Avens Publishing Group
J Forensic Investigation
June 2026 Vol.:13, Issue:1

Comparative Evaluation of Salivary Proteomic Profiling and STR DNA Typing for Forensic Subject Discrimination

Open Access Research Article

Journal of Forensic Investigation

Doan H[1], Hogan C[1], Viray J[2] and Giulivi C[1,3]*

[1]*Department of Molecular Biosciences, School of Veterinary Medicine, Davis, CA 95616,*
[2]*Sacramento District Attorney's Office, Biology Laboratory, Sacramento, CA 95814*
[3]*Medical Investigations of Neurodevelopmental Disorders (M.I.N.D.) Institute, University of California Davis, CA 95817*

***Address for Correspondence**
Cecilia Giulivi, Department of Molecular Biosciences, School of Veterinary Medicine, 1089 Veterinary Medicine Drive, Davis, CA, USA, E-mail: cgiulivi@ucdavis.edu

Submission: 18 March, 2026
Accepted: 29 May, 2026
Published: 03 June, 2026

Keywords

Forensic identification; short tandem repeat (STR) typing; DNA analysis; salivary proteomics; liquid chromatography–tandem mass spectrometry (LC–MS/MS); principal component analysis (PCA); forensic biology; Random Match Probability (RMP)

Abstract

Identifying persons of interest from biological evidence is central to establishing probative value in forensic investigations. Short tandem repeat (STR) DNA typing remains the forensic gold standard due to its statistical robustness and high inter-individual discriminatory power. However, low-template DNA, degradation, and complex mixtures may limit profile clarity. In contrast, forensic proteomics leverages the relative abundance and chemical stability of proteins and may provide complementary biological information.

This study directly compared subject discrimination using STR genotyping and salivary proteomic profiling. Saliva samples from 41 individuals (including three samples from the same individual collected across ~2 years) were analyzed by LC–MS/MS, and protein abundance profiles were evaluated using unsupervised principal component analysis (PCA) with Euclidean distance metrics. DNA genotyping was performed on 16 samples using real-time qPCR quantification, PowerPlex ® Fusion STR amplification, and capillary electrophoresis.

Full STR profiles were obtained for all DNA samples, and identical alleles across all loci confirmed perpetrator identity through population-based Random Match Probability (RMP) calculations. Proteomic PCA of 269 shared proteins explained 20.1% (PC1) and 15.0% (PC2) of total variance; reduction to 10 discriminatory proteins increased explained variance to 27.6% and 20.2%, respectively. Perpetrator-derived proteomic samples formed significantly tighter clusters than unrelated individuals (Mann–Whitney U = 0, p = 1.88 × 10^{-4}). Notably, samples collected approximately two years apart remained more similar to each other than to any other subject in multivariate space.

While STR typing provides deterministic identification through locus-by-locus allele concordance, proteomic profiling captures continuous, biologically dynamic variation. These findings support the concept that salivary proteomics may serve as a complementary forensic tool, particularly when DNA quality or quantity is compromised, though further validation and the development of a statistical framework are required before evidentiary implementation.

Introduction

DNA genotyping has been the "Gold Standard" in forensic science, serving as a critical identification method and a noninvasive means of collecting reference profiles [1]. One of the primary advantages of short-tandem repeat (STR) typing is the visual discrimination provided by electropherogram peaks, which allow analysts to determine the alleles present in a sample, assess potential contributions from multiple individuals, and estimate the number of contributors through peak height ratios at each locus [2-5].

DNA can be extracted from numerous biological materials encountered at crime scenes, including saliva deposited on cigarette butts, glassware, utensils, and gum, as well as from half-eaten food, vomit, and telephone receivers [6]. In sexual assault cases, saliva recovered from the neck, face, breast, and/or genitalia—often in conjunction with genital fluids—can corroborate victim testimony and link offender(s) to the crime scene. In routine forensic workflows, saliva is first identified using presumptive and confirmatory assays targeting α-amylase, including immunochromatographic strip tests and radial diffusion methods, prior to downstream DNA analysis.

High-quality and sufficient-quantity DNA are essential for reliable identification. Samples encountered in forensic investigations are frequently partially degraded, contain low template amounts, and may include PCR inhibitors. Individually or collectively, these factors influence both the efficiency of laboratory processing and the reliability of results [7] DNA degradation or inefficient amplification [3, 4, 6, 8] can compromise interpretability [4, 6, 8, 9] due to intrinsic chemical instability. Biological samples may be exposed to harsh environmental conditions prior to collection, including elevated temperatures, ultraviolet radiation, high humidity, and bacterial nucleases that degrade template DNA [2, 4, 6, 8, 9]. The ribose backbone and nitrogenous bases are particularly susceptible to depurination, oxidation, and hydrolysis [10], whereas proteins and peptides are generally more stable outside physiological environments [11].

Although STR markers provide robust genetic linkage to an individual, they are located within noncoding regions of chromosomes and therefore offer limited phenotypic or contextual information [12]. To address limitations associated with degraded or low-template samples, alternative genetic approaches have been developed, including mitochondrial DNA (mtDNA), single-nucleotide polymorphisms (SNPs) [13], insertion/deletion (indels)

ISSN: 2330-0396

polymorphisms [14], and mini-STRs [15]. DNA typing remains highly effective when contributors can be clearly resolved [3-5, 16] however, complex mixtures and environmentally compromised samples continue to pose interpretive challenges.

Importantly, DNA genotyping does not provide information regarding an individual's disease status, age, lifestyle factors, or recent exposures. In contrast, saliva transfer—such as that occurring in bite marks—may provide insight into sex and age [17], geographical region [18, 19], smoking habits or drug use [20, 21], and disease states through the analysis of salivary analytes, including proteins [22-26]. While such information may enhance investigative leads, it also raises important ethical and privacy considerations that warrant broader discussion.

Collectively, these findings support the exploration of proteomic profiling as either an alternative or complementary identification strategy. Degraded DNA and mixed-source samples remain difficult to resolve [2-5, 27, 28], whereas many salivary proteins are highly conserved across individuals [22, 26, 29], with a subset exhibiting discriminatory potential [30, 31]. Recent proteomic investigations have demonstrated that salivary protein profiles can discriminate between individuals using principal component analysis (PCA), reducing hundreds of detected proteins to a focused subset of highly discriminatory markers that cluster uniquely by subject.

This study, therefore, aims to critically evaluate the comparative and complementary value of STR-based DNA typing and salivary proteomics for subject identification in forensic casework.

