



LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

Development of a Protein-based Human Identification Capability from a Single Hair

K. E. Mason, P. H. Paul, D. S. Anex, B. R. Hart

June 11, 2018

Journal of Forensic Science

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

TITLE

Development of a protein-based human identification capability from a single hair

KEYWORDS

Forensic Science

Human Identification

Genetically Variant Peptide

Single Hair

Proteomics

Random Match Probability

ABSTRACT

Shed-human hair (lacking root nuclear DNA) frequently contributes important information to forensic investigations involving human identification. Detection of genetic variation observed in amino acid sequences of hair proteins provides a new suite of identity markers that augment microscopic hair analysis and mitochondrial DNA sequencing. In this study, a new method that completely dissolves single hairs using a combination of heat, ultrasonication, and surfactants was developed. Dissolved proteins were digested and genetically variant peptide (GVP) profiles were obtained for single hairs (25 mm) via high-resolution nanoflow liquid chromatography-based mass spectrometry and a novel exome-driven bioinformatic approach. Overall, 6,519 unique peptides were identified and a total of 57 GVPs were confirmed. Random match probabilities ranged between 2.6×10^{-2} and 6.0×10^{-9} . The new bioinformatic strategy and ability to analyze GVPs in forensically relevant samples sizes demonstrates applicability of this approach to distinguish individuals in forensic contexts.

Use of human hair in forensics has contributed to investigations aimed at associating individuals with scenes and objects connected with a crime.(1) Current forensic techniques applied to characterization of shed hair that lacks a root include microscopic analysis and mitochondrial DNA (mtDNA) sequencing.(2-4) Studies intended to understand the reliability of microscopic analysis reveal underlying/fundamental issues in the approach.(4-6) For example in 2002, the FBI published a landmark report describing a systematic and comprehensive analysis of a large collection of previous casework to measure the frequency of false-positive associations in microscopic comparisons.(2) The importance of the 2002 FBI report was further echoed on page 121 of the 2016 report from the President’s Council of Advisors on Science and Technology: “Its conclusion is of enormous importance to forensic science, to police, to courts and to juries: When hair examiners conclude in casework that two hair samples are microscopically indistinguishable, the hairs often (1 in 9 times) come from different sources.” (7) The reliance of microscopic hair analysis on subjective systems of correlation (e.g. expert witness opinion) is emphasized in these reports as a major cause of the technology’s shortfalls.(8-10) mtDNA sequencing is an objective alternative that can provide statistic measures of human identity.(2, 3) The direct inheritance of mtDNA from the maternal genetic line (e.g. mother to son) limits the discriminatory power of this methodology.(11, 12) Additional forensic techniques that are more individualizing and possess higher powers of discrimination are needed to increase the amount of information gained during investigations using hair evidence.

Recent development of a new forensic technique that exploits genetically variant peptides (GVPs) found in human hair offers an alternative method that is statistically based and objective. (13) This approach probes genetic variation in humans by associating mutations found in hair proteins (single amino acid polymorphisms; SAPs) to their DNA counterparts

(missense single nucleotide polymorphisms; SNPs) in protein-encoding regions.(13, 14) The initial demonstration of this method characterized GVPs in 10 mg of human hair and implemented a standard shotgun proteomic approach.(13, 14) Specifically, Parker et al. performed identification of potential GVP identity markers through traditional proteomic analysis alone and subsequently confirmed results using Sanger sequencing.(13) This previous study established that GVPs found in hair protein can be a source of human identity markers.

To further develop this technology as a useful forensic technique, the present study aims to maximize the amount of proteomic information obtained from forensically relevant sample sizes (e.g. 1-inch of hair) in contrast to the Parker et al study that analyzed 10 mg of hair for each sample. Utilization of single hairs was achieved by developing a protocol that fully dissolved the hair matrix. This strategy improved the number of peptides detected in hair by increasing the solubility of proteins in the sample and resulted in more complete proteolysis by trypsin. In addition, an exome-driven approach was developed to detect genetically encoded GVPs in contrast to Parker et al study that applied a generic set of GVPs that were confirmed individually by Sanger sequencing. The exome-driven approach provided a genetic “blueprint” of all GVPs expected in each individual and ensured that potential GVP identifications were not be missed during data analysis.

