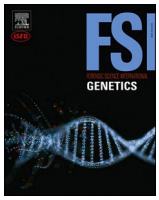
Forensic Science International: Genetics 81 (2026) 103343

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen



An Integrated proteomic workflow for body fluid classification and single amino acid variant identification: Advancing towards body fluid source attribution

Thomas P. Shehata [a,b,c,1], Shirin Alex [d,1,*], Stijn N.C. van Lierop [d], Maarten J. Blom [a], Jantine van de Wetering -Tieleman [e], Nadine Prust [e], Jeroen Demmers [e], Marcel de Puit [d,f]

[a] University of Amsterdam, Spui 21, Amsterdam 1012 WX, the Netherlands
[b] Vrije Universiteit Amsterdam, De Boelelaan 1105, Amsterdam 1081 HV, the Netherlands
[c] Wildlife Forensic Academy, R27, West Coast 7345, South Africa
[d] Netherlands Forensic Institute, Laan van Ypenburg 6, The Hague 2497 GB, the Netherlands
[e] Proteomics Centre, Erasmus MC, University Medical Center, Wytemaweg 80, Rotterdam 3015 CN, the Netherlands
[f] Technische Universiteit Delft, Mekelweg 5, Delft 2628 CD, the Netherlands

ARTICLE INFO

Keywords:
Mass spectrometry
Proteomics
Body fluid classification
Single amino acid variants
Forensic serology

ABSTRACT

A particularly challenging subject in the investigation of forensic human biological traces is analyzing samples containing mixtures of body fluids from multiple donors. Ideally, researchers want to identify each type of body fluid present. However, traditional methods, like mRNA and DNA profiling, often struggle with sensitivity, specificity, and efficiency, especially in complex mixtures. This proof-of-concept study has two primary aims: first, to classify body fluids within a mixture using discriminatory protein markers, and second, to evaluate the feasibility of using single amino acid variants (SAAVs) to trace the source of specific body fluids back to individual donors. To achieve this, we employed proteomic analysis via liquid chromatography-tandem mass spectrometry (LC-MS/MS) in data-independent acquisition (DIA) mode, developing a reliable approach for accurate body fluid classification. Through comprehensive proteomic profiling, we characterized a diverse array of discriminatory proteins present in peripheral blood, semen, saliva, urine, and vaginal fluid. Using advanced data analysis techniques, including t-distributed stochastic neighbor embedding (t-SNE), we demonstrated that these proteins could reliably distinguish between different body fluids, even in mixed samples. Additionally, our findings reveal that SAAVs within certain proteins, such as those in saliva, hold promise for source attribution in a forensic context. Challenges, including contamination and limited sample sizes, highlighted the need for strict quality controls and further large-scale studies. With these improvements, proteomic analysis could greatly enhance body fluid identification, classification, and source attribution in forensic investigations, improving both accuracy and reliability in forensic science.

1. Introduction

Various types of body fluids may be encountered at a crime scene, with blood, saliva, vaginal fluid, and semen as the most tested for [1]. Biological stains are a major form of forensic evidence, and various types of information can be provided by these stains, such as a deoxyribonucleic acid (DNA) profile to attribute individuals as donors of stains, or a deposition pattern with the purpose of drawing inferences about the nature and timing of a crime [2–5]. Moreover, body fluid classification is a key component in forensic investigations as the ability to identify and classify them can be particularly important in cases of sexual assault [3,6–11]. In these cases, associating specific body fluids with a donor, especially in a mixture is often equally or even more crucial.

For many years, body fluid identification has been conducted both at crime scenes and in laboratories using a range of methods. These include (bio)chemical-based tests, such as Rapid Stain Identification Series assays as well as enzyme-linked immunosorbent assays, which exhibit

* Corresponding author.
E-mail address: s.alex@nfi.nl (S. Alex).
[1] These authors contributed equally to this work

https://doi.org/10.1016/j.fsigen.2025.103343
Received 17 March 2025; Received in revised form 11 August 2025; Accepted 18 August 2025
Available online 19 August 2025

limited confirmatory power for body fluid identification and can only be used presumptively as such [2–4,12]. Also, more advanced techniques that are based on spectroscopic (e.g., Raman spectroscopy [13–15] and Fourier transform infrared spectroscopy [16–21]), ribonucleic acid (RNA) [e.g., 1,5,21–23, epigenetic [e.g., 24–26], or microbial markers [e.g., 27–34] have been developed. These advanced techniques offer varying levels of sensitivity, reproducibility, speed, cost, applicability to body fluid mixtures, and mostly specificity [2,35].

When dealing with a mixture of body fluids from a crime scene, determining the types of fluids present and their origins is crucial. Currently, messenger RNA (mRNA) profiling via endpoint reverse transcription polymerase chain reaction addresses the former, albeit with limitations of cost, specificity, and relative instability in degraded samples [3,6,36].

Currently, identifying the origin of each body fluid typically relies on DNA profiling. However, this approach faces challenges when DNA typing of mixed body fluids produces mixed profiles, making it difficult to associate specific fluids to individual donors. In practice, reliable source attribution of body fluids is only feasible when the mixed DNA profile indicates a single male and female donor, and the mixture contains sex-specific body fluids that can be identified by performing a differential extraction [37]. In these cases, linking a donor to a specific body fluid is possible when considering sex-specific body fluids and the presumed involvement of a donor of that sex. This method, however, risks an association fallacy [5,38,39].

A concept to overcome these problems is to both make use of genetic markers for body fluids and the genetic variation within these markers. Such an approach was recently suggested by Neis et al. (2024) and highlighted in the review by Keane et al. (2024) [40,41]. However, this approach would still rely on the availability of non-degraded tissular genetic markers, which may not always be available in casework samples. A promising alternative is a proteomics-based strategy, which can overcome some of the limitations of mRNA and DNA profiling, especially in complex mixtures. Over the last few years, proteomics has provided answers to forensically relevant questions regarding identification and classification of type(s) of body fluid(s) in a pristine or degraded sample, leveraging the fact that not all genes are expressed in every cell type [42]. It has also demonstrated to be able to individualize hair shafts [36,43–47], bones [48], and fingermarks [49].

This proteomics-based method utilizes single amino acid variants (SAAVs) in proteins, which are induced by non-synonymous single nucleotide polymorphisms (nsSNPs). Because proteins encode genetic information through these SAAVs, their detection can contribute to individualization by offering insight into ancestry, disease susceptibility, or potentially even identity (Fig. 1). Building on foundational work by Danielson, Legg, McKiernan, Parker, and others [4,12,36,41,43,45–58], this study explores a proteomics-based approach to identify body fluids in mixtures, using specific protein markers for accurate classification and the utility of genetically variant peptides (GVPs) which contain SAAVs within the forensic context of body fluid source attribution. This approach is particularly valuable when DNA and mRNA analyses are inconclusive or unavailable. As the field advances, it may become possible to infer specific SNP alleles within protein-coding genes that are associated with each body fluid type. The combination of these proteomically-inferred nsSNPs could ultimately allow for the construction of a genetic profile, analogous to SNP or short tandem repeat (STR) genotypes, and usable in calculating random match probabilities (RMPs) [44,59].

Accordingly, this proof-of-concept study aims to both classify body fluids in mixed samples and detect SAAVs within the body fluid proteome using liquid chromatography tandem mass spectrometry (LC-MS/MS).

2. Materials and methods

2.1. Sample collection

In this study, biological samples were obtained from a cohort of 14 donors, comprising 7 males and 7 females (Fig. 2). The age range of the donors fell between 23 and 63 years (mean age, x_{avg} = 34) (Supplementary information 1).

Prior to sample collection, informed consent was obtained from all volunteers through a detailed Informed Consent form and after consultation with a privacy officer of the Netherlands Forensic Institute. The form outlined the study's objectives, the intended use and storage of biological samples and associated data.

Saliva was collected by instructing volunteers to rinse their mouths thoroughly with water and then spit through a sterile 5 mL pipette tip into a sterile Protein LoBind tube (Eppendorf, Hamburg, Germany). For morning urine collection, donors were provided with sterile urine pots and instructed on mid-stream collection. To obtain vaginal fluid samples, volunteers utilized a sterile self-sampling device called Evalyn Brush (Rovers Medical Devices, Oss, The Netherlands). This brush was weighed before and after sampling so that the weight of the contents could be determined. The collected sample was then transferred into a sterile collection tube using plastic tweezers and centrifuged at maximum speed for 2 min to remove the content from the brush. Peripheral blood samples were collected following a venepuncture

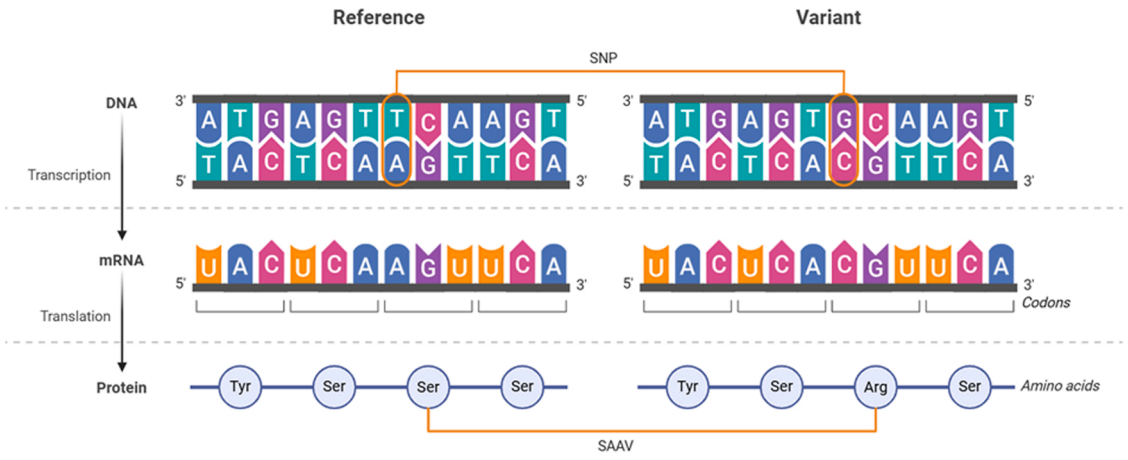


Fig. 1. Genetic Implications of Non-Synonymous Single Nucleotide Polymorphisms (nsSNPs). Substitution of a single nucleotide leads to altered codon assignment and amino acid sequence, exemplified by the conversion from serine (Ser) to arginine (Arg). Detection via proteomic mass spectrometry and *de novo* sequencing is possible due to the resulting shifts in peptide mass and fragmentation spectrum, facilitating inference of nsSNPs. This figure was made using the BioRender software platform.