Materials and Methods

Study design and sample collection

All saliva collection procedures were performed as previously described [30, 31] and were approved by the University of California, Davis Institutional Review Board (IRBNet ID: 1544585-1; approved 4/17/2020). The work described has been carried out in accordance with the World Medical Association's Declaration of Helsinki. Written informed consent was obtained from all participants. A total of 41 saliva samples were collected from female volunteers aged 20-61 years. Eleven samples were collected on October 22, 2020 (B1 cohort), and thirty samples were collected on February 4, 2022 (B2/B3 cohort). Subjects had abstained from eating, drinking, or oral hygiene procedures for 15-30 minutes prior to collection. Participants rinsed with water, allowed unstimulated saliva to pool for 60 seconds, and expectorated 0.1–1 mL into sterile containers. Samples were stored at −20°C until processing. The October 2020 cohort (n = 11) included a designated "crime scene" sample (CH14) and 10 unrelated subjects. The February 2022 cohort included 15 subjects, one of whom corresponded to the perpetrator represented in the 2020 cohort. Saliva samples from this second cohort were divided into two equal aliquots: one untreated (B2) and one treated with starch to remove amylase and glycosylated proteins (B3), as described previously [30]. Untreated samples were labeled CH#, and starch-treated samples were labeled CH#.1. The perpetrator samples were labeled CH14.1 (starch-treated) and CH14.2 (untreated). The original crime scene sample remained labeled CH14.

For DNA genotyping, saliva samples from the 2022 cohort (n = 15) were analyzed along with an additional sample collected from the perpetrator on April 19, 2023 (n = 16 total DNA samples). All samples were stored at -20°C until further processing.

Saliva proteomics

Sample processing – Whole saliva samples were precipitated with four volumes of −20°C analytical-grade acetone (Sigma-Aldrich, St. Louis, MO) and incubated overnight at 4°C. Samples were centrifuged at 16,000 × g for 10 minutes at 4°C. Pellets were washed twice with −20°C acetone, centrifuged under the same conditions, and dried under vacuum for 15 minutes (SpeedVac). Protein pellets were solubilized in 100 µL of 6 M urea in 50 mM ammonium bicarbonate (pH 8.0). Samples were reduced with 2.5 µL of 5 mM dithiothreitol (DTT) for 30 minutes at 37°C and alkylated with 20 µL of 5 mM iodoacetamide (IAA) for 30 minutes in the dark. Excess IAA was quenched with 20 µL DTT for 10 minutes at room temperature. Proteins were digested with a mass spectrometry-grade rLys-C/Trypsin Gold mix (Promega, Madison, WI) at a 1:25 enzyme-to-protein ratio for 4 hours at 37°C. Urea concentration was diluted to <1 M with 50 mM ammonium bicarbonate, and digestion was continued overnight at 37°C. Peptides were desalted using Macro Spin Columns (The Nest Group, Ipswich, MA). Approximately 10–100 µg of peptide digest was subjected to mass spectrometry.

Liquid Chromatography and Tandem Mass Spectroscopy - Peptide digests were randomized prior to analysis and processed at the UC Davis Proteomics Facility using a Q Exactive Orbitrap mass spectrometer (Thermo Scientific) coupled to a Proxeon Easy-nLC II system with a nanospray source.

Peptides were loaded onto a 100 µm × 25 mm Magic C18 (200 Å, 5 µm) trap column and separated on a 75 µm × 150 mm Magic C18 (200 Å, 3 µm) analytical column using a 90-minute gradient at 300 nL/min. Full MS scans were acquired over 300–1600 m/z. The top 15 precursor ions were selected for high-energy collisional dissociation (HCD) using a 2.0 m/z isolation window, 27% normalized collision energy, and 5-second dynamic exclusion.

Protein Identification - MS/MS spectra were searched using X! Tandem (version Alanine 2017.2.1.4) against the HumanFR_crap05292020_rev database (149,657 entries). Search parameters included: (i) Trypsin specificity; (ii) Parent and fragment ion tolerances of 20 ppm; and (iii) Variable modifications: carbamidomethylation (Cys), oxidation (Met, Trp), deamidation (Asn, Gln), N-terminal pyro-Glu formation, ammonia loss, and selenocysteine modifications. Protein identifications were validated using Scaffold (v4.11.1, Proteome Software Inc.). Peptide identifications were accepted at >88% probability to achieve <0.5% false discovery rate (FDR). Protein identifications were accepted at >5% probability, FDR <5%, and required at least two unique peptides, as assigned by the Protein Prophet algorithm. Proteins sharing indistinguishable peptides were grouped according to parsimony principles. Relative protein abundance was estimated using weighted spectral counting.

Salivary protein profiling and statistical analyses- Proteomic data were normalized to the total spectral counts within each sample. Across datasets, 1,169 proteins were identified in the first cohort and 281 in the second; proteins present in all datasets were retained for further analysis (n = 269; (Supplementary Table 1). To correct

ISSN: 2330-0396

for technical variation and non-biological batch effects, data were processed using the EigenMS algorithm [32,33]

Corrected data were subjected to principal component analysis (PCA) using ClustVis [34]. To identify proteins contributing most strongly to inter-subject variability, PCA loading values from the full 269-protein dataset were ranked by absolute magnitude (Supplementary Table 2). The top 10 proteins were selected for reduced-dimensional PCA analysis. Additional statistical analyses were performed using GraphPad Prism 9.1.0. To quantify inter-sample similarity within PCA space, Euclidean distances were calculated using PC1 and PC2 coordinates exported from ClustVis. For each PCA model (full 269-protein dataset and reduced 10-protein dataset), a centroid representing the perpetrator-derived samples (CH14, CH14.1, CH14.2) was calculated by averaging their PC1 and PC2 values. Euclidean distance for each sample was then computed as $\sqrt{[(PC1\_sample - PC1\_centroid)^2 + (PC2\_sample - PC2\_centroid)^2]}$. Distances for perpetrator-derived samples were compared with those of all non-perpetrator samples using the Mann–Whitney U test (two-tailed). Statistical significance was defined as $P < 0.05$.

DNA Analyses

DNA Extraction - DNA was extracted from whole saliva using a modified QIAamp® DNA Blood Mini Kit protocol (QIAGEN, Germantown, MD) following the manufacturer's supplementary instructions for saliva. Briefly, saliva samples (0.1–1.0 mL) were diluted in Dulbecco's phosphate-buffered saline (DPBS), centrifuged at 1,800 × g for 5 minutes at 4°C, and the cell pellets were resuspended in DPBS. Samples were lysed with QIAGEN Protease and Buffer AL at 56°C for 1 hour, followed by ethanol addition and column-based purification using QIAamp spin columns. Wash steps were performed with Buffers AW1 and AW2, and DNA was eluted in 150 µL UltraPure™ water. Extracted DNA was stored at −20°C until analysis.

DNA Quantification- DNA quantification was performed using the Quantifiler® Trio DNA Quantification Kit (Thermo Fisher Scientific, Waltham, MA) on a QuantStudio™ 5 Real-Time PCR System. Standard curves were generated using serial dilutions down to 5 pg/µL. Reactions were prepared according to the manufacturer's instructions and amplified under the following conditions: 95°C for 2 minutes, followed by 40 cycles of 95°C for 9 seconds and 60°C for 30 seconds.

DNA input for downstream amplification was adjusted to 0.5–1.0 ng in a 15 µL reaction volume using amplification-grade water.

STR amplification and Capillary Electrophoresis - STR amplification was performed using the PowerPlex® Fusion 6C System (Promega) following the manufacturer's protocol. PCR was conducted on a Veriti™ 96-Well Thermal Cycler under the following conditions: 96°C for 1 minute; 29 cycles of 96°C for 5 seconds and 60°C for 1 minute; final extension at 60°C for 10 minutes; hold at 4°C.