Keratin and keratin-associated proteins make up the largest proportion of the hair shaft proteome, but numerous intracellular proteins have also been observed.(15, 16) Hair proteins are organized in a well-ordered hierarchical coiled-coil structure that is intricately designed for stability.(16-19) Due to extensive disulfide and isopeptide bonds, hair is an extremely difficult matrix to fully dissolve.(15, 20) Human hair samples of 1 mm length have been analyzed in a previous study aimed at identifying the major proteins observed in such small sample sizes.(21)

In the present study, a combination of surfactants, elevated temperatures, and ultra-sonication were employed to achieve complete dissolution of single 1-inch (25 mm) hair segments to identify potential GVPs. This length of hair was chosen specifically to resemble the average length of trace-hair evidence likely to be found at crime scenes.(22) The optimized procedure developed in this study was finely tuned to successfully dissolve hair proteins while avoiding hydrolysis of peptide bonds that hold proteins together.

Additionally, an exome-driven approach consisting of initial genetic sequencing (whole exome) of a subject's DNA collected separately using blood draws was used to later inform individualized SNP-associated protein databases, thus generating a database of sequences with the SAPs present in the individual. Protein databases created from exome sequenced were then used by the proteomic software to identify potential GVPs in MSMS data collected from specific hair protein extractions. This strategy consequently has a built-in genetic validation component because only DNA-encoded mutations known to be present are identified. By gathering all possible protein mutations observed in a person's DNA, the exome-driven approach greatly expands the opportunity to detect as many GVPs as possible. More GVPs detected within a sample provide additional identity marker that factor into higher degrees of statistical discrimination.

In combination, results from this two-tiered approach provided greater numbers of uniquely identified proteins and peptides than previously published. Additionally, a higher number of GVPs were identified. This article establishes the applicability of this technology to forensic practice by demonstrating successful analysis of sample sizes commonly collected as trace evidence. Additionally, the science of GVPs has been significantly advanced by filling the critical gap between large to small sample sizes, progressing the technology further down the path to implementation in real-world forensic scenarios.

Materials and Methods

Hair Sample Processing

Single-hair samples (1 inch; 25 mm) from each of three individuals were measured and placed into separate Protein LoBind Eppendorf tubes (Sigma-Aldrich, St. Louis, MO). In order to submerge the hair in a minimal liquid volume, each hair was cut into four equal segments before placement in the tube. A volume of 100 μ L of solubilization solution was added to each tube. Solubilization solution contained a buffer (0.05 M ammonium bicarbonate; ABC), reducing agent (0.1 M dithiothreitol; DTT), and surfactant (10 mg/mL sodium dodecanoate; SDD). Samples were then incubated at 70°C in an ultrasonic water bath (Elmasonic P; Elma; Singen, Germany) for 60 min. Throughout incubation an ultrasonic frequency of 37 kHz at a power setting of 100% was applied to maximize dissolving, mixing, and dispersing of the hair matrix. SDD was removed by extraction with an equal volume of acidified ethyl acetate (pH 2-3, 0.75% (v/v) trifluoroacetic acid) as follows. After addition of 100 μ L acidified ethyl acetate to each tube, samples were vortexed (5 s), incubated at 25 °C for 5 min, and centrifuged for 5 min (20,000 x g). Organic phase was removed, discarded to waste, and the extraction process was repeated once. The remaining aqueous phase was then neutralized to pH range 7-8 with 1 M ABC. (23, 24) Carbamidomethylation of reduced cysteines was performed by incubation for 60 min in the dark at 25 °C after addition of 6 μ L of iodoacetamide (1 M). To further solubilize proteins, 1 μ g/mL ProteaseMAX™ Surfactant (Promega; Madison, WI) was added to each sample. Solubilized protein fractions were then concentrated from ~100 to 50 μ L using 10 kD molecular weight spin filter concentrators (Millipore; Hayward, CA). Trypsin (1 μ L of 0.5 μ g/ μ L; Modified, Sequencing Grade, code = TRSEQZ, Worthington Biochemical Corp.; Lakewood, NJ) was added to each protein sample and digestion reaction proceeded at 25°C for 20-22 h with continuous agitation by magnetic-bar stirring. The resulting peptide mixture is then filtered by centrifugation using a spin filter (Durapore™ Membrane; 0.1 μ m pore size; EMD Millipore;