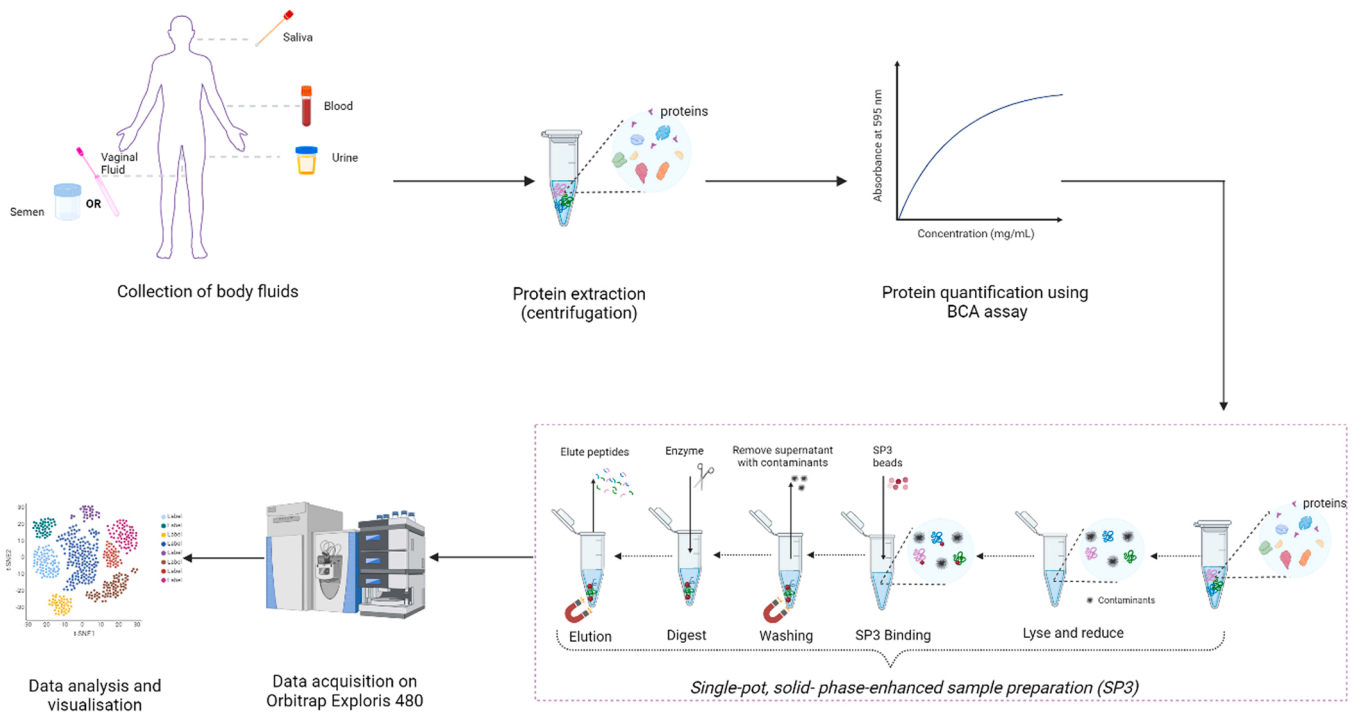


Fig. 2. A simplified depiction of the method workflow. This diagram presents the method employed in this proof-of-concept study, spanning from sample collection to single-pot, solid-phase-enhanced sample preparation, to liquid chromatography tandem mass spectrometry, and finally to data analysis.

procedure. These samples were collected in K2 ethylenediaminetetraacetic acid (EDTA) collection tubes (Becton Dickinson, Plymouth, UK). Semen samples were obtained through a process of automanipulation.

Donors providing urine or vaginal fluid were advised to refrain from engaging in sexual intercourse for a minimum of one day before collecting these body fluids. For volunteers providing seminal fluid, it was recommended to abstain from ejaculation for a period of at least one day before the sampling. In compliance with protocol, individuals providing DNA and saliva samples were advised to abstain from food and beverage consumption for a period of 30 min before the swabbing procedure. Additionally, they were directed to perform a thorough mouth rinse with water prior to the commencement of sample collection.

All collected samples were promptly preserved at a temperature of −25 °C for storage to maintain sample integrity.

2.2. Protein quantification

The protein concentration in the body fluids was assessed using a Thermo Scientific™ Pierce™ bicinchoninic acid (BCA) assay (Thermo Fisher Scientific Inc., Waltham, USA): peripheral blood, saliva, semen, urine, and vaginal fluid samples were diluted 100-fold with Milli-Q (MQ) water to ensure they fell within the working range of the assay (20–2000 µg/mL). Additionally, calibration standards ranging from 0 to 2 mg/mL protein were prepared using a solution of 2.0 mg/mL bovine serum albumin in 0.9 % aqueous sodium chloride containing sodium azide.

Following sample preparation, 2 mL of a 50:1 Pierce™ BCA Protein Assay reagent A to reagent B solution was added to 100 µL of each solution and incubated at 37 °C for 30 min. Subsequently, absorption measurements were taken at 562 nm using a Perkin-Elmer Lambda 35 UV–visible spectrophotometer (PerkinElmer, Inc., Shelton, USA) after chilling the samples to room temperature.

Following the absorption measurements at 562 nm, the absorbance intensity of both the samples and the calibration standards were compared, forming the basis for calculating the protein concentration in the samples (Supplementary information 2). It should be noted that the average protein concentration measured for semen (16.8 ± 6.47 mg/mL) was lower than values typically reported in the literature (35–55 mg/mL) for seminal plasma [60]. This discrepancy may be attributed to multiple factors, including the use of whole semen rather than isolated seminal plasma, absence of a matrix-matched internal control during the BCA assay, and potential matrix effects or sample variability. Future studies could benefit from incorporating a semen-specific positive control or additional sample preparation steps (e.g., centrifugation or desalting) to ensure more accurate protein quantification.

The obtained averages were subsequently utilized to determine appropriate ratios for the body fluid mixtures, as detailed in Table 1. Biological replicates were performed (see sample list in Supplementary information 3). As some body fluids are more heavily protein-concentrated, the volume ratios of body fluids in the mixtures were chosen to maximize the detection of fluid-specific proteins, with both body fluids contributing equally to the protein content in the mixture [2]. Although these ratios may differ from real-world forensic samples, they provide a controlled framework to evaluate the discriminatory power of the proteomic markers.

Table 1
Mixture ratios (volume) for the formation of body fluid mixtures. Body fluid mixtures were made utilizing protein content-based body fluid proportions.

Mix ratio (volume)	Saliva	Urine	Peripheral blood	Semen	Vaginal fluid
Saliva	-				
Urine	1.00: 3.63	-			
Peripheral blood	1.00: 69.0	1.00: 19.0	-		
Semen	1.00: 23.0	1.00: 6.33	3.00: 1.00	-	
Vaginal fluid	1.00: 23.0	1.00: 6.33	3.00: 1.00	1.00: 1.00	-

2.3. Single-pot, solid-phase-enhanced sample preparation

After determining the protein concentration in the various body fluids, an adjusted single-pot, solid-phase-enhanced sample preparation (SP3) method according to Hughes *et al.* (2019) was conducted [61]. Pure body fluids were diluted with a lysis buffer, termed the SP3 reconstitution solution, comprising 50 mM HEPES, pH 8 (Sigma Aldrich, Saint Louis, USA), 1 % (wt/vol) sodium dodecyl sulphate (Acros Organics, Geel, Belgium), 1 % (vol/vol) Triton X-100 (Acros Organics, Geel, Belgium), 1 % (vol/vol) NP-40 (Sigma Aldrich, Saint Louis, USA), 1 % (vol/vol) Tween 20 (Sigma Aldrich, Saint Louis, USA), 1 % (wt/vol) sodium deoxycholic acid (SDC) (Sigma Aldrich, Saint Louis, USA), 5 mM EDTA disodium salt (Sigma Aldrich, Saint Louis, USA), 50 mM NaCl (Sigma Aldrich, Saint Louis, USA), and 1 % (vol/vol) glycerol (Merck Schuchardt OHG, Hohenbrunn, Germany) in MQ water. This yielded a 100 µL solution containing 50 µg of protein for saliva, urine, and mixtures of these body fluids, and 100 µg for peripheral blood, semen, vaginal fluid, and mixtures containing these body fluids.

The samples underwent sonication for 10 min followed by heat inactivation at 90 °C for 10 min and centrifugation at 150 × g for 1 min. Proteins were then cooled to room temperature and reduced by incubation at 50 °C and 1100 rpm for 25 min after adding 5 µL of 100 mM 1,4-dithiothreitol (DTT) (Roche Diagnostics GmbH, Mannheim, Germany). Alkylation was performed by incubating for 20 min at room temperature in the dark after adding 5 µL of 200 mM iodoacetamide (IAA) (Sigma Aldrich, Saint Louis, USA). The alkylation reaction was quenched by adding another 5 µL of 100 mM DTT and incubating for 10 min at room temperature.

2.4. SP3 protein clean-up and digestion

Following reduction and alkylation, a sample volume containing 10 µg of protein was diluted to a final volume of 48 µL in SP3 reconstitution solution. Sera-Mag SpeedBeads Protein A/G (20 mg/mL in sodium azide) (Sigma Aldrich, Saint Louis, USA) were washed three times with 200 µL MQ water on a magnetic rack and reconstituted in MQ water to a final concentration of 50 mg/mL. 2 µL of this solution was pipette-mixed into the sample tubes, resulting in a final volume of 50 µL. To induce protein binding to the beads, 50 µL ethanol (EtOH) (Avantor Performance Materials, Gliwice, Poland) was added, followed by gentle vortexing and incubation at 24 °C, 1000 rpm for 5 min. After incubation, the tubes were transferred onto a magnetic rack to allow the beads to migrate to the tube wall, and the unbound supernatant was removed and discarded. Subsequently, the beads were washed three times by pipette-mixing with 180 µL 80 % EtOH and discarding the supernatant. Prior to digestion, 95 µL 100 mM ammonium bicarbonate (ABC) (Sigma Aldrich, Saint Louis, USA) and 5 µL 10 % SDC were added. 20 mg of Pierce MS-Grade Trypsin/Lys-C Protease Mix (Thermo Fisher Scientific, Waltham, USA) was dissolved in 500 µL 100 mM ABC, and 10 µL of the enzyme solution was then pipette-mixed into the sample tubes and sonicated for 1 min. Incubation occurred overnight at 37 °C, 1000 rpm.

2.5. Peptide purification and concentration

After digestion, the sample tubes were cooled back to room temperature and centrifuged at 16,000 × g for 1 min. The tubes were subsequently transferred onto a magnetic rack, and after the beads settled onto the tube wall, the supernatant was transferred into a fresh Protein LoBind tube. The beads were washed once with 30 µL 100 mM ABC, and the supernatant was added to the earlier collected supernatant. 10 µL of 20 % trifluoroacetic acid (TFA) (Sigma Aldrich, Saint Louis, USA) was added to the tubes to precipitate SDC, followed by centrifugation at 12,500 × g for 15 min. The supernatant was subsequently transferred into a fresh Protein LoBind tube and vacuum-dried for 1.5 h at 45 °C using Eppendorf Concentrator Plus (Eppendorf, Hamburg, Germany). The vacuum-dried samples were stored under −25 °C until desalting and further analysis.

The vacuum-dried samples were reconstituted by adding 100 µL 0.1 % TFA. Following this, the samples underwent vortexing, a 5-minute sonication, and centrifugation for 1 min at 650 × g. Subsequently, desalting of the samples was performed using a StageTip equipped with two Empore octadecyl C18 47 mm extraction disks (Supelco, Bellefonte, USA). The StageTip was initially wetted with 100 µL acetonitrile (ACN) (Sigma Aldrich, Saint Louis, USA), followed by a 10-minute centrifugation at 300 × g. It was then conditioned with 100 µL 0.1 % TFA and centrifuged for 10 min at 300 × g. Next, the digested sample was loaded onto the StageTip and centrifuged for 10 min at 300 × g. After reloading the tryptic digest onto the StageTip, it was centrifuged again for 10 min at 300 × g. Subsequently, the StageTip was washed with 100 µL 0.1 % TFA and centrifuged at 300 × g for 10 min. Finally, the peptides were eluted twice with 50 µL of 70 % ACN in 0.5 % formic acid (FA) (Sigma Aldrich, Saint Louis, USA), followed by centrifugation for 8 min at 300 × g. Following this, the eluates underwent another round of vacuum drying, after which the peptides were resuspended in 25 µL of 2 % ACN in 0.5 % FA.