Amplified products were separated using an Applied Biosystems™ 3500xL Genetic Analyzer with 36 cm capillary arrays and POP-4 polymer. Samples were prepared with Hi-Di™ formamide and WEN ILS 500 size standard, denatured at 95°C for 3 minutes, snap-cooled, and analyzed using the HID36_POP4XL module.

STR profile interpretation- Electropherograms were analyzed using GeneMapper® ID-X v1.6. The analytical threshold for allele calling was set at 100 relative fluorescence units (RFU), consistent with laboratory validation for the PowerPlex® Fusion 6C system. All profiles were single-source. Because participants did not consent to publication of full STR profiles, electropherograms and genotype frequency tables are not presented.

Random Match Probability (RMP) Calculations- Random match probabilities were calculated according to NRC II recommendations (1996) [35] For homozygous loci, the Balding–Nichol's equation $[p^2 + p(1 - p)\theta]$ was applied using $\theta = 0.01$. For heterozygous loci, Hardy–Weinberg expectations (2pq) were used. Allele frequencies were obtained from the NIST-revised U.S. STR population database [36]. Locus-specific probabilities were multiplied across all loci to obtain the overall RMP, and values were reported as 1/RMP.

Results

Salivary Profiling Using Proteomics

Shotgun proteomic analysis identified >2,000 proteins across all samples. For downstream comparative analysis, only proteins detected in all 41 samples were retained (n = 269; (Supplementary Table 1), minimizing missing-value bias and ensuring comparability across subjects.

Protein–protein interaction (PPI) network analysis using STRINGdb demonstrated significant biological connectivity (237 nodes, 2,823 edges; expected edges = 674; average node degree = 23.8; clustering coefficient = 0.49; PPI enrichment $p < 1 \times 10^{-16}$), supporting the biological coherence of the retained dataset (Figure 1A). Unsupervised k-means clustering (k = 2, determined by centroid stability and within-cluster variance) separated the proteins into two principal functional groups (Figure 1A). The larger cluster (n = 207) was enriched for secretory granule lumen and neutrophil degranulation pathways, whereas the smaller cluster (n = 25) was enriched for keratinization and intermediate filament organization. Gene Ontology enrichment confirmed expected salivary molecular functions, including endopeptidase inhibitor and antioxidant activities (Figure 1B–C). Comparison with previously published salivary proteomes demonstrated overlap with known salivary signatures, while identifying 143 proteins not previously reported (Figure 1D), indicating expanded proteomic depth rather than a technical artifact.

Batch Effect Assessment and Correction

Because sample collections occurred approximately 16 months apart, potential batch effects were formally evaluated. **Singular** value decomposition (SVD)-based normalization using EigenMS was applied to partition biological variance from systematic technical variation.

Visualization of uncorrected data demonstrated separation by collection cohort, consistent with batch-driven variance (Figure 2A–C). Following EigenMS correction, this separation was markedly reduced, indicating effective removal of systematic bias while retaining inter-individual variability.

All subsequent analyses were conducted on batch-corrected data

Citation: Doan H, Hogan C, Viray J, Giulivi C. Comparative Evaluation of Salivary Proteomic Profiling and STR DNA Typing for Forensic Subject Discrimination. J Forensic Investigation. 2026; 13(1): 1.

