Final Summary: NIJ Grant 2017-NE-BX-0003, *Bioinformatic Analysis of Big Proteomic Data: A New Forensic Tool to Identify Menstrual Blood & Body Fluid Mixtures*

This summary follows the NIJ Post Award Reporting Requirements issued March 28, 2019 and is divided into the four prescribed sections 1) Purpose 2) Research Design & Methods, 3) Data Analysis & Findings and 4) Implications for Criminal Justice Policy and Practice.

Abbreviations

ABC – ammonium bicarbonate
ACN – acetonitrile
AUROC - area under receiver operating characteristic
BCA – bicinchoninic acid assay
BSA bovine serum albumin
CI – confidence Interval
CD2 – clotted day 2 venous blood
CPLC – combinatorial peptide ligand chromatography
CW2 – clotted week 2 venous blood
DDA – data-dependent acquisition
FDR – false discovery rate
IDA – Information Dependent Acquisition
k-NN - k-nearest neighbor
LC – liquid chromatography
MALDI TOF/TOF – matrix-assisted laser desorption/ionization time-of-flight/time of flight mass spectrometer
MB – menstrual blood
MS – mass spectrometry
MS/MS – tandem mass spectrometry
NP-40 – Nonidet P-40
NYC OCME – New York City Office of Chief Medical Examiner
PBS – phosphate buffered saline
PSM – peptide spectrum match
Q-TOF – quadrupole time of flight mass spectrometer
SDC – sodium deoxycholate
SDS – sodium dodecyl sulfate
SVM - support vector machine
SWATH-MSs – sequential window acquisition of all theoretical mass spectra
TCEP – (tris(2-carboxyethyl)phosphine)
UD2 – unclotted day 2 venous blood
UW2 – unclotted week 2 venous blood
VB – venous blood

1. **Purpose -** The overarching goal of this application is to employ the power of bioinformatics to harness the vast amounts of proteomic mass spectrometry body fluid data to develop a predictive model with strong statistical confidence able to identify menstrual blood and to distinguish it from venous blood, venous blood menstrual blood mixtures, as well as from venous blood mixed with other body fluids such as saliva, semen and vaginal fluid.

1.1. **Statement of Problem** The ability to distinguish menstrual from circulating blood poses distinct problems for forensic scientists. Compared to the more commonly tested forensic body fluids, blood, saliva and semen, which have easily identifiable abundant marker proteins due to the biological functions these proteins perform in their respective body fluids (e.g. hemoglobin, amylase and semenogelin, respectively), menstrual blood is a mixture of the uterine endometrium, vaginal secretions and most abundantly, blood. This poses three problems: i) identifying unique or highly enriched markers in the endometrium (especially that are distinct from those in cells and secretions from the vaginal canal, ii) consistently detecting these inherently low abundant markers in a body fluid that varies daily, and iii) demonstrating that the test is confirmatory with a bioinformatics model capable of giving a predictive value.

1.2. **Objectives:** A confirmatory test for the identification of menstrual blood has been a long-term goal of the forensic community. Such a test would improve just outcomes in our judicial system. Achieving a predictive model with strong statistical confidence that can identify menstrual blood and distinguish it from other body fluids may also help establish predictive models for other body fluid mixture deconvolutions.

2. **Research Design & Methods:** The main objective of this application was to find out if big data can be used to differentiate between menstrual blood and venous blood. To achieve this goal, a) one hundred samples of menstrual and venous blood would have to be collected from donors in order to attain a statistically significant sample size, b) samples would have to be extracted

and processed for HPLC-MS analysis for all detectable proteins, and c) the big data generated from the mass spectrometer would have to be evaluated to create a predictive model. These tasks required the expertise of two institutions: 1) the New York City Office of Chief Medical Examiner for sample collection, extraction and data processing and 2) New York University (**NYU**) for bioinformatic analysis, predictive model creation and testing. Experimental design and execution are described below.

2.1. Sample Selection & Collection – Menstrual blood and venous blood were collected from 100 volunteers. All 100 donors gave one menstrual blood sample from the second day of their period, and one venous blood sample on the same day. In addition, 24 donors gave an additional venous blood sample approximately two weeks following the second day of menses. To evaluate the possibility that clotted venous blood (typical forensic sample) and unclotted venous blood might have different proteomes, both clotted and unclotted venous blood samples were examined. All together we collected 332 samples, 100 menstrual blood and 232 venous blood samples (109 clotted and 123 unclotted). Menstrual blood was collected using menstrual cups, where donors were asked to leave the cup in place for six to eight hours of their second day of menses. Contents were then transferred into 50 mL conical tubes to be returned to the OCME. Venous blood samples were collected using a finger lance into an anticoagulant-coated microcentrifuge tube. All samples were stored at -80^{o}C.