2.6. LC-MS/MS analysis

1 µL of the peptide mixtures – corresponding to an injected mass of approximately 0.4 µg of peptide digest for all pure body fluids and mixtures – were trapped on a 2 cm× 100 µm Pepmap C18 column (Thermo Fisher Scientific, Waltham, USA) and separated on a 30 cm× 75 µm capillary column packed in-house with 1.9 µm ReproSil-Pur C18 beads (Dr. Maisch, Ammerbuch, Germany). This separation was performed at a flow rate of 275 nL/min using an EASY-nLC 1200 (Thermo Fisher Scientific, Waltham, USA) with a linear gradient of 0 – 40 % ACN in 0.1 % FA over 90 min. The eluate was directly introduced into the mass spectrometer via nanospray ionization. To ensure instrument performance and monitor carry-over, a HeLa digest (Thermo Fisher Scientific, Waltham, USA) was used as a system suitability standard, analysed before, between, and after samples. Reagent blanks consisted of 4 % ACN with 0.5 % FA, which was run between each sample to minimize carry-over effects. In cases where additional cleaning was required, a stronger wash, 75 % ACN, was applied. Data for proteomics were acquired on an Orbitrap Exploris 480 mass spectrometer (Thermo Fisher Scientific, Waltham, USA) in data-independent acquisition (DIA) mode. The mass spectrometer operated in positive mode with Tune version 4.2. Precursors were selected in the quadrupole with an isolation window specified in Supplementary information 4 and fragmented with higher-energy collisional dissociation (HCD) using 30 % collision energy. Full MS1 scans were recorded in the range of 175 – 2000 m/z at a resolution of 15,000. MS2 spectra were recorded in profile mode in the Orbitrap at 15,000 resolution with standard automatic gain control. The maximum injection time was set to automatic.

2.7. Data analysis

2.7.1. Protein identification

The DIA raw files, accessible via ProteomeXchange with identifier PXD061543, were converted into the Spectronaut file format HTRMS and then analysed in Spectronaut 18 (v 18.6.231227.55695, Biognosys AG, Schlieren, Switzerland) with default settings. Theoretical m/z values for fragment ions and precursors were used in all spectral libraries. Protein inference, which gave rise to the protein groups, was performed on the principle of parsimony using the ID picker algorithm as implemented in Spectronaut [62].

The settings for DIA data analysis in Spectronaut included peak detection with precision iRT, a correction factor of 1, enabled interference correction on the MS2 level, and disabled cross-run normalization. During the search, one fixed modification was specified, which was carbamidomethylation of cysteine, along with two variable modifications: acetylation at the protein N-terminus and oxidation of

methionine. The maximum allowable number of variable modifications was set to 5. Trypsin was chosen as the enzyme, with a maximum allowance of 2 missed cleavages. Additionally, the minimal and maximal peptide length were established at 7 and 52, respectively [63,64]. The minimal and maximal fragment (m/z) were set at 200 and 3000 m/z, respectively. Peptides were identified via automated database searching of the generated data using Spectronaut against the human UniProt database (v release-2023_02, The UniProt Consortium, Geneva, Switzerland). While certain authors propose that a solitary specific peptide suffices for protein identification [65], others advocate for a "two-peptide rule" [66], while still others advocate for a three-peptide requirement [67]. To mitigate the false positive rate, in this study, a protein was deemed identified if it attained a protein confidence level of $\geq$ 99.0 % with at least three different protein-specific peptides, each exhibiting a confidence level of $\geq$ 99.0 %. Additionally, peptide and protein identifications were accepted if they achieved a false discovery rate of less than 1.0 %.

2.7.2. Body fluid classification

After identifying the proteins, further data analysis was performed using Python 3.10 and the Numpy [68] scikit-learn [69] and pandas libraries [70]. Data visualization was done using the matplotlib [71], plotly [72] and seaborn libraries [73]. First, a t-distributed stochastic neighbor embedding (t-SNE) was conducted to gauge the discriminatory capacity of the dataset. A nonlinear t-SNE was favored over a linear principal component analysis (PCA), primarily due to the intricate and nonlinear relationships typically observed between proteins and body fluids in proteomic data [74]. t-SNE analysis was performed using the following default settings: perplexity set to 30.0, learning rate auto, number of iterations 1000, Euclidean distance metric, output dimensionality 2, random state 42.0, early exaggeration 12.0, and PCA initialization.

To get an idea of how the proteome can be utilized for classifying body fluids, the proteins with the highest discriminatory power between all body fluids in pure samples were pinpointed using the Gini impurity value. The Gini impurity quantifies the likelihood of misclassifying a randomly selected element from a set if it were labelled randomly and independently, based on the distribution of labels within the set (Eq. 1). It achieves its minimum value of zero when all instances within the node belong to a single target category. In this case, this entails that the value would be minimum for a protein that only occurs in one type of body fluid and not in others. Afterwards, the top 10 proteins with lowest Gini impurity, highest relative occurrence in a certain body fluid, and highest abundance for each body fluid, respectively, were collated. Additionally, the occurrence of these proteins in mixtures containing the respective body fluid was documented.

$$Gini(D) = 1 - \sum_{i=1}^{k} p_i^2 \tag{1}$$

Eq. 1. **The Gini equation.** The Gini equation considers a dataset D that contains samples from k classes. The probability of the samples belonging to class i at a given node can be denoted as p_i. Then, the Gini impurity of D can be defined.

Due to the limited sample size, it was not feasible to employ a multilabel classification model for classifying the encountered body fluids within the mixtures. Alternatively, proteins exhibiting the highest discriminatory power (i.e., Gini impurity equal to zero) were compared to those identified in the mixtures, following the application of the one, two, and three-peptide rule. These rules were selected for comparison due to the possibility that peptides from a highly abundant protein in one body fluid may suppress peptides from less abundant proteins (possessing high discriminatory power) of another body fluid within the mixture. Subsequently, the two body fluids with the highest relative occurrence of identified discriminatory proteins in the mixture were designated as the 'predicted' body fluids. For this approach, it was assumed that the mixture indeed contained two distinct body fluids. Recall (Eq. 2), precision (Eq. 3), support, and an F1-score (Eq. 4) were computed for this method.

$$Recall = \frac{True\ positive}{True\ positive + False\ negative} \tag{2}$$

Eq. 2. **The equation for the recall (sensitivity).** Recall measures the proportion of true positive predictions among all actual positive instances in the dataset.

$$Precision = \frac{True\ positive}{True\ positive + False\ positive} \tag{3}$$

Eq. 3. **The equation for the precision.** Precision measures the proportion of true positive predictions among all positive predictions made by the model.

$$F1score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

Eq. 4. **The general form for the F1 score.** The F1-score is the harmonic mean of precision and recall.

2.7.3. Source attribution of body fluids

SAAVs were identified by querying a custom-built database that integrates variant data from the variant viewer accessible via UniProt. Only variants associated with proteins exhibiting a Gini impurity score of zero were included in the database, as outlined in Supplementary information 5. The search settings were similar to those outlined in *2.7.1 Protein identification*. There are practical chemical constraints on the effective number of usable genetically variant peptides (GVPs) for source attribution. Firstly, certain variants share identical or similar masses, such as I → L, K → Q, and *vice versa*. Secondly, chemical modifications or variants with equivalent mass shifts to genetic variants must be excluded. These include N → D (deamidation), Q → E (deamidation), M → F (oxidation), P → I or L (oxidation), C → S (oxidation) and K → R (formylation) [48,54,75–77]. After the database search was performed, SAAVs identified in the mixtures were compared to those identified in the pure body fluids in the respective body fluid class. The individual providing the pure body fluid with the highest similarity of variants to those in the mixture was designated as the donor, allowing for an accuracy assessment. Subsequently, the allele frequencies in a European population from dbSNP (National Center for Biotechnology Information, Bethesda, USA) of the nsSNPs corresponding to SAAVs with the lowest Gini impurity score were collated to give an indication of the power of discrimination [78].

3. Results and discussion

3.1. Protein identification

In total, 39,540 different peptides were identified across all pure samples. Following a database search utilizing these peptides and applying the "three-peptide rule", 3856 different proteins were identified [67]. The distribution of identified proteins per body fluid is depicted in Fig. 3. The numbers presented in this figure do not account for the possibility that the presence of specific proteins in a particular unmixed body fluid might be attributed to traces of another body fluid. For instance, considering the presence of traces of semen or vaginal fluid in urine, and *vice versa*, could lead to such occurrences, which could also be observed in forensic samples.

Contamination by polyethylene glycol (PEG), $(\text{-}CH_2\text{-}CH_2\text{-}O\text{-})_n$, was detected in the mass spectra of 21 out of 70 pure and mixed samples, evidenced by the observed + 44 Da repeat (Supplementary Information 6). Possible sources of this contamination include PEG leaching from the plastic Eppendorf tubes or collection pots used or the inadvertent use of highly polymerized forms of Triton X-100, NP-40, or Tween-20 during the SP3 sample preparation. The latter explanation is considered the

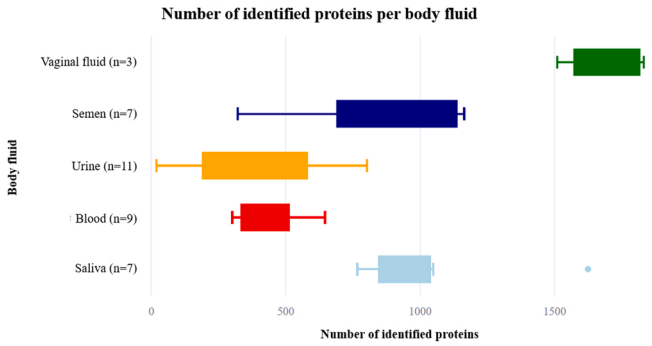


Fig. 3. A box plot of the number of proteins identified in body fluids. This box plot depicts the number of proteins identified in five different body fluids with *n* samples: peripheral blood (red), semen (dark blue), saliva (light blue), urine (yellow), and vaginal fluid (green). A protein was marked as identified if at least 3 different peptides were identified.

most probable, given that the same Eppendorf tubes were used for all samples. This contamination may have led to the suppression of signals from certain (low abundant) peptides. Consequently, it was hypothesized that contaminated samples would result in the identification of fewer proteins overall. As a result, the decision was made to refrain from filtering out proteins based on signal intensity. This decision assumed that the likelihood of identifying lower abundant proteins in all samples from the same body fluid would be low. Consequently, the probability of these lower abundant proteins emerging as proteins with high discriminatory power was also deemed low, implying that these proteins would not significantly contribute to the identification of body fluids in the final analysis.

3.2. Identification of discriminatory proteins

The analysis of peripheral blood identified a set of 17 proteins with a Gini impurity score of 0 (Table 2), indicating their significant discriminatory capability in distinguishing this body fluid. For saliva, urine, seminal fluid, and vaginal fluid, the corresponding numbers of proteins with a Gini impurity score of 0 were 75, 79, 282, and 352, respectively. This indicates their significant discriminatory capability in distinguishing the corresponding body fluid.

For peripheral blood these proteins exhibited varying relative occurrences in pure samples of peripheral blood with *eukaryotic translation initiation factor 4B* being present in most (55.6 %) pure blood samples. Interestingly, the occurrence of these proteins in mixtures containing peripheral blood was variable, with most proteins being absent in a majority of the samples. However, with the exception of *eukaryotic translation initiation factor 4B*, these proteins showed minimal occurrence in mixtures that did not contain peripheral blood, suggesting their specificity to this fluid in this experiment.

For saliva, eight of the discriminatory proteins consistently exhibited a relative occurrence of 1 in pure saliva samples, underscoring their strong association with this body fluid. While their presence in mixtures containing saliva varied, certain proteins, such as *carbonic anhydrase 6*, *protein LEG1 homolog, deleted in malignant brain tumours 1 protein*, and *lactoperoxidase*, were consistently present in all mixtures containing saliva and absent in those without, highlighting their specificity in this research.