ISSN: 2330-0396

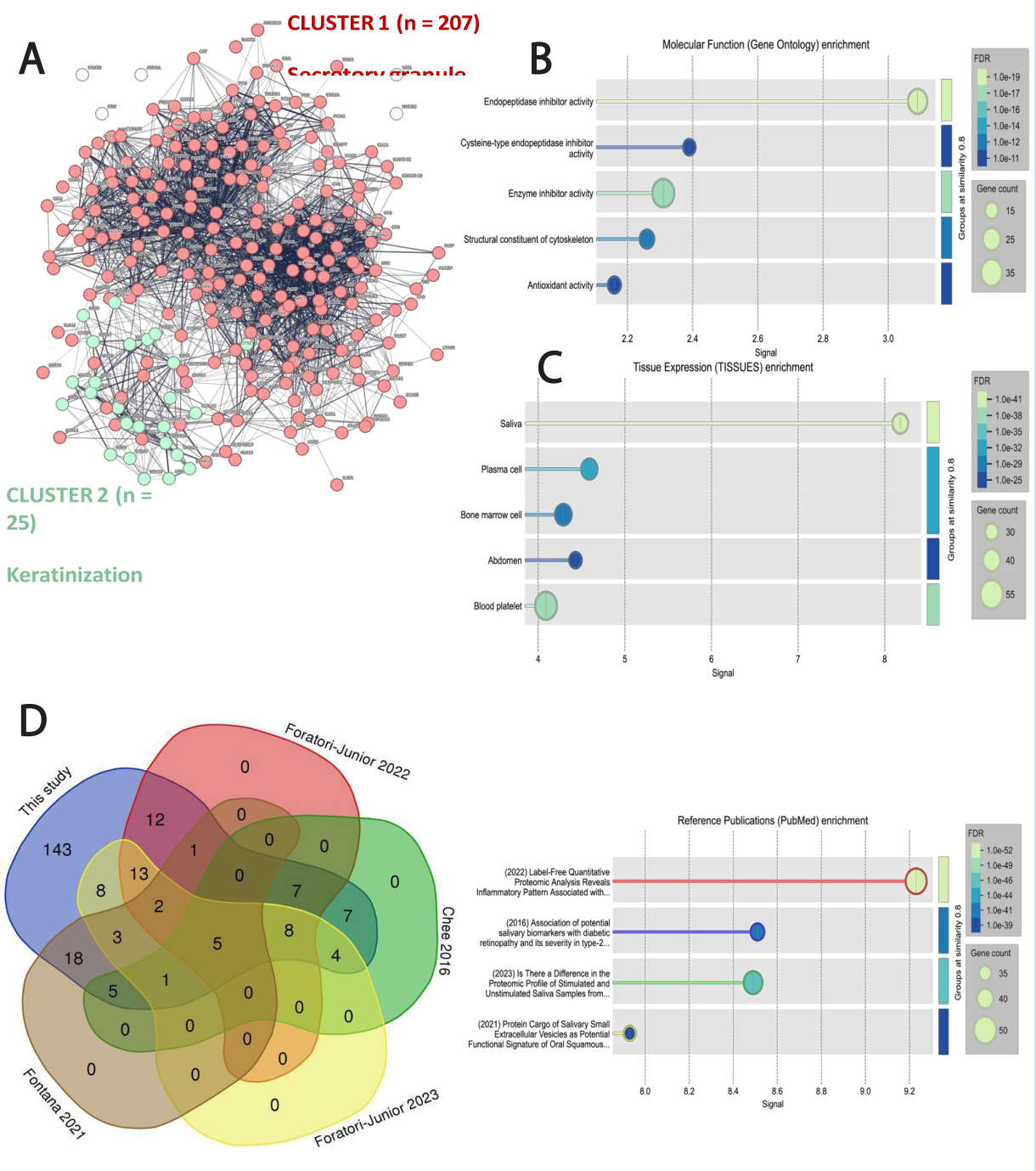

Figure 1: Functional characterization and literature comparison of the salivary proteome dataset. Panel A. Protein–protein interaction (PPI) network analysis of the 269 proteins retained for comparative analysis, generated using STRINGdb with whole-genome background enrichment. The network comprised 237 nodes and 2,823 edges (expected edges = 674), with an average node degree of 23.8 and an average local clustering coefficient of 0.49. The observed PPI enrichment was highly significant ($p < 1 \times 10^{-16}$), indicating greater connectivity than expected by chance. Nodes represent individual proteins, and edges represent high-confidence interactions. Unsupervised k-means clustering (k = 2) identified two principal functional clusters.
Panel B. Gene Ontology (GO) enrichment analysis of molecular function for the network proteins. The top five enriched terms are shown for clarity. Terms were grouped by similarity (similarity ≥ 0.8) and ranked by signal strength. Minimum term count = 2; minimum signal and strength thresholds = 0.01. Panel C. GO enrichment analysis for tissue-associated terms derived from the same protein set. The top four enriched tissue categories are shown. Parameters were identical to those used in panel B. Panel D. Literature overlap analysis. Left: Venn diagram illustrating overlap between the present dataset and four previously published salivary proteomic studies. Right: Enrichment of overlapping proteins across PubMed-indexed datasets; the top four overlapping studies are shown (38-41).

ISSN: 2330-0396

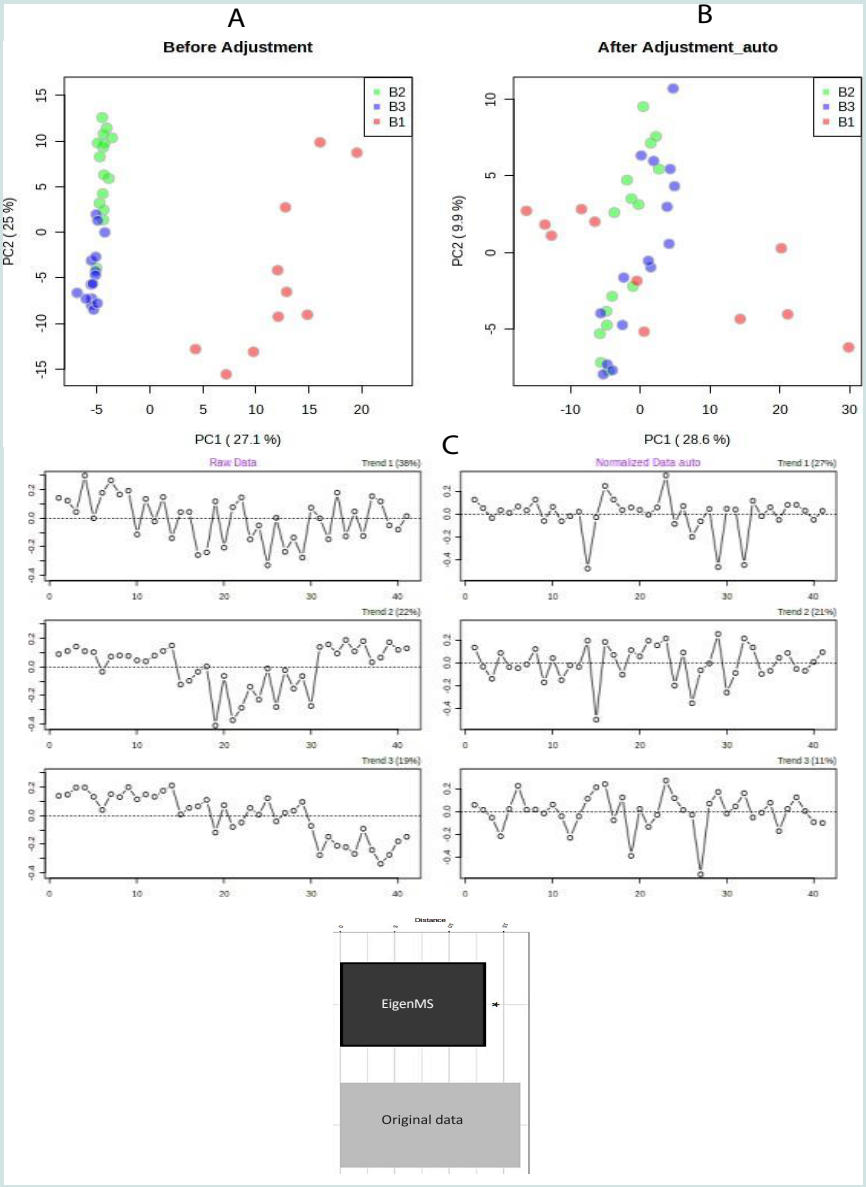

Figure 2: Assessment of batch effects and normalization of salivary proteomic data. Panel A. Principal component analysis (PCA) of salivary proteomic profiles before and after batch correction. Samples are color-coded by collection group: B1 (red; collected in 2020), B2 (green; collected in 2022 without starch treatment), and B3 (blue; collected in 2022 with starch pre-treatment to remove amylase and glycosylated proteins). The left panel shows clustering of raw data prior to normalization, with separation largely driven by the collection cohort. The right panel shows the PCA after EigenMS normalization, demonstrating reduced cohort-driven separation and improved sample integration. Percent variance explained by each principal component is indicated on the axes.
Panel B. Singular value decomposition (SVD) analysis used to identify systematic technical trends in the dataset. The left panel displays dominant variance components in the raw data,revealing structured batch-associated patterns. The right panel shows the corresponding components after EigenMS normalization, indicating attenuation of systematic bias.
Panel C: Quantification of inter-sample distances across experimental conditions before and after normalization. Bars represent mean pairwise Euclidean distances within and between groups; the asterisk indicates a statistically significant reduction in batch-associated separation following EigenMS correction (statistical test described in Methods).

that were log-transformed (ln[x+1]) and Pareto-scaled to stabilize variance and reduce dominance by highly abundant proteins.

Principal Component Analysis and Subject Discrimination

Unsupervised principal component analysis (PCA) was performed using singular value decomposition on the batch-corrected dataset (n = 41 samples; 269 proteins). PCA was used solely for dimensionality reduction and visualization, without incorporating class labels.

In the full proteome dataset, the first two principal components explained 20.1% (PC1) and 15.0% (PC2) of the total variance, respectively (Figure 2D). The three samples from the same perpetrator (CH14, CH14.1, CH14.2) clustered closely together in this reduced-

ISSN: 2330-0396

dimensional space and were separated from the majority of other subjects. Euclidean distances in PC1–PC2 space were calculated relative to the centroid of the perpetrator samples (CH14, CH14.1, CH14.2). Perpetrator-derived samples showed substantially smaller distances (0.0481 ± 0.0327) than non-perpetrator samples (0.524 ± 0.234; Mann–Whitney U = 0, $p = 1.88 \times 10^{-4}$), supporting statistically tighter clustering of perpetrator samples. The closest non-perpetrator sample (CH12.1; distance = 0.137) remained separated from the perpetrator cluster. Complete Euclidean distance values for all samples in both PCA models are provided in (Supplementary Table 3). Although PC1 and PC2 together accounted for 35.1% of total variance, clustering was observed without supervised modeling, indicating that subject-level variance contributed measurably to the dominant components.