Hayward, CA), transferred into fresh vials, and subsequently analyzed by mass spectrometry. Processing of bulk samples was performed on the same number of sample replicates using hair from the same individuals. The procedure published in Parker et al. was followed directly as written in supplemental methods section (S1 Methods, Treatment I) using 10 mg of hair.(13)

Data Acquisition

Data acquisition was performed using Thermo Scientific Q Exactive Plus Orbitrap mass spectrometer coupled to an Easy-nLC 1000 liquid chromatograph (Thermo Scientific; Waltham, MA; USA). The system was equipped with a trap column (Acclaim™ PepMap™ 100, 75 µm x 20 mm, 3 µm C18 particles) that preceded an Easy-Spray™ column (ES801, 50 µm x 150 mm, 2 µm C18 particles) (Thermo Scientific; Waltham, MA; USA). After trap equilibration (12 µL mobile phase A; 0.1% formic acid in water) and column equilibration (5 µL mobile phase A), 2 µL of sample was loaded onto the trap column using 12 µL mobile phase A. Peptides were separated by reversed-phase liquid chromatography using a mobile phase A and a mobile phase B (0.1% formic acid in acetonitrile) in a 105-min gradient elution program (ramping from 5 to 20% B in 75 min, 20 to 35% B in 15 min, and holding at 90% B for 15 min) at a flow rate of 300 nL/min. Nano-electrospray ionization was achieved in positive mode with a voltage of 1.9 kV. Mass spectral data were obtained using a “top-10” data-dependent collection strategy in which an initial MS scan over the range of m/z 380-1800 and a resolution of 70,000 was used to select 10 precursor ions for subsequent MSMS scans. An isolation width of m/z 2 and a resolution of 17,500 was used for MSMS scans. Precursor ions were dynamically excluded from being selected again as precursor ions for MSMS for 24 s (time parameter chosen to achieve one to four spectra in chromatographic peak). Blank samples were used to assess and minimize potential carry over between samples.

Genetically Variant Peptide Prediction

Exome-driven GVP identification hinges on matched mutated protein sequences for each subject. Subjects provided a hair sample and a blood sample. DNA was isolated from blood corresponding to each hair sample to obtain genomic DNA that was not compromised during hair formation. Full exome sequencing data (10-0111_ACE Research Exome with Secondary Analysis; 8 Gb; Alignment, Variant Calling and Annotation; © 2018 Personalis Inc.; Menlo Park, CA) was obtained to detect the presence of SNPs in genes of interest. Specifically, exome sequencing revealed the presence of missense SNPs in the hair genes of each subject. These SNPs cause amino acid substitutions in the encoded protein amino acid sequence that can be detected in the form of GVPs. Variant call format (VCF) files provided by Personalis, Inc. were filtered to include only missense variants found in 627 genes commonly identified by hair proteomics. (13, 15, 24) This subset of target genes and SNPs were additionally filtered to those with a Genome Analysis Toolkit (GATK; © Broad Institute 2018) filter grade of PASS or better. Exome data was acquired with coordinates referenced to Genome Reference Consortium Human Build 37 (GRCh37.5 or hg19) and the Bioconductor package Variant Annotation (25) was used to convert to the more current GRCf38 coordinates in R (<https://www.ncbi.nlm.nih.gov/grc>). (26, 27) Resulting files were annotated using the web-based Ensembl Variant Effects Predictor (VEP; <https://uswest.ensembl.org/Tools/VEP>) (28) with identifiers including gene, transcript, location of genetic mutation, and corresponding amino acid substitution for each SNP. In cases when VEP did not return an identifier (e.g., SNP rsID) the mutation was identified using a shortened version of the corresponding Human Genome Variation Society (HGVS) (29) specification. SNP-associated transcripts were translated into protein amino acid sequences reflecting the amino acid substitutions by Ensembl Genome Browser (<http://www.ensembl.org>) (30) using the R package biomaRt (31). Resulting mutated sequences were combined with their non-mutated counterparts and the UniProtKB reference

human proteome (The UniProt Consortium; <https://www.uniprot.org>) (32) as FASTA formatted files, each having unique accessions in FASTA headers, and used as individualized protein databases for each subject.