In the case of urine, most discriminatory proteins exhibited high occurrences in pure urine samples. Notably, their occurrence in mixtures containing urine varied, with certain proteins showing notable presence, such as *matrix remodelling-associated protein 8*, *osteopontin*, and *cubilin*. Moreover, the discriminatory proteins' occurrence in mixtures not containing urine was minimal, emphasizing their specificity to this body fluid in this study.

For seminal fluid, 18 of these proteins, including 10 highlighted in Table 2, displayed a relative occurrence of 1 in pure seminal fluid samples, indicating their exclusive association with this bodily fluid. Notably, all these proteins exhibited minimal occurrence in mixtures lacking seminal fluid, while being predominantly present in mixtures containing semen.

Finally, for vaginal fluid, 92 of the discriminatory proteins consistently exhibited a relative occurrence of 1 in pure vaginal fluid samples, indicating their exclusive association with this bodily fluid. Although their presence in vaginal fluid-containing mixtures varied, certain proteins, such as *15-hydroxyprostaglandin dehydrogenase [NAD(+)]*, and *short-chain dehydrogenase/reductase family 9 C member 7*, were notably prevalent. Moreover, most of these displayed minimal occurrence in mixtures lacking vaginal fluid, highlighting their specificity to this particular body fluid in this study.

These findings underscore the discriminatory potential of specific proteins in classifying and distinguishing different body fluids in pure and mixed samples.

In this study we successfully identified several widely recognized proteins employed for proteomics-based body fluid classification, including *semenogelin 1* and *semenogelin 2* for semen, *cystatin-SA* and *alpha-amylase 1* for saliva, and *hemoglobin subunit beta*, *hemoglobin subunit alpha 1*, and *hemoglobin subunit alpha 2* for peripheral blood [2,55–58,79,80,81]. However, these proteins did not rank among those demonstrating the highest discriminatory power in our analysis. This could be explained by the fact that these high-abundance proteins, while identifiable, are present in trace concentrations in other body fluids, making them less discriminatory from a qualitative perspective. Therefore, reliance on the presence of a single protein for body fluid classification is discouraged. Consequently, in the subsequent paragraphs, a combination of markers will consistently be employed for body fluid classification.

3.3. Body fluid classification in body fluid mixtures

The t-SNE projection below separates the pure body fluids into quite distinct clusters quite effectively (Fig. 4). Each cluster represents a specific body fluid type, as labelled in the figure (saliva, blood, urine, semen, vaginal fluid). This visual separation suggests that the t-SNE dimensionality reduction has overall successfully captured key differences and similarities between the body fluid samples. This indicates body fluid classification of pure body fluids is possible using a proteomics-based methodology. The clustering of a single (female) urine sample with blood samples may be attributed to the presence of traces of blood in the urine, possibly due to menstruation. Moreover, one saliva sample was observed to cluster near the vaginal fluid cluster. Given the limited number of samples in this proof-of-concept study, it remains unclear whether this borderline positioning represents a true misclassification or simply reflects natural biological variability. Future studies with larger sample sizes will be necessary to determine the reproducibility of this observation. Nonetheless, this finding underscores the potential for overlapping proteomic profiles among certain body fluids, which should be considered when interpreting classification outcomes.

After demonstrating the discriminatory potential of proteins from five distinct body fluids using t-SNE, a body fluid prediction was performed for 32 body fluid mixtures. Overall, 75 % of the mixtures were accurately classified when requiring at least one peptide for protein identification. Misclassification, amounting to 9 %, occurred due to random guesses, when only identifying proteins were found for one or no body fluid. Adhering to the two and three peptide rule, an accurate classification of 65.6 % was achieved. The two-peptide rule resulted in a 12.5 % misclassification rate, while the three-peptide rule had a 15.6 % misclassification rate due to random guessing. Fully leveraging the proteomic data, including Gini impurity scores higher than 0, through the implementation of a multilabel classification model, could significantly mitigate these misclassification rates. However, achieving this would necessitate a larger sample size. In properly classified mixtures,

Table 2
The ten most discriminating proteins per pure body fluid. The presentation includes the top ten most discriminating proteins per pure body fluid, ranked by their Gini impurity scores. It further details their relative occurrence in pure samples of the corresponding body fluid, their occurrence in mixtures containing the corresponding body fluid, and their occurrence in mixtures not containing the corresponding body fluid. The proteins are organized based on their Gini impurity scores, followed by their occurrence in pure body fluid samples.

Body fluid	Protein	Gini impurity	Occurrence in		
			pure samples of this BF	mixtures containing this BF	mixtures not containing this BF
Seminal fluid	Cysteine-rich secretory protein 1	0	7/7	11/14	0/18
	Laminin subunit alpha−5	0	7/7	11/14	0/18
	ADP-ribosyl cyclase/cyclic ADP-ribose hydrolase 1	0	7/7	8/14	0/18
	72 kDa type IV collagenase	0	7/7	11/14	0/18
	Limbic system-associated membrane protein	0	7/7	10/14	0/18
	Matrilin−2	0	7/7	11/14	0/18
	A disintegrin and metalloproteinase with thrombospondin motifs 1	0	7/7	10/14	0/18
	Carboxypeptidase Z	0	7/7	10/14	0/18
	Phospholipid hydroperoxide glutathione peroxidase GPX4	0	7/7	10/14	0/18
	Ras-related protein Rab−3B	0	7/7	9/14	0/18
Saliva	Carbonic anhydrase 6	0	7/7	14/14	0/18
	Protein LEG1 homolog	0	7/7	14/14	0/18
	Deleted in malignant brain tumours 1 protein	0	7/7	14/14	0/18
	Lactoperoxidase	0	7/7	14/14	0/18
	BPI fold-containing family A member 2	0	7/7	13/14	0/18
	Keratin, type I cuticular Ha6	0	7/7	11/14	0/18
	Chitinase−3-like protein 2	0	7/7	9/14	0/18
	Liver carboxylesterase 1	0	7/7	9/14	0/18
	Mucin−7	0	6/7	13/14	0/18
	Cystatin-D	0	6/7	12/14	0/18
Peripheral blood	Eukaryotic translation initiation factor 4B	0	5/9	3/14	1/18
	SEC14-like protein 4	0	4/9	0/14	0/18
	Pregnancy zone protein	0	3/9	4/14	0/18
	Hepatocyte growth factor activator	0	3/9	1/14	0/18
	Immunoglobulin lambda-like polypeptide 5	0	2/9	1/14	0/18
	Apolipoprotein C-I	0	2/9	2/14	0/18
	Acylphosphatase−1	0	2/9	0/14	0/18
	Adiponectin	0	2/9	0/14	0/18
	UV excision repair protein RAD23 homolog A	0	2/9	0/14	0/18
	Tubulin-specific chaperone A	0	2/9	1/14	0/18
Vaginal fluid	Cysteine-rich C-terminal protein 1	0	3/3	3/8	0/24
	15-hydroxyprostaglandin dehydrogenase [NAD(+)]	0	3/3	5/8	0/24
	Short-chain dehydrogenase/reductase family 9 C member 7	0	3/3	5/8	1/24
	Serine/threonine-protein kinase RIO3	0	3/3	2/8	0/24
	Serine/threonine-protein phosphatase 4 catalytic subunit	0	3/3	2/8	0/24
	Protein FAM83A	0	3/3	2/8	1/24
	Glycogen [starch] synthase, liver	0	3/3	5/8	0/24
	Syntaxin-binding protein 5	0	3/3	0/8	0/24
	Translocon-associated protein subunit alpha	0	3/3	5/8	1/24
	Heterogeneous nuclear ribonucleoprotein D-like	0	3/3	0/8	0/24
Urine	Osteopontin	0	8/11	8/14	0/18
	Cubilin	0	8/11	8/14	0/18
	Low-density lipoprotein receptor-related protein 2	0	8/11	6/14	0/18
	Roundabout homolog 4	0	8/11	2/14	0/18
	Endosialin	0	8/11	3/14	0/18
	Vasorin	0	7/11	5/14	0/18
	Mannan-binding lectin serine protease 2	0	7/11	7/14	0/18
	Matrix remodelling-associated protein 8	0	7/11	9/14	0/18
	Urokinase-type plasminogen activator	0	7/11	7/14	0/18
	Angiopoietin-related protein 2	0	7/11	2/14	0/18

there were no instances of random guesses.

When mandating at least one peptide for protein identification, an average sample precision of 87.5 % was attained (Fig. 5). Interestingly, implementing the two and three peptide rule led to a reduced average sample precision of 82.8 %, suggesting a preference for utilizing more identifying proteins over fewer, yet more certain, identifying proteins. Comparable trends were observed for the recall score, with saliva exhibiting an exceptionally high recall rate of 100 %. When balancing recall and precision by means of the F1-score, consistent patterns emerged: blood, semen, and saliva displayed the highest F1-score when considering at least one peptide sufficient for identification, while urine and vaginal fluid favoured the requirement of at least two different peptides.

Upon further investigation of the misclassified mixtures, a clear pattern was identified: blood was the body fluid most often not detected, occurring in 71.4 % of the cases. This was followed by semen, which was undetected in 21.4 % of the cases, and urine, which was undetected in 7.1 % of the cases. One of the main reasons for blood not being detected, could be because many of the proteins found in blood are also present in other body fluids. This significant overlap of proteins can make it

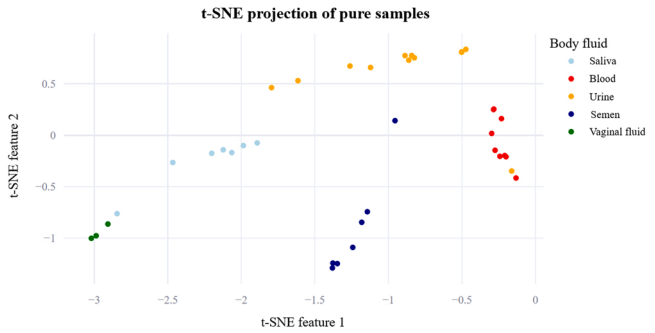


Fig. 4. A t-SNE projection of the pure samples. This scatterplot depicts the results of applying the t-distributed Stochastic Neighbour Embedding (t-SNE) algorithm to the generated proteomic dataset containing pure samples of peripheral blood (red), semen (dark blue), saliva (light blue), urine (yellow), and vaginal fluid (green).

difficult to distinguish blood, leading to it being either undetected or misclassified [82].

It should be noted that the protein concentration of semen measured in this study (16.8 ± 6.47 mg/mL) was considerably lower than typical literature values for seminal plasma (35–55 mg/mL) [60]. While several potential causes were addressed in the Materials & Methods section, this underestimation may have influenced mixture classification. Specifically, an underestimation of semen protein concentration would lead to a higher volume of semen being added to the mixtures to achieve equal total protein content. As a result, semen may be overrepresented in the mixtures relative to its true forensic occurrence, potentially biasing the body fluid classification toward detection of semen. Although the approach was designed to equalize protein contributions from each fluid, this nuance should be considered when interpreting the mixture classification outcomes, especially in cases involving semen.