Identification of Discriminatory Proteins

To identify proteins contributing most strongly to inter-subject variability, PCA loadings from the full dataset were examined (Supplemental Table 2). The top 10 proteins with the highest absolute loading values were selected as candidate discriminatory markers (AMY1A, GC, IGHA2, IGKC, JCHAIN, KRT13, KRT14, MUC5B, SPPR3, ZG16B).

In the reduced 10-protein PCA model (Figure 3B), PC1 and PC2 explained 27.6% and 20.2% of the total variance, respectively (47.8% cumulative). The increase in explained variance after feature reduction suggests that the selected proteins capture a greater proportion of subject-associated variability than the full proteome dataset. The three perpetrator-derived samples (CH14, CH14.1, CH14.2) formed a compact cluster in PC space. To quantify this separation, Euclidean distances were calculated in PC1–PC2 space relative to the centroid of the perpetrator samples. Perpetrator-derived samples exhibited significantly smaller distances to their centroid (0.0633 ± 0.0105) compared with all non-perpetrator samples (0.6269 ± 0.1616; Mann–Whitney U = 0, $p = 1.88 \times 10^{-4}$). The closest non-perpetrator sample (CH6.1; distance = 0.248) remained substantially separated from the perpetrator cluster. These results quantitatively confirm the enhanced discriminatory resolution observed visually in (**Figure 3C**). Collectively, these findings demonstrate that salivary proteomic profiling retains subject-specific structure capable of discriminating a crime scene sample from unrelated individuals, even after correction for batch effects and dimensionality reduction. The observed clustering and quantitative separation in PCA space suggest that a focused subset of discriminatory proteins enhances resolution beyond that achieved with the full proteome. To contextualize the forensic utility of this approach, we next compared proteomic-based discrimination with conventional STR DNA genotyping performed on the same cohort.

DNA STR Genotyping and PCA of Genotypic Frequencies STR Profile Quality and Interpretation

All saliva samples yielded single-source, full STR profiles using the PowerPlex® Fusion 6C system. Stutter peaks were observed at expected positions (N ± 4 repeat units) and were identified according to established analytical criteria. Complete 1/RMP values calculated for all four U.S. populations and for the single combined population are provided in (Supplementary Table 5A, 5B), respectively. Several samples displayed minor artifacts consistent with pull-up peaks. Sample A3 showed a 102 RFU pull-up at locus D12S391 associated with a strong FGA allele (>1800 RFU). Sample A13 displayed three pull-up peaks at Amelogenin, D3S1358, and SE33, attributable to strong neighboring loci. These artifacts were below true allele peak

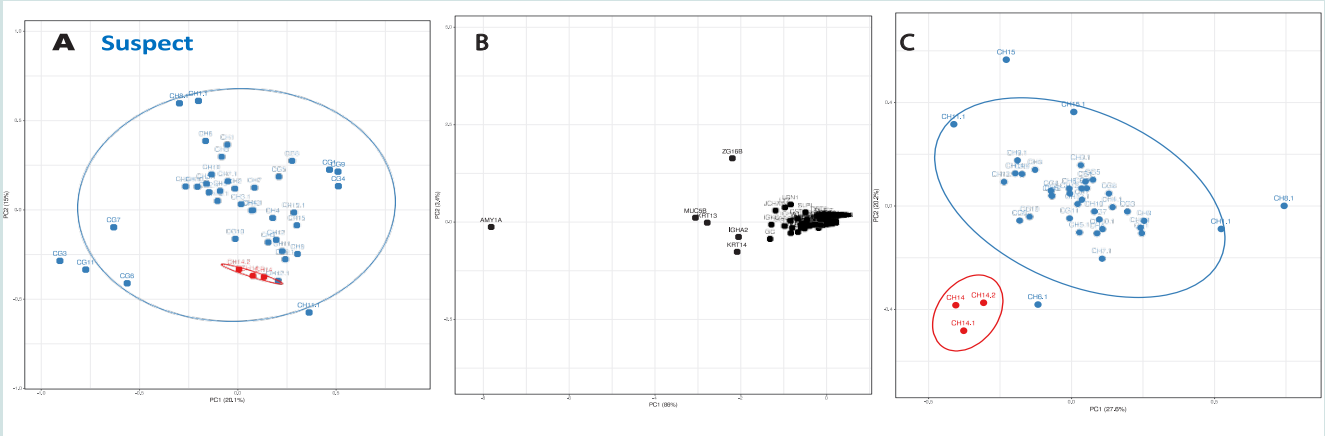

Figure 3: Identification of discriminatory salivary proteins and subject separation in reduced-dimensional space.
Panel A. Principal component analysis (PCA) of all 41 salivary samples using the full set of 269 shared proteins after batch correction and preprocessing (ln[x+1] transformation and Pareto scaling). Samples corresponding to unrelated individuals ("suspects") are shown in blue, and perpetrator-derived samples (CH14, CH14.1, CH14.2) are shown in red. Ellipses represent 95% confidence intervals. PC1 and PC2 explained 20.1% and 15.0% of the total variance, respectively.
Panel B. PCA loadings from the full dataset were examined to identify proteins contributing most strongly to inter-subject variability. The top 10 proteins with the highest absolute loading values were selected as candidate discriminatory markers. Panel C. PCA of salivary samples using the 10 selected proteins (AMY1A, GC, IGHA2, IGKC, JCHAIN, KRT13, KRT14, MUC5B, SPPR3, ZG16B).
Reduction to this subset increased explained variance (PC1 = 27.6%, PC2 = 20.2%) and resulted in tighter clustering of perpetrator-derived samples relative to unrelated individuals. Ellipses represent 95% confidence intervals.

ISSN: 2330-0396

intensities and did not interfere with genotype interpretation. Sample A1 exhibited a potential tri-allelic pattern between loci D12S391 and D19S433, observed as an off-ladder allele. Re-amplification reproduced the tri-allelic signal (243 RFU) with expected stutter patterning. GeneMapper® ID-X flagged locus D12S391; however, the tri-allelic pattern did not alter single-source profile interpretation. To our knowledge, tri-allelic patterns at this locus are rare and not commonly reported in public databases. All profiles were suitable for statistical evaluation.

PCA of Genotypic Frequencies (All 23 Loci)

To explore variance structure among DNA profiles, a two-dimensional principal component analysis (PCA) was performed using locus-specific genotypic frequencies calculated under Hardy–Weinberg or Balding–Nichol's expectations ($\theta = 0.01$), based on NIST allele frequency data (Hill et al., 2013). One-population frequency estimates were used to maintain uniform scaling across samples for PCA visualization.

Using all 23 loci (Figure 4A), PC1 and PC2 explained 18.0% and 16.5% of total variance, respectively (34.5% cumulative). PCA revealed substantial overlap among samples, reflecting the relatively small variance in genotypic frequencies across individuals. For example, samples A14 and A16 (perpetrator-derived samples) overlapped completely in PC space, consistent with identical STR profiles. However, additional samples (e.g., A7) occupied proximal positions in PCA space despite having distinct genotypes. Also, additional sample pairs exhibited limited spatial resolution in reduced-dimensional space (A1–A2, A3–A11, A5–A10; (Figure 4A). The limited dispersion observed in PCA likely reflects the constrained numerical range of genotypic frequencies derived from a finite population database (n = 1,036 individuals), resulting in relatively small between-profile variance (average variance = 0.003686).