Genetically Variant Peptide Identification

PEAKS Studio 7.5 (Bioinformatics Solutions Inc.; Waterloo, Ontario, Canada) protein identification software was used to search each RAW data file to determine the specific proteins and GVPs that were identified in each sample (see supplemental for RAW datafile downloads). Search settings included variable post-translational modifications including oxidation of methionine, deamidation of asparagine and glutamine, and carbamidomethylation of cysteine. Monoisotopic mass error tolerance of 30 ppm was used for precursor ion identifications and a 0.05 Da for the fragment ions masses. Individualized mutated databases were uploaded into PEAKS software to identify GVPs associated with SNPs observed in subject's exome. A decoy database generated within the software was used to determine the false discovery rate (FDR) of protein identifications. Protein and peptide identifications (IDs) were filtered by a 1% FDR. Protein IDs were then additionally filtered to those having two or more unique peptides detected. Remaining peptides with a reported mass error of less than ± 5 ppm from the mean (the mean mass error for each analysis case) were retained. Peptides having SNP-associated GVPs and their corresponding reference counterpart were retained if they were unique to a single human gene (unique genetic coordinates). Unique criteria was tested against the UniprotKB human set of transcripts and Ensembl Genome Browser (<http://www.ensembl.org>) (30) containing over 172,000 transcripts. GVP peptide hits were further annotated with information retrieved from the Genome Aggregation Database (gnomAD; <http://gnomad.broadinstitute.org/>; © Broad Institute 2018) (33), allowing for calculation of the population frequency.

Random-match Probability

Random-match probability (RMP) is calculated based on detection of variant GVPs. To compute RMP from detected hits, assumptions including: i) The reference peptide is not reliably detected, hence the population frequency for the mutation as computed from the gnomAD (<http://gnomad.broadinstitute.org/>; © Broad Institute 2018; (33)) is used, here population frequency is the fraction of mutation carriers within a population, as opposed to an allele frequency; ii) Negligible correlation of mutations between chromosomes; and, iii) Within a chromosome, a mutation with an allele frequency of less than 1×10^{-3} is independent, and there is negligible correlation between mutations separated by more than 2×10^5 base-pairs.

Results and Discussion

Mass spectrometric analysis of single-hair (SH) samples resulted in 578 unique protein identifications overall. There were 6,519 gene-specific peptides contributing to these identifications. Bulk-hair (BK) samples (see Materials and Methods) processed for comparison had 531 unique proteins identified with 3,887 gene-specific peptides total. Roughly two-thirds (314) of the protein identifications were shared between the two sample sets processed with different protocols. The SH set had 264 unique protein identifications and the BK set had 217 (Figure 1). On average, SH samples had 603 ± 195 ($\bar{x} \pm \sigma$) protein identifications and BK samples had 485 ± 123 . Unique peptide identifications averaged $3,570 \pm 874$ and $2,182 \pm 595$ for SH and BK samples, respectively (Table 1).

There were 28 keratins and 29 keratin-associated proteins detected in SH samples. Gene ontologies (GOs) were obtained using UniProt (The UniProt Consortium; <https://www.uniprot.org/>) database.(32) There were 513 molecular function GO annotations assigned to the protein IDs in SH-data sets and 465 in the BK. Comparing the two sets of protein IDs, the majority of GO categories had proportional differences of less than 2%. For

categories copper chaperone and translation regulator the same protein IDs were observed in the two sets. The molecular function designations that had differences greater than 2% between the datasets include SH samples having 7% more IDs with catalytic activity and BK samples having 5% more proteins designated to have structural molecule activity (Figure 2). In both sample sets, the GO annotation of binding (defined as the selective, non-covalent, often stoichiometric, interaction of a molecule with one or more specific sites on another molecule) was the most populated molecular function designation for the set of proteins.

A total of 57 genetically variant peptides (GVPs) were identified across all 12 samples (representing three subjects each processed in duplicate by SH and BK methods) using the described exome-driven bioinformatic approach to identify GVPs. Together, GVP peptides were observed in 40 unique protein-encoding genes, located on 11 different chromosomes. SH samples derived from the same three subjects as the BK samples resulted in an overall higher number of total GVP identifications, 114 and 71 respectively (Figure 3). There were 26 GVPs observed in SH samples that were not detected in bulk-hair samples and 5 GVPs only observed in the BK sample set. An overlap of 26 GVPs were observed in at least one sample from each sample sets (combined SH data vs. combined BK data). There were nine exome-predicted GVP markers detected in all expected instances (includes all replicates and subjects where associated SNP was detected in corresponding exome) consistently in both SH and BK samples, see Supplemental Table 1 for detected peptide sequences.

