fluid Type					
Metric	Blood	Fecal	Saliva	Semen	Urine M	Vaginal Fld
Precision	92.4%	100.0%	97.6%	71.1%	58.6%	93.4%
Sensitivity/Recall	93.6%	99.9%	99.3%	64.9%	39.1%	97.0%
F-1 Score	93.0%	99.9%	98.4%	67.9%	46.9%	95.2%

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.