PCA After Locus Reduction

To evaluate whether a subset of loci could improve separation in PCA space, loci contributing minimally to principal component loadings were removed. Absolute loading values from principal components through PC7 were examined to identify loci contributing minimally to variance structure (Supplementary Table 4), yielding a cumulative variance threshold of 83.3%. Six loci (D10S1248, D16S539, D2S1338, vWA, D5S818, SE33) were excluded, leaving 17 loci for reanalysis (Figure 4B).

The reduced-locus PCA demonstrated modest improvement in cluster separation; however, overlap among unrelated individuals remained, indicating that PCA of genotypic frequencies provides limited discriminatory resolution compared with conventional STR matching criteria.

Random Match Probability (RMP)

For forensic comparison, locus-specific genotypic frequencies were multiplied across all loci to calculate Random Match Probabilities (RMP) in accordance with NRC II guidelines. Because donor ethnicity was unknown, 1/RMP values were calculated using allele frequencies from the four major U.S. populations. Using this approach, only one exact match was observed: the perpetrator's samples A14 and A16 showed identical alleles at all loci, confirming their genetic identity. No unrelated sample produced a matching STR profile. Unlike proteomic profiling, which demonstrated measurable clustering in multivariate space under exploratory dimensionality reduction, STR identification relies on deterministic locus-by-locus comparison and population- based probability calculations, yielding unambiguous profile matching when full allelic concordance is present. The greater dispersion observed in proteomic PCA likely reflects continuous quantitative variability in protein abundance across individuals, whereas STR genotypic frequencies are constrained by discrete allele categories and bounded population-frequency distributions.

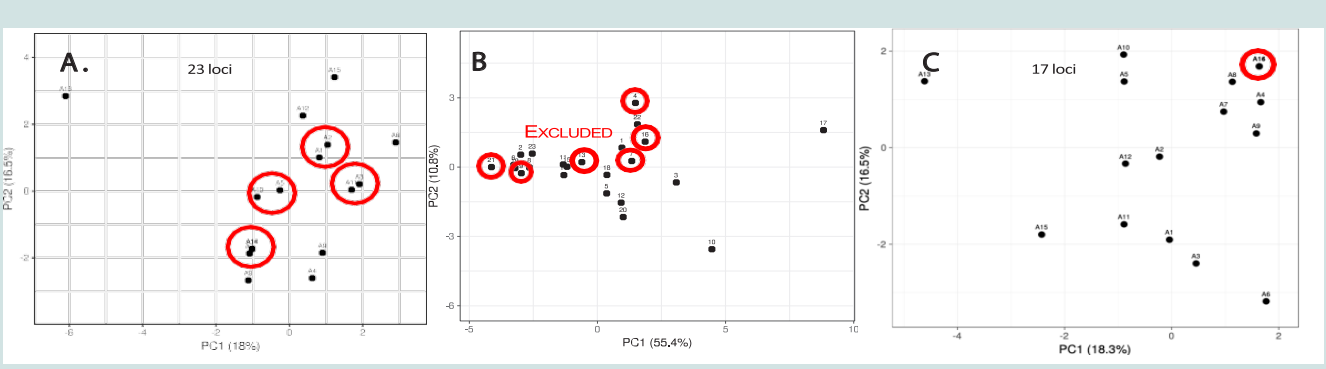

Figure 4: Principal component analysis (PCA) of STR genotypic frequencies.
Panel A: PCA of subjects using genotypic frequencies from all 23 STR loci. PC1 and PC2 explained 18.0% and 16.5% of total variance, respectively. Several sample pairs exhibited limited separation in reduced-dimensional space, including A1–A2, A3–A11, A5–A10, and the cluster A7–A14–A16 (circled in red). Notably, perpetrator-derived samples A14 and A16 overlapped completely, consistent with identical STR profiles. PCA was used for exploratory visualization only and does not substitute for locus-by-locus comparison or Random Match Probability (RMP) calculations.
Panel B: PCA loadings plot showing the contribution of individual STR loci to variance in principal component space. Loci with minimal loading contributions (D10S1248, D16S539, D2S1338, vWA, D5S818, SE33) are circled in red. These loci were excluded for exploratory reanalysis in panel
C. Low loading contributions in the PCA space do not imply reduced forensic discriminatory power.
Panel C: PCA of subjects using the reduced set of 17 STR loci selected based on loading magnitude. Modest improvement in sample dispersion was observed; however, overlap among unrelated individuals remained. Perpetrator-derived samples A14 and A16 (circled in red) remained superimposed, reflecting identical allele profiles across loci.

ISSN: 2330-0396

Discussion

Comparative Forensic Utility of Proteomics and STR Typing

The present study highlights fundamental methodological differences between salivary proteomic profiling and conventional STR DNA genotyping. STR typing operates through deterministic allele concordance across defined loci, with statistical weight expressed as a population-based Random Match Probability (RMP). When full allelic concordance is observed, as in samples A14 and A16, identity is established within the statistical framework of population genetics through RMP calculations. In contrast, proteomic profiling relies on multivariate quantitative variation in protein abundance, capturing subject-specific biological structure in reduced-dimensional space. The PCA and Euclidean distance analyses demonstrated that salivary protein signatures retain measurable discriminatory structure even after correction for batch effects, with perpetrator-derived samples forming statistically tighter clusters than unrelated individuals. The reduced PCA was intended to visualize variance structure after feature reduction and does not constitute independent validation of discriminatory performance, as the same dataset was used for feature selection and visualization.

Importantly, these approaches address different aspects of forensic identification. STR typing provides categorical identity confirmation but is limited in its ability to resolve degraded samples, low-template DNA, or complex mixtures, and does not convey phenotypic or physiological information. Proteomic profiling, while inherently multivariate and exploratory in its current implementation, captures biologically informative variation that may complement DNA typing, particularly when DNA quantity or quality is compromised. The enhanced separation observed with a reduced discriminatory protein panel further suggests that targeted proteomic markers may improve resolution. The loci removed in the reduced PCA model were selected solely on the basis of low loading contributions within the exploratory principal component framework of this dataset. Importantly, low PCA loading does not imply reduced forensic discriminatory power. Several of the excluded loci (e.g., SE33, D2S1338, vWA) are highly polymorphic and contribute substantially to match probability calculations in standard STR analysis. Their limited contribution in the present PCA likely reflects cohort-specific genotype distribution and the constrained variance structure of population-based frequency values rather than the intrinsic weakness of the markers. Thus, rather than serving as a replacement for STR genotyping, salivary proteomics may function as a complementary evidentiary layer— providing additional discriminatory structure or contextual biological information under conditions where traditional DNA-based approaches encounter analytical challenges.