Numbers of GVPs detected between the SH and BK samples when comparing the two processing protocols between each subject-specific duplicate (Figure 4) had substantial differences, ranges of ± 16 and 30% of the highest minimum (15 for SH) and the highest

maximum (24 for SH). Results from combined subject-duplicates averaged from 15 to 24 for SH samples and 9 to 17 for BK samples. Standard deviation between GVPs observed in duplicate samples for the BK analysis was slightly lower (average 1) compared to SH standard deviation (average of 2).

Random match probabilities (RMPs) were calculated for each sample by combining each GVP-associated SNP population frequency with the product rule. The lowest RMP calculated was 6.0×10^{-9} for SH sample A, individual 3. RMPs calculated from SH samples had a broad range, spanning from 2.0×10^{-2} to 6.0×10^{-9} . Similarly, BK sample RMPs ranged from 1.4×10^{-1} to 2.0×10^{-7} . Variation in the RMPs calculated for each set of GVPs detected is a result of the differences the population frequencies of corresponding SNPs. Specifically, if two GVPs are identified and correspond to low population frequencies, the resulting RMP will be low (outcome of applying the product rule). Whereas if in a sample from the same subject two GVPs are detected but one having a higher population frequency, the resulting RMP will be higher and less discriminating. Moreover, additional GVPs will contribute additional factors to the product-rule calculation of RMP and lower its value, resulting in an increased level of discrimination.

Comparison of three individuals in biological duplicates for single-hair (SH) vs. bulk-hair (BK) protocols revealed that SH preparations resulted in 20% more unique proteins and 40% more unique peptides. The ability of SH sample preparation to fully dissolve hair likely released additional proteins into solution, ultimately increasing the number of proteins accessible to proteolysis by trypsin and subsequent identification.(34, 35) Due to the incomplete dissolution of the hair matrix during BK sample processing as indicated by the pellet left over after proteolysis, proteomic results likely represent the proteome found in the more readily dissolved hair shaft cortex (15, 36) whereas SH data represents a more comprehensive hair proteome. The number

of observed unique peptides and proteins directly impacts the ability to detect GVPs. Greater diversity observed in protein identifications (IDs) increases the probability for additional GVP observations. Increased unique peptides observed in SH data indicates an overall higher protein sequence coverage and subsequently higher chance of detecting regions containing genetic variation. Overall, comparable results were obtained from both SH and BK processed samples with other published research.(15, 35, 37-39) The two different processing techniques resulted in a substantial proportion of the same protein IDs as previous publications.(13-15)

Differences between the datasets were observed in the proportion of protein IDs designated with gene ontologies having structural-molecule activity (20% of SH- and 25% of BK proteins; 21% in Lee et al. 2006) and catalytic activity categories (44% of SH and 37% of BK proteins). The higher representation of the structural functional class of proteins (e.g. keratins and keratin-associated) in BK samples as compared to SH observed in the current study are in line with the established understanding that the soluble portion of the hair shaft (cortex) is largely made up of keratins.(15)

Average standard deviation between duplicates in number of unique protein IDs was slightly higher for SH sample data than for the BK sample data (\pm 15% and 6%, respectively; Table 1). This expected observation may suggest greater homogeneity in BK-sample sets. Protein ID differences between SH samples provide a closer look at biological variation of individual hairs from the same subject. Biological variation among individual hairs originating from the same individual may arise from different proteins being integrated into hair shafts. For example, the nonspecific enzyme transglutaminase is understood to play a key role in hair formation by creating isopeptide linkages between random proteins.(15, 35, 37) This may result in the integration of different proteins into individual hair shafts. In forensic contexts, GVPs would be chosen to represent a generalized panel that could be implemented to analyze unknown

samples. GVPs would be selected from proteins whose abundances do not significantly change with body location. For example, they would be chosen from the most abundant keratins and other hair proteins that are important for hair structure and consistently detected with high sequence coverage.