2.2. Protein Extraction & Analysis - Menstrual blood samples were homogenized in one milliliter of phosphate buffered saline (**PBS**) and aliquoted into two milliliter tubes with the total volume noted for each sample. For routine protein extraction, samples were solubilized in 10 volumes of 50 mM ammonium bicarbonate (**ABC**) with 1% sodium deoxycholate (**SDC**). All samples were quantified using the bicinchoninic acid (**BCA**) protein assay with BSA as a standard. Samples were reduced with tris(2-carboxyethyl) phosphine, alkylated with iodoacetamide, and digested

overnight with trypsin. Following digestion, 500 ng of sample were loaded on each of the two independent C-18 HPLC columns, eluted with 60 minute 6%-40% ACN concave gradients, followed by SWATH MS.

3. Data Analysis & Findings

3.1. Cohort Demographics The cohort of donors ranged from ages 18-49. The average and median age was 28. Ethnicity of the donors (self-reported) consisted of 53% White, 19% Asian, 16% Hispanic, 7% Black and 5% mixed. Fifty-six percent of the donors described their menses as normal flow, 27% as light, 15% as heavy and two individuals preferred not to say. Thirty-nine percent of the donors used hormonal contraceptive, 39% used non-hormonal methods, and 22% preferred not to say.

3.2. Library Generation:

3.3. A menstrual/venous blood peptide library was generated in order to perform SWATH MS acquisition on all samples. The library was created from a subset of menstrual and venous blood samples using an independent data acquisition (**IDA**) method. Ten sets (a total of 60 samples) were analyzed twice. Each set consisted of one individual's menstrual blood sample and two venous blood samples - one taken two days after menses and the other taken two weeks after menses. Samples from each donor were prepared by two methods: first, using total protein extraction, and second, using a depletion method (combinatorial peptide ligand chromatography, **CPLC**) to enrich for low abundance proteins. Proteins and peptides were identified using ProteinPilot 5.0 software (Sciex) searched against a publicly available database of human proteins (UniProt Human database, April 2019). The SWATH library was generated with both menstrual and venous blood samples from a total of 134 IDA runs. The library consisted of 4,584 peptides and 545 proteins. Library underwent stringent refinement and

filtering ensuring high-quality sequence matches, allowing for a less conservative false discovery rate (**FDR**).

3.4. Data Processing: All 332 samples, 100 menstrual blood and 232 venous blood (123 unclotted and 109 clotted) were processed using an XIC window of 10 minutes, and an XIC width of 75 ppm. XIC 10 was run using the library we built with ProteinPilot 5.0. Number of peptides per proteins was set to one thousand, number of transitions per peptide was set to six, peptide confidence interval to 95% and the FDR threshold was eliminated.

3.5. Predictive Model to Distinguish Menstrual & Venous Blood

Four supervised learning algorithms, k-nearest neighbor (**k-NN**), logistic regression, support vector machine (**SVM**) and decision trees were used on the IDA data and compared to assess their distinguishing performance in building predictive models. These four models were trained and evaluated using ten iterations of five-fold cross validation. For the analysis, all runs of both venous and menstrual blood from single donors were always partitioned together. Mean classification performance metrics for the ten iterations of cross-validation on the four models were calculated. The gradient boosted tree models (XGBoost, **XGB**) gave the highest scores for all the metrics (all > 0.9). The gradient decision tree model also gave the highest score for sensitivity and specificity with a mean area under the curve of 0.992. The gradient decision tree algorithm gave the best performance and was selected to build the predictive model with all available SWATH data. This algorithm combines multiple decision trees to create a stronger model and is well-suited for data with highly correlated variables. It was implemented with XGBoost in R software.

Samples were assigned into training sets and testing sets based on random assignment of subjects, such that both samples (menstrual blood and venous blood during menses) from a

single individual will be included in either the training or testing set according to a randomization procedure.

Protein level area without FDR filtering was log10 transformed. Model training and evaluation was carried out in five different settings using a subset: all samples, menstrual blood and clotted venous day 2 and week 2 (MB, CD2, and CW2), menstrual and unclotted venous (MB, UD2 and CW2), menstrual and day two venous (MB, CD2 and UD2), and menstrual and week two venous (MB, CW2 and UW2). For each setting, samples were divided into five folds based on subject ID, such that all blood samples (menstrual or venous) collected from one subject was assigned to only one of the folds. Machine learning models were built to predict menstrual blood versus venous blood. All models were evaluated with five-fold cross-validation. The mean and standard error of area under receiver operating characteristic (AUROC) which represents the relation between sensitivity and specificity, and area under precision-call curve (AUPR), which measure the information retrieval in the positive class (in this case menstrual blood), were calculated across the five folds to evaluate model performance.