3.4. Source attribution of body fluids

Source attribution of body fluids was exclusively undertaken for

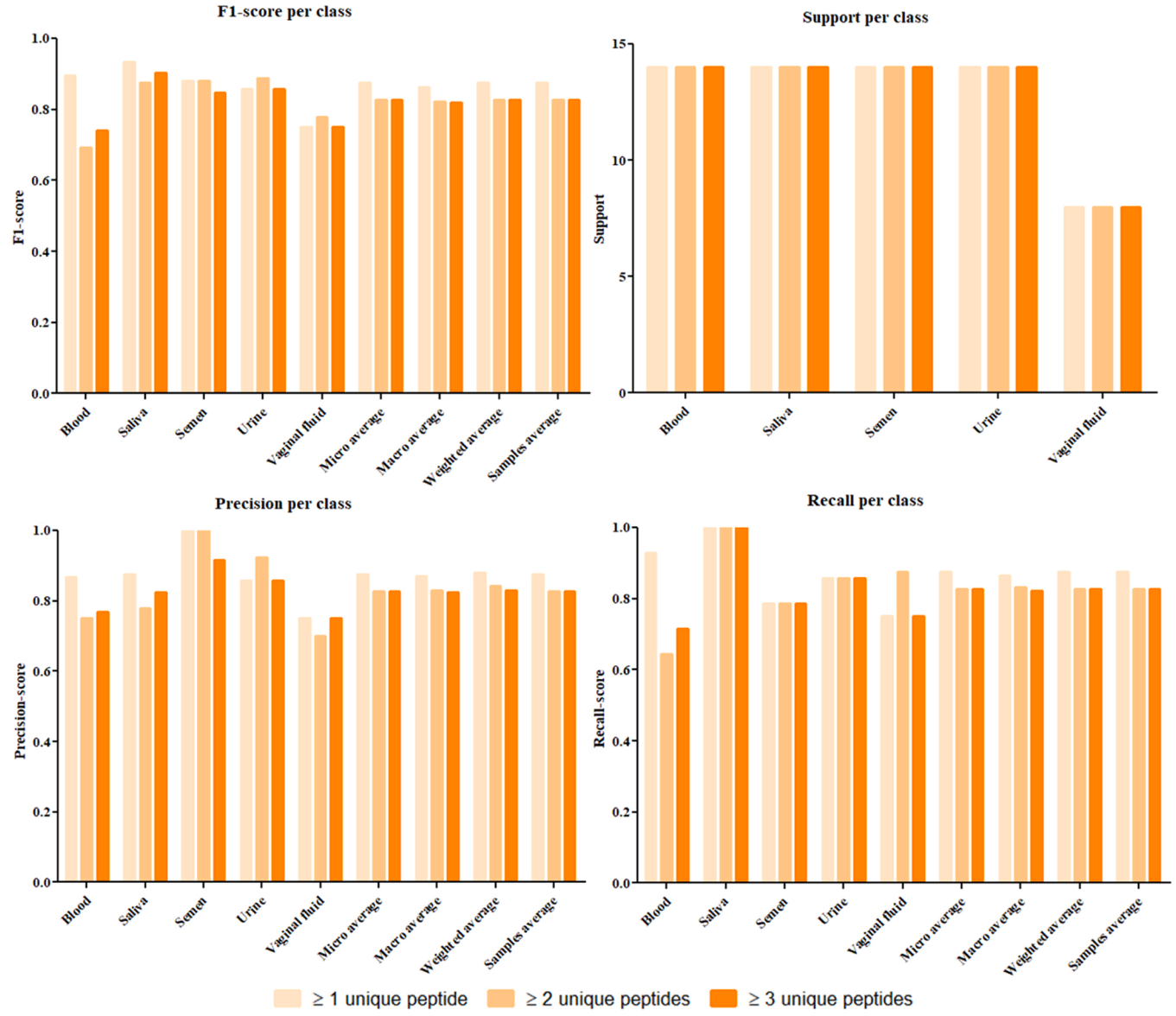


Fig. 5. The F1-score, support, precision, and recall per class for body fluid classification in a mixture of two body fluids. The F1-score (upper left), support (upper right), precision (lower left), and recall (lower right) for the identification of different body fluids in a mixture of two body fluids using the 1, 2, and 3 peptide rule for protein identification (light to dark orange, respectively).

mixtures containing saliva (n = 11). This selection was based on the accurate classification of saliva for all 11 mixtures containing saliva. In the 7 pure saliva samples and 11 mixtures in total 288 SAAVs were identified, excluding those attributable to chemical modifications or indistinguishable from other SAAVs via mass spectrometry (Supplementary information 7). Comparing the SAAVs found in saliva-specific proteins within the mixtures to those in the 7 pure saliva samples revealed that 71.4 % of the saliva in the mixtures could be correctly attributed to the corresponding pure sample from the same individual. No discernible trend was observed in the misclassified samples. However, it is important to note that comparing lists of (un)identified SAAVs may not yield comprehensive insights, as not all peptides were identified in every sample. This variability could stem from factors such as sample preparation (e.g., incomplete digestion or peptide loss during extraction), instrumental constraints (e.g., high limit of detection for low-abundance peptides or variations in ionization efficiency), and data analysis considerations and is coherent with other proteomic studies [83,84]. Furthermore, source attribution within a closed dataset may not adequately represent source attribution within the wider population. However, in many cases the activity and not the actors will be disputed, for which this will not be an issue.

What would be more useful is to calculate a likelihood ratio (LR) expressing the strength of evidence for the origin of the saliva, using the random match probability (RMP) of the complete set of SAAVs. However, several limitations currently prevent this. First, the statistical relationship between SAAV-based RMPs and those derived from single-nucleotide polymorphisms (SNPs) – for which population allele frequencies are well characterized – is not yet well established. Additionally, it is unknown whether the donors were homozygous or heterozygous at each locus, which influences LR calculations. In scenarios without monoallelic expression, heterozygous individuals may produce both protein variants, resulting in bi-allelic expression [85]. In Table 3, we present an overview of the most discriminating SAAVs – those uniquely present in one individual's pure saliva sample and consistently detected in mixtures from the same donor – along with their associated nsSNPs and allele frequencies, which serve as a rough proxy for SAAV frequency. Assuming independence between variants, multiplying these frequencies yields a relatively low RMP for this specific case, underscoring the potential of SAAVs for forensic source attribution. This shows the high potential of SAAVs for source attribution of body fluids in mixtures.

However, many of the observed SAAVs, for example those listed in Table 3, were rare, which limits their current forensic utility in probabilistic models. To enable LR-based approaches in future casework, it will be essential to focus on genetically variant peptides (GVPs) with higher minor allele frequencies. Common variants are statistically more robust, more likely to be reproducible across different datasets, and crucially, allow for the use of population-based frequency databases. These are fundamental requirements for calculating reliable likelihood ratios and moving toward operational forensic implementation. Future research should aim to refine peptide filtering strategies, expand population proteomic references, and determine the reproducibility and

Table 3
Examples of single amino acid variants identified in proteins with a Gini impurity score of zero. This table collates the protein, single amino acid variant and corresponding non synonymous single nucleotide polymorphism with variant ID. The nsSNP frequency for a European population is also listed. Green indicates the presence of the variant and red the absence of the variant in the pure and mixed samples.

Protein accession code	SAAV	nsSNP	Variant ID	Frequency nsSNP (Allele frequency database, sample size European subgroup)	Saliva 1	Saliva 2	Saliva 3	Saliva 5	Saliva 8	Saliva 10	Saliva 13	Saliva 1 – Semen 4	Saliva 2 – Semen 6	Saliva 2 – Urine 13	Saliva 3 – Semen 8	Saliva 3 – Vaginal fluid 12	Saliva 5 – Blood 10	Saliva 8 – Semen 1	Saliva 8 – Blood 5	Saliva 8 – Blood 9	Saliva 8 – Vaginal fluid 11	Saliva 10 – Urine 12
Q8NES3	N243S	A>G	rs939411540	0.00000 (ALFA, n=19224)	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
A0A075B7D0	T112A	T>C	rs746763011	0.0081 (ALFA, n=4462)	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Q6P5S2	P54T	C>A	rs760569836	0.00015 (ALFA, n=32646)	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
P02810	R119C	C>T	rs200488155	0.00157 (ALFA, n=12080)	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
P23083	S96T	C>A,G,T	rs188093056	0.00050 (ALFA, n=18118)	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
A0A075B6S5	G79E	C>G,T	rs1573025401	-	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
P01602	S82A	A>C,G,T	rs751093913	0.0000 (ALFA, n=6962)	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
A0A087WSY6	N52H	A>C,T	rs1387785387	0.00007 (ALFA, n=13446)	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
P19961	W218*	G>A	rs1570647535	-	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
P19961	G428R	G>A	rs745947383	0.000000 (ExAC, n=68804)	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Q8WWA0	G16R	C>T	rs760854751	0.0001 (ALFA, n=9690)	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●

inheritance patterns of commonly observed SAAVs.

4. Limitations

4.1. Body fluid classification

An important limitation of the applied method for body fluid classification is that it likely provides a more optimistic view than reality, as the variability in the data may be lower than expected in the broader population. Furthermore, because the proteins identified in the mixed samples come from the same individuals as the proteins in the pure samples, the diversity of the dataset is constrained, potentially leading to an overestimation of the recall. Additionally, the mixture ratio was configured such that roughly half of the proteins originated from each of the two body fluids in the mixture, a composition that does not comprehensively mirror all real-world forensic scenarios. Moreover, the method assumes that the mixture comprises only two body fluids and is limited to a closed set of five body fluids for selection, potentially leading to inflated recall rates for all classes. This means that the results may not be fully representative of more complex scenarios where there is greater variability in the samples and a larger number of possible body fluids present.

As with any analytical method, both DIA (Data-Independent Acquisition) and DDA (Data-Dependent Acquisition) come with their respective advantages and disadvantages [86]. One of the key strengths of DIA compared to DDA is its greater dynamic range, which allows for the detection of both high- and low-abundance peptides within a single run. This leads to improved consistency and reproducibility, which is especially beneficial when analysing complex biological samples, such as mixtures of body fluids. However, within a forensic context, DIA presents quite a few limitations. While its consistent acquisition, targets all precursor ions within a specified mass range and provides broad coverage, this less selective approach can introduce ambiguity, particularly in complex samples. This makes data analysis and interpretation much more challenging. A significant limitation of DIA is its use of a wider precursor isolation window compared to DDA. This increases the likelihood of contaminant peptides being coeluted and cofragmented with target peptides. We encountered this issue in our own dataset, where contaminant signals interfered with confident identification. It's important to note that while DDA provides more targeted and confident identifications, the broader and less selective nature of DIA can increase ambiguity and the number of false positives. This is an important concern in forensic contexts where accuracy and reliability are crucial.

It is important to consider these limitations when interpreting the findings and when considering the future use of the method in broader forensic contexts.

4.2. Source attribution of body fluids

Since in its early stages, this application of forensic proteomics will bring forth several critical considerations compared to the 'gold standards' of DNA and mRNA analysis. The data that we present are based on samples limited in number and composition, so there is no clear view yet on some of the analytical characteristics (such as robustness, sensitivity and reproducibility) that are needed for acceptance and ultimately validation of the method. Additionally, our understanding of the exact changes of SAAVs over time *in* and *ex vivo* and forensic value remains limited, a pivotal factor for its forensic applicability. Moreover, the specificity of the method has yet to be thoroughly evaluated, to alleviate concerns regarding the potential for false-positive results. Furthermore, the absence of an established ethical and legal framework further complicates the adoption of forensic proteomics in practice as a complementary option for DNA and mRNA profiling. However, due to the robustness of the protein backbone we believe there could be scenarios in the future where protein based information especially regarding inference of SAAVs could complement information obtained from DNA and mRNA analysis.