The observed number of GVPs between the two datasets were higher in SH data than in BK data. This observation demonstrated that GVP-marker analysis is applicable to sample sizes that are three orders of magnitude less than previously reported (10 mg to ~85 ug). (13) The difference in the standard deviation observed between BK and SH samples indicated that in practice, single-hair samples may provide different sets of GVPs depending on body location, hair length, or age. GVPs that are consistently observed from hair across all body locations would be targeted as candidates for a panel of markers that could be used in forensic applications.

Results revealed that higher numbers of detected GVPs does not necessarily result in lower random match probabilities. Specifically, comparing subject 2B and 3B BK results, there was an outcome of 10 GVPs for both, however that RMPs calculated were six orders of magnitude different (1.4×10^{-1} and 2.0×10^{-7}). In the case of Subject 2, there were 7 GVPs that were not detected in the B duplicate sample that were detected in the A duplicate. This may be a function of the data-dependent detection strategy used in this study. For discovery purposes, data-dependent acquisition is the ideal approach because it does not allow highly abundant peptides to mask the presence of lower abundant peptides. However, because the data acquisition software is choosing which spectra to collect throughout the run and is dependent on the ions detected at a particular time, ion detection may be variable between samples and even between technical replicates. The observed variation within the data highlights the need for further development of a generalized panel that includes a set of GVPs that can be searched for in all

samples using a targeted detection strategy. This panel should be specifically designed to include GVP markers that together, yield low the RMPs. Additionally, each GVP within the panel is strategically chosen to ensure observation of a subset in each sample because they have a high enough probability of being observed in random samples (high enough population frequency). Targeted selected-ion monitoring (t-SIM; also known as selected-reaction monitoring) MS methods can be applied to peptides of interest across individual samples to increase sensitivity.(40, 41) Implementation of this MS approach offers a potential avenue to decrease the variation of observed GVPs among SH samples. Operationally, the GVP panel could be used in combination with synthetic peptide standards and MS targeted methods as a kit that could be implemented in any forensic laboratory. tSIM-MS, or other mass spectrometric techniques are also necessary to further evaluate the expected variability of detected GVP panels from sample to sample. Inherent variation in the specific GVPs detected or not detected in each sample parallels occasions of incomplete STR profiles commonly observed in standard forensic genetic sequencing practice used to analyze genomic DNA found at a crime scene.

Exome-driven discovery is an efficient approach to developing a panel of GVP markers that are genetically valid. This exome-driven approach to GVP discovery is advantageous for initial surveys of proteomic results however, additional confirmation is required to establish false positive rate of SNP allele calls and GVP detection. Sanger sequencing serves as an alternative measurement to exome next generation sequencing and can be used for genetic validation. tSIM-MS methods can be applied to peptides of interest across individual samples to gauge the reliability of each GVP as an identity marker to this aim. Reliable GVPs that are observed consistently in samples in which they are expected become candidates for a final GVP panel. Synthetic peptides can be used to obtain additional confirmation on MS peptide identifications. These peptides standards provide various metrics such as retention time, MS feature

quantification, and MSMS fragmentation patterns that can be documented for comparison to experimental data. In operational practice, synthetic peptides can be isotopically labeled and spiked into experimental samples, providing an internal standard for GVP identifications.

This study established that protein-based forensics applied to hair evidence is a viable approach to human identification by demonstrating GVP markers can be detected in forensically relevant sample sizes of a single inch of hair. This approach used in combination with microscopic comparisons can enhance the information available to forensic scientists when analyzing hair evidence. Further research is necessary to establish if protein-based human identification can be implemented to discriminate unknown individuals, however the observation of GVPs in a sample found at a crime scene can exclude a suspect from a crime when they correspond to SNPs not present in their DNA. Statistical evaluation of GVP panels circumvent any ambiguity in panel profiles by providing the associated score (RMP) that is easily understood. Additionally, this technology introduces a new non-subjective method for hair analysis that can be successfully applied to forensically relevant sample sizes. Protein-based human identification has the potential to be a powerful tool used by forensic experts in the future.