Practical and Analytical Considerations

The PowerPlex® Fusion system has been validated to generate full STR profiles from as little as 0.125 ng of template DNA [37], underscoring the high sensitivity of modern STR typing. In contrast, proteomic profiling required milligram-scale total protein input for LC–MS/MS analysis. Both workflows required approximately one working day from extraction to analytical output; however, DNA typing relies on multiple proprietary kits, locus-specific primers, fluorescent dyes, and capillary electrophoresis instrumentation. Proteomic workflows depend primarily on LC–MS/MS instrumentation and downstream computational analysis. Although LC–MS/MS instrumentation represents a substantial capital investment, it enables high-dimensional biological characterization beyond identity testing alone.

When evaluating analytical robustness, STR typing benefits from decades of developmental validation, population database construction, and standardized statistical interpretation. In the present study, full allelic concordance between A14 and A16 provided categorical confirmation of identity. Proteomic profiling, while able to cluster perpetrator-derived samples distinctly from unrelated individuals, remains inherently multivariate and probabilistic in its current implementation.

Temporal stability further distinguishes the two approaches. STR genotypes remained identical across time points, as expected for germline DNA markers. Notably, despite sampling intervals spanning approximately 2 years, perpetrator-derived proteomic profiles remained significantly closer to one another than to any unrelated individual in PCA space, as quantified by centroid-based Euclidean distance analysis, indicating that intra-individual similarity exceeded inter-individual variability under the present analytical framework. Such small variability may reflect biological influences, including age, environmental exposure, and physiological state.

Nevertheless, even after correction for batch effects and reduction to discriminatory protein subsets, perpetrator-derived samples maintained statistically significant clustering relative to unrelated individuals.

Principal component analysis revealed that proteomic data accounted for a greater proportion of variance in the dominant components (PC1 = 27.6%, PC2 = 20.2%) than STR genotypic frequency PCA (PC1 = 18.0%, PC2 = 16.5%). This difference likely reflects the continuous quantitative nature of protein abundance compared with the constrained frequency range of allele-based genotypes. However, STR discriminatory power arises not from multivariate dispersion but from the multiplicative combination of locus-specific genotype probabilities across loci. Consequently, reduced separation in PCA space does not imply reduced forensic discrimination for STR typing, as evidentiary weight is derived from locus-by-locus probability calculations rather than dimensional variance structure. For forensic proteomics to achieve courtroom viability, statistical frameworks analogous to those used in DNA typing will be required. The identification of genetically variable peptides and construction of population frequency databases may allow calculation of likelihood ratios or RMP-like statistics, placing proteomic evidence on a comparable statistical footing with STR analysis.

Limitations

Several limitations of the present study should be acknowledged. First, the cohort size was modest (n = 41 for proteomics; n = 16 for DNA comparison), limiting the generalizability of the findings and precluding robust population-level statistical modeling of proteomic variability. While clear clustering was observed for perpetrator-derived samples, larger, more diverse cohorts will be necessary to assess false-positive rates, inter-individual overlap, and classification stability.

ISSN: 2330-0396

Second, proteomic profiling was evaluated using principal component analysis and Euclidean distance metrics, which are exploratory and visualization-oriented techniques. Although statistically significant separation was observed, PCA does not constitute a predictive or classification model. Future studies should incorporate supervised machine learning approaches with cross-validation or external validation cohorts to quantify classification accuracy, sensitivity, and specificity.

Third, salivary protein expression is influenced by biological variables including age, sex, circadian rhythm, diet, health status, and environmental exposures. Although batch correction (EigenMS) was applied to mitigate technical variation, biological variability across time points was observed in longitudinal samples. This temporal variability underscores the need to characterize intra-individual stability over extended intervals before proteomic profiling can be considered a deterministic identification method.

Fourth, protein abundance was estimated using spectral counting, which provides semi-quantitative measurements. While sufficient for comparative profiling, more precise quantification methods (e.g., MS1 intensity-based quantification or targeted proteomics) may improve reproducibility and discriminatory resolution.

Fifth, STR PCA analyses were conducted using one-population allele frequency estimates to standardize visualization. While appropriate for exploratory multivariate comparison, this simplification reduces population-specific accuracy and does not reflect standard forensic reporting practices. Importantly, identity conclusions were based solely on full allele concordance and RMP calculations.

Finally, proteomic profiling currently lacks established population databases and widely accepted statistical frameworks analogous to Random Match Probability or likelihood ratios used in forensic DNA typing. The development of frequency databases for genetically variable peptides and standardized interpretive guidelines will be essential before proteomic evidence can be considered for courtroom application.

Concluding remarks

This study demonstrates that salivary proteomic profiling captures subject-specific biological structure that can discriminate a crime scene sample from unrelated individuals in multivariate space. After correction for batch effects and dimensionality reduction, perpetrator-derived samples formed statistically tighter clusters than non-perpetrator samples, and quantitative Euclidean distance analysis confirmed significant separation in principal component space. These findings indicate that salivary protein abundance patterns retain measurable subject-associated structure under exploratory multivariate analysis. In contrast, conventional STR DNA typing provided categorical identity confirmation through full allelic concordance and population-based Random Match Probability calculations. As expected, STR profiles remained temporally stable and yielded unambiguous matching when alleles were identical across loci.

The comparative results highlight fundamental differences between the two approaches: STR genotyping is deterministic and population-statistically validated, whereas proteomic profiling is continuous, biologically dynamic, and currently exploratory. However, proteomic analysis offers potential advantages in contexts where DNA quantity or quality is compromised and may provide additional contextual biological information not accessible through noncoding STR markers.

Further development of standardized workflows, larger population datasets, quantitative validation studies, and statistical interpretive frameworks will be necessary before forensic proteomics can approach the evidentiary maturity of STR DNA typing. Nevertheless, the present findings support the concept that salivary proteomics may serve as a complementary forensic tool, augmenting traditional DNA-based identification rather than replacing it.

Acknowledgments

This study was supported by discretionary funds (C.G.).

Conflict of Interest: No potential financial and non-financial competing interests that could directly or indirectly undermine the objectivity, integrity, and value of this publication through a possible influence on the judgments and actions of authors regarding objective data presentation, analysis, and interpretation were found. C.G. is an Editorial Board Member of Scientific Reports (Nature Publishing Company). She received compensation as Field Chief Editor for Frontiers in Molecular Biosciences and honoraria from participating in NIH peer review meetings. C.H. and H.D. have no conflict of interest to report.

Highlights

- Salivary proteomes retain subject-specific multivariate structure.
- STR typing provides deterministic identity via allele concordance.
- PCA-based proteomics shows intra-individual similarity over time.
- Proteomics may complement DNA in low-template conditions.
- Comparative analysis clarifies the strengths and limitations of both methods.

References

1. Chatterjee S (2019) Saliva as a forensic tool. J Forensic Dent Sci 11: 1-4.
2. Taupin JM (2019) Interpreting Complex Forensic DNA Evidence. 1 ed: CRC Press 2019: 11-14.
3. Butler JM (2015) The future of forensic DNA analysis. Philos Trans R Soc Lond B Biol Sci 370: 20140252.
4. Butler JM (2012)Advanced Topics in Forensic DNA Typing: Methodology: Elsevier.
5. Butler JM (2011) Forensic DNA testing. Cold Spring Harb Protoc.2011:1438-1450.
6. Butler JM (2010) Fundamentals of Forensic DNA Typing. Fundamentals of Forensic DNA Typing. 2010: 1-500.
7. Carrasco PA, Brizuela CI, Rodriguez IA, Munoz S, Godoy ME, et al. (2017) Histological transformations of the dental pulp as possible indicator of post mortem interval: a pilot study. Forensic Sci Int 279: 251-257.
8. Bright JA, Taylor D, Curran JM, Buckleton JS (2013) Degradation of forensic DNA profiles. Australian Journal of Forensic Sciences 45: 445-449.
9. Merkley ED (2019) Introduction to Forensic Proteomics. In: Merkley ED, editor. ACS Symposium Series. 1339. Washington, DC: American Chemical Society 2019: 1-8.