5. Future recommendations

5.1. Body fluid classification

To address the limitations and build upon our findings, several avenues for future research are suggested. First, conducting larger-scale studies to validate the findings and assess the reproducibility of proteomic-based body fluid classification methods across diverse sample populations. Second, optimising sample preparation and implementing rigorous quality control measures to mitigate contamination and ensure the reliability of proteomic data. This approach ensures the maximization of information obtained from the sample, allowing for the development of a targeted strategy, such as Single Reaction Monitoring (SRM), to enhance the sensitivity and detection of SAAVs. Third, developing advanced classification algorithms, such as machine learning models, to improve the accuracy of body fluid classification, especially in the context of mixed samples. These models could make use of the full proteome including quantitative data. Additionally, although McKiernan *et al.* (2021) and Zaarour *et al.* (2025) conducted valuable work on body fluid classification in mixed (forensic casework) samples, and Davidovics *et al.* (2022) conducted research on the sensitivity of vaginal fluid in semen and *vice verse*, establishing the limits of detection of body fluids within other body fluids is imperative to examine whether the method is applicable in real-world forensic scenarios [56,57,87]. Fourth, establishing ethical and legal frameworks for the integration of proteomic analysis into forensic investigations, addressing privacy concerns and ensuring compliance with existing regulations: although proteomic data share many privacy concerns with DNA profiling, additional ethical considerations may arise due to the potential for revealing additional sensitive health-related information. An established ethical framework for proteomics should therefore address both privacy and data security, ensuring that proteomic profiles are handled with the same rigor as DNA data while acknowledging the unique aspects of protein expression. Furthermore, future studies could incorporate other forensically relevant body fluids like sweat, tears, cerebrospinal fluid, and stool. Expanding the inclusion of various cell types helps mitigate the occurrence of "out-of-distribution" instances, a term in artificial intelligence denoting data points diverging notably from the training data distribution, thus reducing false positive rates. Finally, working towards standardization of protocols and methodologies for proteomic analysis in forensic laboratories to facilitate comparison and reproducibility of results.

5.2. Source attribution of body fluids

Concerning source attribution of body fluids, future research into its forensic applicability is imperative. Initially, the stability of single amino acid variants over time requires examination, as it could impact the validity of proteomic source attribution. Explicitly evaluating the impact of aging and degradation on protein and SAAV stability and the reliability for body fluid classification and source attribution by analyzing samples stored under conditions that simulate real forensic scenarios is recommended. For example, examining fresh and long-term aged samples could provide valuable insights into the time frame over which proteomic markers remain reliable for both body fluid classification and source attribution. Second, investigation into whether the formation of single amino acid variants is exclusively attributed to single nucleotide polymorphisms or if they could arise post-transcriptionally is warranted in order to understand whether it would be able to attribute a proteomic profile using DNA as reference. As this is a proof-of-concept study, validation of the identified SAAVs through comparison with DNA-based genotypes or other forms of proteomic validation (e.g., chromatographic retention time) was not performed. Future studies could incorporate such validation strategies, including genomic

confirmation, as described in the research by Parker et al. (2016), Mason et al. (2018) and Parker et al. (2021) [44,48,54]. Thirdly, it is essential to ascertain the impact of the formation of single nucleotide polymorphisms in the DNA over time on the reliability of source attribution utilizing single amino acid variants. Fourth, determining whether the random match probability of single amino acid variants is equal to that of single nucleotide polymorphisms is essential when dealing with an open set of suspects. Fifth, it is crucial to ascertain the forensic discriminatory value of single amino acid variants and the impact of homozygosity and heterozygosity on the source attribution of body fluids using mass spectrometry based proteomics. Lastly, future studies should prioritize SAAVs with higher minor allele frequencies to enable robust and probabilistic interpretation in forensic applications.

By addressing these recommendations, future research can advance the application of proteomic analysis in forensic science, enhancing the accuracy and reliability of body fluid classification and source attribution in criminal investigations.

6. Conclusion

This study aimed to explore the potential of proteomic analysis in forensic investigations for identifying, classifying, and source attributing body fluids. Employing liquid chromatography tandem mass spectrometry in data-independent acquisition mode, we characterized a wide array of proteins present in various body fluids, including peripheral blood, semen, saliva, urine, and vaginal fluid. Through t-distributed stochastic neighbour embedding analysis, we demonstrated the discriminatory power of proteins from these body fluids, allowing for effective classification of body fluids. Our findings indicate that utilizing a proteomic-based approach can yield high precision and recall rates in classifying body fluids from mixed samples. However, we observed a decrease in precision when applying stricter criteria for protein identification, emphasizing the trade-off between precision and inclusivity in forensic proteomics. Furthermore, the misclassification rates were notably high, especially for blood. This issue could potentially be addressed by employing a multilabel classification model after increasing the sample size, making full use of the obtained proteomic data.

In this study, the potential of utilizing single amino acid variants for body fluid source attribution is demonstrated. A total of 294 single amino acid variants were identified across the samples, all exhibiting forensic discriminatory power. However, the exact forensic discriminatory power remains unknown and validation is needed.

Despite the promising preliminary results, several challenges and limitations were identified. Contamination by polyethylene glycol in mass spectra highlights the need for stringent quality control measures in sample preparation. Additionally, the small sample size limited the application of multilabel classification models, likely affecting the accuracy of body fluid classification. Notably, due to the limited sample size and thus not being able to apply a multilabel classification model, we were unable to fully utilize data – and thus to fully assess the classification power – indicating for example proteins present exclusively in two body fluids and absent in others, which still possess significant discriminatory power. Additionally, the representativeness of both classification and source attribution to the broader population could be subject to questioning, rendering the representativeness of this proof-of-concept study unknown.

All in all, this proof-of-concept study has shown the potential of mass spectrometry-based proteomics for body fluid classification and body fluid source attribution leveraging single amino acid variants. This approach proves valuable in scenarios where RNA and DNA analyses prove impractical or impossible.

Human research ethics

The conducted study is a participant-based experiment, indicating that ethical considerations are of primary importance. These ethical considerations have been taken into account in all aspects of this research. First, physical harm, psychological, social, and legal harm were also considered prior to the design of this study. Moreover, participation to this study was completely voluntarily: all participants were able to withdraw from, or leave, the study at any point without the feeling of obligation and without giving a reason. Participation to this study did only occur if the participant was fully informed about the study that was going to be conducted and if they had given their consent beforehand. If consent was given, personal data were collected and handled securely and with the utmost respect to the participants complying with the regulations as specified in the "EU General Data Protection Regulation 2016/679" (2016). Besides, the data of the people participating in this study were stored anonymously throughout the entire study, as defined in article 4 (1) of the "EU General Data Protection Regulation 2016/679" (2016). The right of the participants to request deletion of their personal data were enabled through data pseudonymization. Moreover, all personal information of the participant that was not relevant for the results of this study was not available to scientific staff not involved in the experiments. Furthermore, it is important to emphasize a conflict of interest was and should be avoided at all times. This occurs when the study is performed for academic, financial, or personal purposes rather than to benefit the participant of interest. This also means that any data obtained with this study were and will not be shared with others unless the participants have given explicit consent for this. At last, during this study steps were taken to avoid (self-)plagiarism and misconduct.

CRediT authorship contribution statement

van Lierop Stijn N.C: Visualization, Software, Data curation. **Shirin Alex:** Writing – review & editing, Writing – original draft, Methodology, Investigation. **Shehata Thomas:** Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Marcel de Puit:** Writing – review & editing, Supervision. **Jeroen Demmers:** Writing – review & editing, Software, Resources, Methodology, Conceptualization. **Nadine Prust:** Writing – review & editing, Software, Resources, Methodology, Investigation. **Wetering-Tieleman Jantine van de:** Writing – review & editing, Resources, Methodology, Investigation. **Blom Maarten:** Writing – review & editing, Supervision.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT, an artificial intelligence developed by OpenAI, in order to assist in their redaction. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Firstly, we would like to thank all participants that participated voluntarily in this study for their time and effort. Also, we would like to thank Isabelle Kohler for fulfilling her role as supervisor, her help and advice. At last, a big thank you to Wouter Doff from the Proteomics Center of the Erasmus MC for his expert help and advice regarding the bioinformatics analysis.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fsigen.2025.103343.