References

1. Maxfield MG, Babbie ER. Research methods for criminal justice and criminology: Nelson Education, Boston, MA: Cengage, 2014;198-228.
2. Houck MM, Budowle B. Correlation of microscopic and mitochondrial DNA hair comparisons. *J J Forensic Sci.* 2002;47(5):1-4.
3. Melton T, Dimick G, Higgins B, Lindstrom L, Nelson K. Forensic mitochondrial DNA analysis of 691 casework hairs. *J Forensic Sci.* 2005;50(1):JFS2004230-8.
4. Robertson JR. Forensic examination of hair: CRC Press, Philadelphia, PA: Taylor & Francis, 2002; 120-128.
5. Wennig R. Potential problems with the interpretation of hair analysis results. *Forensic Sci Int.* 2000;107(1-3):5-12.
6. Smith CAS, Goodman PD. Forensic hair comparison analysis: nineteenth century science or twentieth century snake oil. *Columbia Human Rights Law Rev.* 1995;27:227.
7. President's Council of Advisors on Science and Technology. Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods. Washington, DC: Executive Office of the President. 2016 Sept.
8. Wendler C, Bridgeman B, Cline F, Millett C, Rock J, Bell N, et al. The Path Forward: The Future of Graduate Education in the United States. Princeton, NJ: Council of Graduate Schools and Educational Testing Service. 2010 Apr.

9. Murphy E. The new forensics: Criminal justice, false certainty, and the second generation of scientific evidence. 95 Calif Law Rev. 721, 2007;95:3.
10. Moriarty JC, Saks MJ. Forensic science: Grand goals, tragic flaws, and judicial gatekeeping. Judges J. 2005;44:16.
11. Melton T. Mitochondrial DNA: Profiling. In: Allan Jamieson and Scott Bader, Glasgow, UK, editors, Wiley Encyclopedia of Forensic Science. Sussex, UK, John Wiley & Sons Ltd, 2016;245-250.
12. Keastle FA, Kittles RA, Roth AL, Ungvarsky EJ. Database limitations on the evidentiary value of forensic mitochondrial DNA evidence. Am Crim Law Rev. 2006; 22:169-174.
13. Parker GJ, Leppert T, Anex DS, Hilmer JK, Matsunami N, Baird L, et al. Demonstration of protein-based human identification using the hair shaft proteome. PLoS One. 2016; 11(9):e0160653.
14. Wu P-W, Mason KE, Durbin-Johnson BP, Salemi M, Phinney BS, Rocke DM, et al. Proteomic analysis of hair shafts from monozygotic twins: Expression profiles and genetically variant peptides. PROTEOMICS. 2017; 17(13-14).
15. Lee YJ, Rice RH, Lee YM. Proteome analysis of human hair shaft: from protein identification to posttranslational modification. Mol Cell Proteomics. 2006; 5(5):789-800.
16. Wolfram LJ. Human hair: a unique physicochemical composite. J Am Acad Dermatol. 2003;48(6):S106-S14.

17. Yu Y, Yang W, Wang B, Meyers MA. Structure and mechanical behavior of human hair. *Mater Sci Eng C Mater Biol Appl.* 2017;73:152-63.
18. Harkey MR. Anatomy and physiology of hair. *Forensic Sci Int.* 1993;63(1-3):9-18.
19. Kreplak L, Doucet J, Briki F. Unraveling double stranded α -helical coiled coils: An x-ray diffraction study on hard α -keratin fibers. *Biopolymers.* 2001;58(5):526-33.
20. Greenberg CS, Birckbichler PJ, Rice RH. Transglutaminases: multifunctional cross-linking enzymes that stabilize tissues. *FASEB J.* 1991;5(15):3071-7.
21. Carlson TL, Moini M, Eckenrode BA, Allred BM, Donfack J. Protein extraction from human anagen head hairs 1-millimeter or less in total length. *Biotechniques.* 2018;64(4):170-6.
22. Garn SM. Types and distribution of the hair in man. *Ann N Y Acad Sci.* 1951;53(3):498-507.
23. Rice RH, Green H. The cornified envelope of terminally differentiated human epidermal keratinocytes consists of cross-linked protein. *Cell.* 1977;11(2):417-22.
24. Rice RH. Proteomic analysis of hair shaft and nail plate. *J Cosmet Sci.* 2011;62(2):229.
25. Obenchain V, Lawrence M, Carey V, Gogarten S, Shannon P, Morgan M. VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants. *Bioinformatics.* 2014;30(14):2076.

26. Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, et al. Modernizing reference genome assemblies. *PLoS Biol.* 2011;9(7):e1001091.
27. Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen H-C, Kitts PA, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* 2017; 27(5):849-864.
28. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The ensembl variant effect predictor. *Genome Biol.* 2016;17(1):122.
29. Den Dunnen JT, Antonarakis SE. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. *Hum Mutat.* 2000;15(1):7.
30. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, et al. Ensembl 2018. *Nucleic Acids Res.* 2017;46(D1):D754-D61.
31. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics.* 2005;21(16):3439-40.
32. UniProt C. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 2016;45(D1):D158-D69.
33. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature.* 2016; 536(7616):285-91.

34. Carrió MaM, Corchero JL, Villaverde A. Proteolytic digestion of bacterial inclusion body proteins during dynamic transition between soluble and insoluble forms. *Biochim Biophys Acta*. 1999;1434(1):170-6.
35. Laatsch CN, Durbin-Johnson BP, Rocke DM, Mukwana S, Newland AB, Flagler MJ, et al. Human hair shaft proteomic profiling: individual differences, site specificity and cuticle analysis. *PeerJ*. 2014;2:e506.
36. Yang F-C, Zhang Y, Rheinstädter MC. The structure of people's hair. *PeerJ*. 2014 2014/10/14;2:e619.
37. Thibaut S, Cavusoglu N, de Becker E, Zerbib F, Bednarczyk A, Schaeffer C, et al. Transglutaminase-3 Enzyme: A Putative Actor in Human Hair Shaft Scaffolding? *J Invest Dermatol*. 2009;129(2):449-59.
38. Hashimoto K. The structure of human hair. *Clin Dermatol*. 1988;6(4):7-21.
39. Barthélemy NR, Bednarczyk A, Schaeffer-Reiss C, Jullien D, Van Dorsselaer A, Cavusoglu N. Proteomic tools for the investigation of human hair structural proteins and evidence of weakness sites on hair keratin coil segments. *Anal Biochem*. 2012;421(1):43-55.
40. Lange V, Picotti P, Domon B, Aebersold R. Selected reaction monitoring for quantitative proteomics: a tutorial. *Mol Syst Biol*. 2008;4(1):222.
41. Shi T, Su D, Liu T, Tang K, Camp DG, Qian WJ, et al. Advancing the sensitivity of selected reaction monitoring-based targeted quantitative proteomics. *Proteomics*. 2012;12(8):1074-92.

Tables

Table 1. Identified Proteins and Peptides

Subject ID	Sample ID	Proteins	Peptides
<i>Single-Hair</i>			
1	A	698	4126
	B	622	3976
2	A	622	3754
	B	900	4571
3	A	401	2505
	B	377	2485
<i>Bulk-Hair</i>			
1	A	531	2802
	B	598	2713
2	A	544	2540
	B	577	2059
3	A	321	1480
	B	338	1496

*Number of unique proteins and peptides identified in each sample duplicate. Identifications are based on PEAKS software searches.

Figure Legends

FIG 1. Venn diagram of shared and unique protein identifications in single-hair (left) and bulk-hair (right) samples., **FIG 2.** Gene ontology (GO) percentages of identified proteins for each group of samples processed with different protocols. Asterix indicates difference of >5% between the two methods. Uniprot (www.uniprot.org) was the source of GO annotations., **FIG 3.** Genetically variant peptide panel. GVP markers detected in biological duplicates (indicated by A and B) for single-hair (SH) and bulk (BK) processed samples originating from three individuals (1-3) are indicated by black squares when present, and white when not detected. Associated SNP accessions (rsIDs), gene name (Gene), chromosome number (Chr#), genetic location (Location), and gnomAD population frequency (PF) are labeled. The rsIDs that were consistently detected in all expected samples are indicated by bold text and those consistent across single-hair samples specifically are indicated by an asterisk., **FIG 4.** Calculated random match probability (RMP) as a function of the number of detected genetically variant peptides (GVPs) in each subject panel. Each data point is labeled by subject number (1-3) and sample duplicate identifier (A or B). Bulk-hair data is designated with black circles and single-hair data is grey. RMPs were calculated using the product rule, with details found in Materials and Methods section.