ISSN: 2330-0396

10. Lindahl T (1993) Instability and decay of the primary structure of DNA. Nature 362: 709-715.

11. Parker GJ, McKiernan HE, Legg KM, Goecker ZC (2021) Forensic proteomics. Forensic Sci Int Genet 54: 102529.

12. Fan H, Chu JY (2007) A brief review of short tandem repeat mutation. Genomics Proteomics Bioinformatics 5: 7-14.

13. Brenner CH, Weir BS (2003) Issues and strategies in the DNA identification of World Trade Center victims. Theor Popul Biol 63: 173-178.

14. Pereira R, Phillips C, Alves C, Amorim A, Carracedo A, et al. (2009) A new multiplex for human identification using insertion/deletion polymorphisms. Electrophoresis 30: 3682-3690.

15. Holland MM, Cave CA, Holland CA, Bille TW (2003) Development of a quality, high throughput DNA analysis procedure for skeletal samples to assist with the identification of victims from the World Trade Center attacks. Croat Med J 44: 264-272.

16. Gill P, Gusmao L, Haned H, Mayr WR, Morling N, Parson W, et al. (2012) DNA commission of the International Society of Forensic Genetics: Recommendations on the evaluation of STR typing results that may include drop-out and/or drop-in using probabilistic methods. Forensic Sci Int Genet. 6: 679-688.

17. Kalipatnapu P, Kelly RH, Rao KN, van Thiel DH (1983) Salivary composition: effects of age and sex. Acta Med Port 4: 327-330.

18. Cho HR, Kim HS, Park JS, Park SC, Kim KP, Wood TD, et al. (2017) Construction and characterization of the Korean whole saliva proteome to determine ethnic differences in human saliva proteome. PLoS ONE 12: e0181765.

19. Jain S, Ahmad Y, Bhargava K (2018) Salivary proteome patterns of individuals exposed to High Altitude. Arch Oral Biol 96: 104-112.

20. Chatterjee S (2019)Saliva as a forensic tool. J Forensic Dent Sci 11: 1-4.

21. Toennes SW, Steinmeyer S, Maurer HJ, Moeller MR, Kauert GF (2005) Screening for drugs of abuse in oral fluid--correlation of analysis results with serum in forensic cases. J Anal Toxicol 29: 22-27.

22. Al Kawas S, Rahim ZH, Ferguson DB (2012) Potential uses of human salivary protein and peptide analysis in the diagnosis of disease. Arch Oral Biol 57:1-9.

23. Lee YH, Zhou H, Reiss JK, Yan X, Zhang L, Chia D, et al. (2011) Direct saliva transcriptome analysis. Clin Chem 57: 1295-302.

24. Rajshekar M, Tennant M, Thejaswini BDS (2014) Salivary biomarkers and their applicability in forensic identification. Sri Lanka Journal of Forensic Medicine, Science & Law 4:10.

25. Range H, Leger T, Huchon C, Ciangura C, Diallo D, Poitou C, et al. (2012) Salivary proteome modifications associated with periodontitis in obese patients. J Clin Periodontol 39: 799-806.

26. Sivadasan P, Gupta MK, Sathe GJ, Balakrishnan L, Palit P, Gowda H, et al. (2015) Human salivary proteome--a resource of potential biomarkers for oral cancer. J Proteomics 127: 89-95.

27. McCord BO, Funes K, Zoppis M, Meadows Jantz SL (2011) An Investigation of the Effect of DNA Degradation and Inhibition on PCR Amplification of Single Source and Mixed Forensic Samples2011: 66.Hughes-Stamm SR, Ashton KJ, van Daal A (2011) Assessment of DNA degradation and the genotyping success of highly degraded samples. Int J Legal Med 125: 341-348.

28. Siqueira WL, Salih E, Wan DL, Helmerhorst EJ, Oppenheim FG (2008) Proteome of human minor salivary gland secretion. J Dent Res 87: 445-450.

29. Smith H, Giulivi C (2024) Starch treatment improves the salivary proteome for subject identification purposes. Forensic Sci Med Pathol 20: 117-128.

30. Thomas C, Giulivi C (2021) Saliva protein profiling for subject identification and potential medical applications. Medicine in Omics 3: 100012.

31. Karpievitch YV, Nikolic SB, Wilson R, Sharman JE, Edwards LM (2014) Metabolomics data normalization with EigenMS. PLoS ONE 9: e116221.

32. Karpievitch YV, Taverner T, Adkins JN, Callister SJ, Anderson GA, Smith RD, et al. (2009) Normalization of peak intensities in bottom-up MS-based proteomics using singular value decomposition. Bioinformatics 25: 2573-2580.

33. Metsalu T, Vilo J (2015) ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. Nucleic Acids Res 43:W566-570.

34. (U.S.) NRC (1996) The evaluation of forensic DNA evidence. Committee on DNA Forensic Science: an Update., Update. NRCUSCoDFSa, editors. Washington, D.C.: National Academy Press 1996.

35. Hill CR, Duewer DL, Kline MC, Coble MD, Butler JM (2013) US. population data for 29 autosomal STR loci. Forensic Sci Int Genet 7: e82-e83.

36. Cisana S, Cerri N, Bosetti A, Verzeletti A, Cortellini V (2017) PowerPlex(R) Fusion 6C System: evaluation study for analysis of casework and database samples. Croat Med J 58: 26-33.

37. Foratori-Junior GA, Ventura TMO, Grizzo LT, Carpenter GH, Buzalaf MAR, Sales-Peres SHC (2022) Label-Free Quantitative Proteomic Analysis Reveals Inflammatory Pattern Associated with Obesity and Periodontitis in Pregnant Women. Metabolites 12: 1091.

38. Chee CS, Chang KM, Loke MF, Angela Loo VP, Subrayan V (2016) Association of potential salivary biomarkers with diabetic retinopathy and its severity in type-2 diabetes mellitus: a proteomic analysis by mass spectrometry. PeerJ 4: e2022.

39. Foratori-Junior GA, Ventura TMO, Grizzo LT, Jesuino BG, Castilho A, Buzalaf MAR, et al. (2023) Is There a Difference in the Proteomic Profile of Stimulated and Unstimulated Saliva Samples from Pregnant Women with/without Obesity and Periodontitis? Cells 12: 1389.

40. Fontana S, Mauceri R, Novara ME, Alessandro R, Campisi G (2021) Protein Cargo of Salivary Small Extracellular Vesicles as Potential Functional Signature of Oral Squamous Cell Carcinoma. Int J Mol Sci 22: 11160.