References

[1] M. Bauer, RNA in forensic science, Forensic Sci. Int. Genet. 1 (2007) 69–74, https://doi.org/10.1016/J.FSIGEN.2006.11.002.
[2] H. Yang, B. Zhou, H. Deng, M. Prinz, D. Siegel, Body fluid identification by mass spectrometry, (n.d.). https://doi.org/10.1007/s00414-013-0848-1.
[3] C. Lynch, R. Fleming, RNA -based approaches for body fluid identification in forensic science, WIREs Forensic Sci. 3 (2021), https://doi.org/10.1002/WFS2.1407.
[4] J.H. An, K.J. Shin, W.I. Yang, H.Y. Lee, Body fluid identification in forensics, BMB Rep. 45 (2012) 545–553, https://doi.org/10.5483/BMBREP.2012.45.10.206.
[5] T. Sijen, S. Harbison, On the identification of body fluids and tissues: a crucial link in the investigation and solution of crime, Genes (Basel) 12 (2021), https://doi.org/10.3390/GENES12111728.
[6] P. Danaher, R.L. White, E.K. Hanson, J. Ballantyne, Facile semi-automated forensic body fluid identification by multiplex solution hybridization of NanoString® barcode probes to specific mRNA targets, Forensic Sci. Int. Genet. 14 (2015) 18–30, https://doi.org/10.1016/J.FSIGEN.2014.09.005.
[7] K. Virkler, I.K. Lednev, Analysis of body fluids for forensic purposes: from laboratory testing to non-destructive rapid confirmatory identification at a crime scene, Forensic Sci. Int. 188 (2009) 1–17, https://doi.org/10.1016/J.FORSCIINT.2009.02.013.
[8] R.I. Fleming, S. Harbison, The development of a mRNA multiplex RT-PCR assay for the definitive identification of body fluids, Forensic Sci. Int. Genet. 4 (2010) 244–256, https://doi.org/10.1016/J.FSIGEN.2009.10.006.
[9] E. Hanson, S. Ingold, C. Haas, J. Ballantyne, Targeted multiplexed next generation RNA sequencing assay for tissue source determination of forensic samples, Forensic Sci. Int. Genet. Suppl. Ser. 5 (2015) e441–e443, https://doi.org/10.1016/j.fsigss.2015.09.175.
[10] J. Juusola, J. Ballantyne, Messenger RNA profiling: a prototype method to supplant conventional methods for body fluid identification, Forensic Sci. Int. 135 (2003) 85–96, https://doi.org/10.1016/S0379-0738(03)00197-X.
[11] A. Lindenbergh, M. De Pagter, G. Ramdayal, M. Visser, D. Zubakov, M. Kayser, T. Sijen, A multiplex (m)RNA-profiling system for the forensic identification of body fluids and contact traces, Forensic Sci. Int. Genet. 6 (2012) 565–577, https://doi.org/10.1016/J.FSIGEN.2012.01.009.
[12] C.O. Brown, C.G. Westring, P.B. Danielson, K.M. Legg, Saliva identification in forensic samples by automated microextraction and intact mass analysis of statherin, J. Forensic Sci. 69 (2024) 640–650, https://doi.org/10.1111/1556-4029.15445.
[13] S. Boyd, M.F. Bertino, S.J. Seashols, Raman spectroscopy of blood samples for forensic applications, Forensic Sci. Int. 208 (2011) 124–128, https://doi.org/10.1016/J.FORSCIINT.2010.11.012.
[14] K. Virkler, I.K. Lednev, Raman spectroscopic signature of semen and its potential application to forensic body fluid identification, Forensic Sci. Int. 193 (2009) 56–62, https://doi.org/10.1016/J.FORSCIINT.2009.09.005.
[15] T. Casey, E. Mistek, L. Halámková, I.K. Lednev, Raman spectroscopy for forensic semen identification: method validation vs. Environmental interferences, Vib. Spectrosc. 109 (2020), https://doi.org/10.1016/j.vibspec.2020.103065.
[16] A. Takamura, K. Watanabe, T. Akutsu, T. Ozawa, Soft and robust identification of body fluid using Fourier transform infrared spectroscopy and chemometric strategies for forensic analysis, Sci. Rep. 2018 81. 8 (2018) 1–10, https://doi.org/10.1038/s41598-018-26873-9.
[17] C.-M. Orphanou, L. Walton-Williams, H. Mountain, J. Cassella, The detection and discrimination of human body fluids using ATR FT-IR spectroscopy, Forensic Sci. Int. 252 (2015) e10–e16, https://doi.org/10.1016/J.FORSCIINT.2015.04.020.
[18] S. Sharma, R. Chophi, R. Singh, Forensic discrimination of menstrual blood and peripheral blood using attenuated total reflectance (ATR)-Fourier transform infrared (FT-IR) spectroscopy and chemometrics, Int. J. Leg. Med. 134 (2020) 63–77, https://doi.org/10.1007/S00414-019-02134-W.
[19] S. Zha, X. Wei, R. Fang, Q. Wang, H. Lin, K. Zhang, H. Zhang, R. Liu, Z. Li, P. Huang, Z. Wang, Estimation of the age of human semen stains by attenuated total reflection Fourier transform infrared spectroscopy: a preliminary study, Forensic Sci. Res. 5 (2019) 119–125, https://doi.org/10.1080/20961790.2019.1642567.
[20] A. Takamura, L. Halamkova, T. Ozawa, I.K. Lednev, Phenotype profiling for forensic purposes: determining donor sex based on Fourier transform infrared spectroscopy of urine traces, Anal. Chem. 91 (2019) 6288–6295, https://doi.org/10.1021/ACS.ANALCHEM.9B01058.
[21] E. Mistek-Morabito, I.K. Lednev, Discrimination of menstrual and peripheral blood traces using attenuated total reflection Fourier transform-infrared (ATR FT-IR) spectroscopy and chemometrics for forensic purposes, Anal. Bioanal. Chem. 413 (2021) 2513–2522, https://doi.org/10.1007/S00216-021-03206-W.
[22] T. Sijen, Molecular approaches for forensic cell type identification: on mRNA, miRNA, DNA methylation and microbial markers, Forensic Sci. Int. Genet. 18 (2015) 21–32, https://doi.org/10.1016/J.FSIGEN.2014.11.015.
[23] S. Harbison, R. Fleming, Forensic body fluid identification: state of the art, Res. Rep. Forensic Med. Sci. 6 (2016) 11–23, https://doi.org/10.2147/RRFMS.S57994.
[24] M. Bauer, D. Patzelt, Evaluation of mrna markers for the identification of menstrual blood, J. Forensic Sci. 47 (2002) 1278–1282, https://doi.org/10.1520/JFS15560J.
[25] H.Y. Lee, J.H. An, S.E. Jung, Y.N. Oh, E.Y. Lee, A. Choi, W.I. Yang, K.J. Shin, Genome-wide methylation profiling and a multiplex construction for the identification of body fluids using epigenetic markers, Forensic Sci. Int. Genet. 17 (2015) 17–24, https://doi.org/10.1016/J.FSIGEN.2015.03.002.
[26] J. Antunes, Q. Gauthier, V. Aguiar-Pulido, G. Duncan, B. McCord, A data-driven, high-throughput methodology to determine tissue-specific differentially methylated regions able to discriminate body fluids, Electrophoresis 42 (2021) 1168–1176, https://doi.org/10.1002/ELPS.202000217.
[27] J.H. An, A. Choi, K.J. Shin, W.I. Yang, H.Y. Lee, DNA methylation-specific multiplex assays for body fluid identification, Int. J. Leg. Med. 127 (2013) 35–43, https://doi.org/10.1007/S00414-012-0719-1.
[28] C.C.G. Benschop, F.C.A. Quaak, M.E. Boon, T. Sijen, I. Kuiper, Vaginal microbial flora analysis by next generation sequencing and microarrays; can microbes indicate vaginal origin in a forensic context? Int. J. Leg. Med. 126 (2012) 303–310, https://doi.org/10.1007/S00414-011-0660-8.
[29] S.L. Leake, M. Pagni, L. Falquet, F. Taroni, G. Greub, The salivary microbiome for differentiating individuals: proof of principle, Microbes Infect. 18 (2016) 399–405, https://doi.org/10.1016/J.MICINF.2016.03.011.
[30] J.Y. Jung, H.K. Yoon, S. An, J.W. Lee, E.R. Ahn, Y.J. Kim, H.C. Park, K. Lee, J. H. Hwang, S.K. Lim, Rapid oral bacteria detection based on real-time PCR for the forensic identification of saliva, Sci. Rep. 8 (2018), https://doi.org/10.1038/S41598-018-29264-2.
[31] F.C.A. Quaak, M.L.M. de Graaf, R. Weterings, I. Kuiper, Microbial population analysis improves the evidential value of faecal traces in forensic investigations, Int. J. Leg. Med. 131 (2017) 45–51, https://doi.org/10.1007/S00414-016-1390-8.
[32] T. Akutsu, H. Motani, K. Watanabe, H. Iwase, K. Sakurada, Detection of bacterial 16S ribosomal RNA genes for forensic identification of vaginal fluid, Leg. Med. (Tokyo). 14 (2012) 160–162, https://doi.org/10.1016/J.LEGALMED.2012.01.005.
[33] C. Díez López, D. Montiel González, C. Haas, A. Vidaki, M. Kayser, Microbiome-based body site of origin classification of forensically relevant blood traces, Forensic Sci. Int. Genet. 47 (2020), https://doi.org/10.1016/J.FSIGEN.2020.102280.
[34] C. Díez López, A. Vidaki, A. Ralf, D. Montiel González, D. Radjabzadeh, R. Kraaij, A.G. Uitterlinden, C. Haas, O. Lao, M. Kayser, Novel taxonomy-independent deep learning microbiome approach allows for accurate classification of different forensically relevant human epithelial materials, Forensic Sci. Int. Genet. 41 (2019) 72–82, https://doi.org/10.1016/J.FSIGEN.2019.03.015.
[35] H.E. McKiernan, C.O. Brown, L.C. Arantes, P.B. Danielson, K.M. Legg, NextGen serology: leveraging mass spectrometry for Protein-Based human body fluid identification, ACS Symp. Ser. 1339 (2019) 47–80, https://doi.org/10.1021/BK-2019-1339.CH004.
[36] Z.C. Goecker, K.M. Legg, M.R. Salemi, A.W. Herren, B.S. Phinney, H.E. McKiernan, G.J. Parker, Alternative LC-MS/MS platforms and data acquisition strategies for proteomic genotyping of human hair shafts, J. Proteome Res. 20 (2021) 4655–4666, https://doi.org/10.1021/ACS.JPROTEOME.1C00209.
[37] K. Yoshida, K. Sekiguchi, N. Mizuno, K. Kasai, I. Sakai, H. Sato, S. Seta, The modified method of two-step differential extraction of sperm and vaginal epithelial cell DNA from vaginal fluid mixed with semen, Forensic Sci. Int. 72 (1995) 25–33, https://doi.org/10.1016/0379-0738(94)01668-U.
[38] P. Gill, Misleading DNA evidence: reasons for miscarriages of justice, Int. Comment. Evid. 10 (2012) 55–71, https://doi.org/10.1515/ICE-2014-0010.
[39] G.E. Meakin, B. Kokshoorn, | Roland, A.H. Van Oorschot, B. Szkuta, Evaluating forensic DNA evidence: connecting the dots, Wiley Interdiscip. Rev. Forensic Sci. 3 (2021) e1404, https://doi.org/10.1002/WFS2.1404.
[40] M. Neis, T. Groß, H. Schneider, P.M. Schneider, C. Courts, Comprehensive body fluid identification and contributor assignment by combining targeted sequencing of mRNA and coding region SNPs, Forensic Sci. Int. Genet. 73 (2024) 103125, https://doi.org/10.1016/J.FSIGEN.2024.103125.
[41] R.Elise Keane, R.Jayne Tidy, G.J. Parker, J.P.A. Gummer, C. Priddis, C.P.Joel A. Gummer, Mass spectrometry based proteomics: changing the impact of protein analysis in forensic science, Wiley Interdiscip. Rev. Forensic Sci. 6 (2024) e1516, https://doi.org/10.1002/WFS2.1516.
[42] M.R. Wilkins, J.C. Sanchez, A.A. Gooley, R.D. Appel, I. Humphery-Smith, D. F. Hochstrasser, K.L. Williams, Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it, Biotechnol. Genet. Eng. Rev. 13 (1996) 19–50, https://doi.org/10.1080/02648725.1996.10647923.
[43] J.A. Milan, P.W. Wu, M.R. Salemi, B.P. Durbin-Johnson, D.M. Rocke, B.S. Phinney, R.H. Rice, G.J. Parker, Comparison of protein expression levels and proteomically-inferred genotypes using human hair from different body sites, Forensic Sci. Int. Genet. 41 (2019) 19–23, https://doi.org/10.1016/J.FSIGEN.2019.03.009.
[44] G.J. Parker, T. Leppert, D.S. Anex, J.K. Hilmer, N. Matsunami, L. Baird, J. Stevens, K. Parsawar, B.P. Durbin-Johnson, D.M. Rocke, C. Nelson, D.J. Fairbanks, A. S. Wilson, R.H. Rice, S.R. Woodward, B. Bothner, B.R. Hart, M. Leppert, Demonstration of Protein-Based human identification using the hair shaft proteome, PLoS One 11 (2016), https://doi.org/10.1371/JOURNAL.PONE.0160653.
[45] L.A. Catlin, R.M. Chou, Z.C. Goecker, L.A. Mullins, D. Silva, R.R. Spurbeck, G. J. Parker, C.M. Bartling, Demonstration of a mitochondrial DNA-compatible workflow for genetically variant peptide identification from human hair samples, Forensic Sci. Int. Genet. 43 (2019) 102148, https://doi.org/10.1016/J.FSIGEN.2019.102148.
[46] R.N. Franklin, N. Karim, Z.C. Goecker, B.P. Durbin-Johnson, R.H. Rice, G.J. Parker, Proteomic genotyping: using mass spectrometry to infer SNP genotypes in