Two types of models, gradient boosting trees (XGBoost) and random forests were used to perform the binary classification tasks. Models on all datasets were built with the following hyper-parameters: XGBoost, max depth=6, eta=0.2, n rounds=10; random forests, n tree=101. Other hyper-parameters were set to default of the R implementation (packages 'XGBoost' and 'random forests'). Feature importance was assessed in both models by averaging the importance score (average gain for XGBoost and average decreased mean accuracy for random forests) for each feature across the five cross-validation models.

3.6. Limits of Detection & Analysis – Menstrual blood and venous blood samples from two individuals were used to make serial menstrual blood:venous blood dilutions (1:2.5, 1:5, 1:10, 1:100, 1:1000) in order to determine the minimum MS data necessary to accurately predict a

menstrual blood or venous blood sample. Sample preparation and mass spectrometry analysis was performed, and the approximate limit of detection was established to be 1:5. The data were analyzed by identifying if the important features in the predictive model are identified in the dilutions. An intensity vs dilution plot was to see if the important proteins from the predictive model were present in the data from these dilutions. To evaluate the predictive power of the proteomic data at different sample concentrations, full model XGBoost and random forests were built using all samples except for the ones collected from the subjects that were used to make the serial dilution samples. Proteins that ranked higher than twenty in the importance table in both the XGBoost and random forests models were selected as the top features. Exemplary important features (sp|P05109|S10A8_HUMAN, sp|P62805|H4_HUMAN and sp|P13646|K1C13_HUMAN) were detected in serial dilutions.

3.7. Models for Mixture Prediction - Using the limit of detection (1:5) a series of mixtures (1:1, 1:5, 5:1) of venous blood and menstrual blood were created in order to determine if a classification model will correctly identify the major components, and to what extent the relative amounts present in the mixture can be distinguished. In addition, a series of mixtures of venous blood and semen, venous blood and saliva, and venous blood and vaginal fluid were created to determine if the presence of semen, saliva, or vaginal fluid affect prediction accuracy. Sample preparation and mass spectrometry analysis were performed as with all other samples. Samples were run in triplicate. Gradient boosting trees (XGBoost) and random forests models were constructed with all mixture data to predict whether menstrual blood was present in the sample, regardless of what else it might be in the mixture. Both the boosted tree and random forest models performed well, with mean AUPR and AUROC > 0.95. In addition, linear regression was performed between protein level intensity (log transformed) and the mix concentration (1:1, 1:5, 5:1) of the venous blood and menstrual blood mixtures. A few proteins with high R2

values (>0.5) can be used to estimate the mixture ratio of venous blood and menstrual blood mixtures.

3.8. Conclusions – The major goals of this work were to i) determine if bioinformatic analysis of big proteomics mass spectrometry data can be used to identify body fluids and mixtures, and, if so, ii) to identify the limit of detection of these differences. The preliminary results suggested that proteomic quantitation may be utilized to accurately distinguish menstrual from venous blood, regardless of coagulation status and time of collection. Based on model performance, clotted venous blood is modestly more distinguishable from menstrual blood than unclotted venous blood (random forests AUROC: clotted 0.99±0.004, unclotted 0.968±0.013), but the difference is not statistically significant. Archetypal important features (sp|P05109|S10A8_HUMAN, sp|P62805|H4_HUMAN and sp|P13646|K1C13_HUMAN) were detected in serial dilutions. A 1:10 dilution shows a stronger separation in the model predicting menstrual blood vs venous blood, specifically for protein sp|P13646|K1C13_HUMAN). The other two important feature proteins have less distinguishing power on their own between the dilutions.

4. Implications for Criminal Justice Policy & Practice – The ability to identify menstrual blood has important implications in the criminal justice system where blood stains at a crime scene may be ascribed to be a female victim's period or where a violent sexual assault with vaginal trauma may be claimed as consensual intercourse with a woman during menses. A confirmatory test for the identification of menstrual blood has been a long-term goal of the forensic community. Such a test would improve just outcomes in our judicial system. These data also suggest that mixtures of other body fluids, e.g. semen and saliva, may also be deconvoluted using similar models.