pigmented and non-pigmented hair, Forensic Sci. Int. 310 (2020) 110200, https://doi.org/10.1016/J.FORSCIINT.2020.110200.
[47] P.W. Wu, K.E. Mason, B.P. Durbin-Johnson, M. Salemi, B.S. Phinney, D.M. Rocke, G.J. Parker, R.H. Rice, Proteomic analysis of hair shafts from monozygotic twins: expression profiles and genetically variant peptides, Proteomics 17 (2017), https://doi.org/10.1002/PMIC.201600462.
[48] K.E. Mason, D. Anex, T. Grey, B. Hart, G. Parker, Protein-based forensic identification using genetically variant peptides in human bone, Forensic Sci. Int. 288 (2018) 89–96, https://doi.org/10.1016/J.FORSCIINT.2018.04.016.
[49] T. Borja, N. Karim, Z. Goecker, M. Salemi, B. Phinney, M. Naeem, R. Rice, G. Parker, Proteomic genotyping of fingermark donors with genetically variant peptides, Forensic Sci. Int. Genet. 42 (2019) 21–30, https://doi.org/10.1016/J.FSIGEN.2019.05.005.
[50] K.M. Legg, R. Powell, N. Reisdorph, R. Reisdorph, P.B. Danielson, Verification of protein biomarker specificity for the identification of biological stains by quadrupole time-of-flight mass spectrometry, Electrophoresis 38 (2017) 833–845, https://doi.org/10.1002/ELPS.201600352.
[51] K. Van Steendam, M. De Ceuleneer, M. Dhaenens, D. Van Hoofstat, D. Deforce, Mass spectrometry-based proteomics as a tool to identify biological matrices in forensic science, Int. J. Leg. Med. 127 (2013) 287, https://doi.org/10.1007/S00414-012-0747-X.
[52] M. Zhao, Y. Yang, Z. Guo, C. Shao, H. Sun, Y. Zhang, Y. Sun, Y. Liu, Y. Song, L. Zhang, Q. Li, J. Liu, M. Li, Y. Gao, W. Sun, A comparative proteomics analysis of five body fluids: plasma, urine, cerebrospinal fluid, amniotic fluid, and saliva, Proteom. Clin. Appl. 12 (2018), https://doi.org/10.1002/PRCA.201800008.
[53] S. Kamanna, J. Henry, N.H. Voelcker, A. Linacre, K.P. Kirkbride, Direct identification of forensic body fluids using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry, Int. J. Mass Spectrom. 397–398 (2016) 18–26, https://doi.org/10.1016/J.IJMS.2016.01.002.
[54] G.J. Parker, H.E. McKiernan, K.M. Legg, Z.C. Goecker, Forensic proteomics, Forensic Sci. Int. Genet. 54 (2021) 102529, https://doi.org/10.1016/J.FSIGEN.2021.102529.
[55] C.O. Brown, B.L. Robbins, H.E. McKiernan, P.B. Danielson, K.M. Legg, Direct seminal fluid identification by protease-free high-resolution mass spectrometry, J. Forensic Sci. 66 (2021) 1017–1023, https://doi.org/10.1111/1556-4029.14646.
[56] H.E. McKiernan, P.B. Danielson, C.O. Brown, M. Signaevsky, C.G. Westring, K. M. Legg, Developmental validation of a multiplex proteomic assay for the identification of forensically relevant biological fluids, Forensic Sci. Int. 326 (2021), https://doi.org/10.1016/J.FORSCIINT.2021.110908.
[57] L. Zaarour, M. Padula, R.A.H. van Oorschot, D. McNevin, Mass spectrometry-based proteomics for source-level attribution after DNA extraction, Forensic Sci. Int. Genet. 74 (2025) 103168, https://doi.org/10.1016/J.FSIGEN.2024.103168.
[58] K.M. Legg, R. Powell, N. Reisdorph, R. Reisdorph, P.B. Danielson, Discovery of highly specific protein markers for the identification of biological stains, Electrophoresis 35 (2014) 3069–3078, https://doi.org/10.1002/ELPS.201400125.
[59] F. Oldoni, K.K. Kidd, D. Podini, Microhaplotypes in forensic genetics, Forensic Sci. Int. Genet. 38 (2019) 54–69, https://doi.org/10.1016/J.FSIGEN.2018.09.009.
[60] B. Pilch, M. Mann, Large-scale and high-confidence proteomic analysis of human seminal plasma, Genome Biol. 7 (2006) 1–10, https://doi.org/10.1186/GB-2006-7-5-R40/FIGURES/4.
[61] C.S. Hughes, S. Moggridge, T. Müller, P.H. Sorensen, G.B. Morin, J. Krijgsveld, Single-pot, solid-phase-enhanced sample preparation for proteomics experiments, 2018 141, Nat. Protoc. 14 (2018) 68–85, https://doi.org/10.1038/s41596-018-0082-x.
[62] B. Zhang, M.C. Chambers, D.L. Tabb, Proteomic parsimony through bipartite graph analysis improves accuracy and transparency, J. Proteome Res. 6 (2007) 3549–3557, https://doi.org/10.1021/PR070230D/SUPPL_FILE/PR070230DSI20070622_013053.PDF.
[63] F. Trulsson, V. Akimov, M. Robu, N. van Overbeek, D.A.P. Berrocal, R.G. Shah, J. Cox, G.M. Shah, B. Blagoev, A.C.O. Vertegaal, Deubiquitinating enzymes and the proteasome regulate preferential sets of ubiquitin substrates, 2022 131, Nat. Commun. 13 (2022) 1–17, https://doi.org/10.1038/s41467-022-30376-7.
[64] R.G. Huffman, A. Leduc, C. Wichmann, M. Di Gioia, F. Borriello, H. Specht, J. Derks, S. Khan, L. Khoury, E. Emmott, A.A. Petelski, D.H. Perlman, J. Cox, I. Zanoni, N. Slavov, Prioritized mass spectrometry increases the depth, sensitivity and data completeness of single-cell proteomics, Nat. Methods 20 (2023) 714–722, https://doi.org/10.1038/S41592-023-01830-1;SUBJMETA=1647,2067,250,262,296,631;KWRD=INNATE+IMMUNITY,MASS+SPECTROMETRY,PROTEOMIC+ANALYSIS.
[65] N. Gupta, P.A. Pevzner, False discovery rates of protein identifications: a strike against the Two-Peptide rule, J. Proteome Res. 8 (2009) 4173, https://doi.org/10.1021/PR9004794.
[66] S. Carr, R. Aebersold, M. Baldwin, A. Burlingame, K. Clauser, A. Nesvizhskii, The need for guidelines in publication of peptide and protein identification data: working group on publication guidelines for peptide and protein identification data, Mol. Cell. Proteom. 3 (2004) 531–533, https://doi.org/10.1074/MCP.T400006-MCP200.
[67] S.D. Keene, T.M. Greco, L. Parastatidis, S.H. Lee, E.G. Hughes, R.J. Balice-Gordon, D.W. Speicher, H. Ischiropoulos, Mass spectrometric and computational analysis of cytokine-induced alterations in the astrocyte secretome, Proteomics 9 (2009) 768, https://doi.org/10.1002/PMIC.200800385.
[68] C.R. Harris, K.J. Millman, S.J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N.J. Smith, R. Kern, M. Picus, S. Hoyer, M.H. van Kerkwijk, M. Brett, A. Haldane, J.F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, T.E. Oliphant, Array programming with NumPy, 2020 5857825, Nat 585 (2020) 357–362, https://doi.org/10.1038/s41586-020-2649-2.
[69] F. Pedregosa Fabianpedregosa, V. Michel, O. Grisel Oliviergrisel, M. Blondel, P. Prettenhofer, R. Weiss, J. Vanderplas, D. Cournapeau, F. Pedregosa, G. Varoquaux, A. Gramfort, B. Thirion, O. Grisel, V. Dubourg, A. Passos, M. Brucher, M. Perrot, Édouardand, Édouard Duchesnay, Fré Duchesnay Edouardduchesnay, Scikit-learn: machine learning in python Gaël varoquaux bertrand thirion vincent dubourg alexandre passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu perrot, J. Mach. Learn. Res. 12 (2011) 2825–2830. ⟨http://scikit-learn.sourceforge.net⟩. accessed June 25, 2024.
[70] The pandas development team, pandas-dev/pandas: Pandas (v2.2.2), (2024). https://doi.org/Https://doi.org/10.5281/zenodo.10957263.
[71] J.D. Hunter, Matplotlib: a 2D graphics environment, Comput. Sci. Eng. 9 (2007) 90–95, https://doi.org/10.1109/MCSE.2007.55.
[72] Plotly Technologies Inc., Collaborative data science, (2015). ⟨https://plot.ly⟩.
[73] M.L. Waskom, Seaborn: statistical data visualization, J. Open Source Softw. 6 (2021) 3021, https://doi.org/10.21105/JOSS.03021.
[74] M. Rahmatbakhsh, A. Gagarinova, M. Babu, Bioinformatic analysis of temporal and spatial proteome alternations during infections, Front. Genet. 12 (2021) 667936, https://doi.org/10.3389/FGENE.2021.667936/BIBTEX.
[75] B. Ghesquière, K. Gevaert, Proteomics methods to study methionine oxidation, Mass Spectrom. Rev. 33 (2014) 147–156, https://doi.org/10.1002/MAS.21386.
[76] C. Solazzo, J. Wilson, J.M. Dyer, S. Clerens, J.E. Plowman, I. Von Holstein, P. Walton Rogers, E.E. Peacock, M.J. Collins, Modeling deamidation in sheep α-keratin peptides and application to archeological wool textiles, Anal. Chem. 86 (2014) 567–575, https://doi.org/10.1021/AC4026362.
[77] J. Jeong, Y. Jung, S. Na, J. Jeong, E. Lee, M.S. Kim, S. Choi, D.H. Shin, E. Paek, H. Y. Lee, K.J. Lee, Novel oxidative modifications in redox-active cysteine residues, Mol. Cell. Proteom. 10 (2011), https://doi.org/10.1074/MCP.M110.000513.
[78] S.T. Sherry, M. Ward, K. Sirotkin, dbSNP-Database for single nucleotide polymorphisms and other classes of minor genetic variation, Genome Res. 9 (1999) 677–679, https://doi.org/10.1101/GR.9.8.677.
[79] S. Dasari, L. Pereira, A.P. Reddy, J.E.A. Michaels, X. Lu, T. Jacob, A. Thomas, M. Rodland, C.T. Roberts, M.G. Gravett, S.R. Nagalla, Comprehensive proteomic analysis of human cervical-vaginal fluid, J. Proteome Res. 6 (2007) 1258–1268, https://doi.org/10.1021/PR0605419.
[80] G. Zegels, G.A.A. Van Raemdonck, E.P. Coen, W.A.A. Tjalma, X.W.M. Van Ostade, Comprehensive proteomic analysis of human cervical-vaginal fluid using colposcopy samples, Proteome Sci. 7 (2009) 1–16, https://doi.org/10.1186/1477-5956-7-17/FIGURES/6.
[81] S.J. Li, M. Peng, H. Li, B.S. Liu, C. Wang, J.R. Wu, Y.X. Li, R. Zeng, Sys-BodyFluid: a systematical database for human body fluid proteome research, Nucleic Acids Res. 37 (2009) D907–D912, https://doi.org/10.1093/NAR/GKN849.
[82] J. Mathew, P. Sankar, M. Varacallo, Physiology, Blood Plasma, StatPearls, 2023. ⟨https://www.ncbi.nlm.nih.gov/books/NBK531504/⟩ (accessed April 7, 2024).
[83] H. Yang, B. Zhou, M. Prinz, D. Siegel, Proteomic analysis of menstrual blood, Mol. Cell. Proteom. 11 (2012) 1024, https://doi.org/10.1074/MCP.M112.018390.
[84] Y.V. Karpievitch, A.D. Polpitiya, G.A. Anderson, R.D. Smith, A.R. Dabney, Liquid chromatography mass spectrometry-based proteomics: biological and technological aspects, Ann. Appl. Stat. 4 (2010) 1797–1823, https://doi.org/10.1214/10-AOAS341.
[85] R.M. John, M.A. Surani, Imprinted genes and regulation of gene expression by epigenetic inheritance, Curr. Opin. Cell Biol. 8 (1996) 348–353, https://doi.org/10.1016/S0955-0674(96)80008-1.
[86] L. Gent, M.E. Chiappetta, S. Hesketh, P. Palmowski, A. Porter, A. Bonicelli, E. C. Schwalbe, N. Procopio, Bone proteomics method optimization for forensic investigations, cite this, J. Proteome Res. 23 (2024) 1844–1858, https://doi.org/10.1021/acs.jproteome.4c00151.
[87] R. Davidovics, Y.L. Saw, C.O. Brown, M. Prinz, H.E. McKiernan, P.B. Danielson, K. M. Legg, High-throughput seminal fluid identification by automated immunoaffinity mass spectrometry, J. Forensic Sci. 67 (2022) 1184–1190, https://doi.org/10.1111/1556-4029.14975.