Forensic Science International: Genetics 42 (2019) 21–30

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

Proteomic genotyping of fingermark donors with genetically variant peptides

Trevor Borja[a,1], Noreen Karim[a,b,1], Zachary Goecker[a], Michelle Salemi[c], Brett Phinney[c], Muhammad Naeem[b], Robert Rice[a], Glendon Parker[a,*]

[a] Department of Environmental Toxicology, University of California - Davis, Davis, CA, United States
[b] Department of Biotechnology, Quaid-i-Azam University Islamabad, Pakistan
[c] Proteomic Core Facility, Genome Center, University of California - Davis, Davis, CA, United States

ARTICLE INFO

Keywords:
Fingermarks
Genetically variant peptide
Non-synonymous SNPs
Epidermal corneocytes
Proteomics
Mass spectrometry
Proteomic genotyping
Genetically variable peptide

ABSTRACT

Proteomic genotyping detects single amino acid polymorphisms to infer the genotype of corresponding non-synonymous SNPs. Like any DNA genotype, these inferences can be used to estimate random match probability. Fingermarks are a common source of biological evidence that is sample limited and a highly variable source of identifying DNA. Genetically variant peptides from fingermarks, that contain single amino acid polymorphisms, are an additional source of identifying genetic information. To discover these peptide biomarkers epidermal corneocytes from 9 subjects were isolated, processed, digested with trypsin and applied to mass spectrometry. The resulting proteomic and matching exome datasets were used to discover, characterize and validate 60 genetically variant peptides. An average of 28.8 ± 4.4 genetically variant peptides were detected from each subject resulting in a total of 264 SNP allele inferences with 260 true and 4 false positives, a false discovery rate of 1.5%. Random match probabilities were estimated using the genotype frequencies from the matching major populations in the 1000 Genomes Project. Estimates ranged up to a value of 1 in 1.7×10^8, with a median probability of 1 in 2.4×10^6. Furthermore, the proteomically-inferred genotypes are likely to be compatible with the STR-based random match probability estimates since the closest STR locus was 2.2 Mb from the nearest GVP-inferred SNP. This project represents a novel mode of genetic information that can be obtained from fingermarks and has the potential to complement other methods of human identification including analysis of ridge patterns or touch DNA.

1. Introduction

Fingermarks are a source of identifying information that can provide a unique link between an individual and probative items or location [1–8]. Fingermarks are created when material is transferred from skin to a solid surface. While the physical amount of material is highly sample limited, the transfer often contains forensically useful information, either in the form of a two-dimensional friction ridge pattern, or informative molecules such as small compounds or genetic material [1,4,8–11]. The transfer is highly variable. The amount of material transferred is dependent on a donor's biology, such as their shedder status and level of sweat production, or behavior, such as the frequency and duration of transferring DNA- and protein-rich cells from other body regions or tissues, and the amount of time since the donor last removed biological material from the skin surface [8,10,11]. Contextual factors such as contact time, pressure, shear forces, and the physical nature of the receiving surface also affect the level and pattern of material deposition [8,10]. Fingermarks have a high surface area and are more likely to be degraded [12]. In this framework of sample limitation, variability, and degradation, the development of additional sources of genetic information may significantly enhance the reliability and scope of identifying fingermark information.

Recently protein has been demonstrated to be a carrier of genetic information in the form of single amino acid polymorphisms (SAPs), the result of non-synonymous SNPs [13]. Detection of these SAPs in genetically variant peptides (GVPs) allows for the inference of

Abbreviations: GVPs, genetically variant peptides; SNP, single nucleotide polymorphism; RMP, random match probability

* Corresponding author at: Glendon Parker, Department of Environmental Toxicology, University of California – Davis, One Shields Ave, Davis, CA, 95616, United States.
E-mail address: gjparker@ucdavis.edu (G. Parker).

[1] These authors contributed equally towards the manuscript.

https://doi.org/10.1016/j.fsigen.2019.05.005
Received 11 March 2019; Received in revised form 9 May 2019; Accepted 26 May 2019
Available online 28 May 2019
1872-4973/ © 2019 Elsevier B.V. All rights reserved.

corresponding SNP alleles in the donating individual. These are typically common DNA variants, well documented in extant genetic databases and present in every population. In aggregate a profile of proteomically-inferred SNP alleles can be used, just like any SNP or STR-based genotype, to estimate random match probabilities (RMPs) [13–15]. This approach has been demonstrated in hair shafts and more recently in bone protein [13,16]. Protein is a major component of fingermarks [1,8,12,17–22]. Therefore, it is a potential source of genetic information that can be used to infer SNP genotype [13]. Because protein is chemically more stable than DNA, it is possible that proteomic genotyping may be a major source of SNP genotype information in degraded samples [23–25]. Since the inferred SNPs are autosomal they can potentially be incorporated into SNP or STR-based RMPs.

This project sought to discover, characterize and validate genetically variant peptides that could be expected to occur in the fingermark proteome, specifically the proteins in epidermal corneocytes [26–28]. Once this was achieved we then sought to place these data into a broader context. This included estimation of RMPs and development of likelihood values for different population sources, the incorporation of rare GVPs, and the potential of a single RMP estimate using both GVP-inferred SNP alleles and STR genotypes. The validated peptide biomarkers that were identified, characterized and validated in this study provide a starting point for additional studies that extract genetic information from the protein fraction of fingermarks.

2. Methods

2.1. Sample procurement

The study was approved by the Institutional Review Board of the University of California, Davis (IRB#217868-14) and Quaid-i-Azam University, Islamabad (IRB #216) prior to the study and informed consent was obtained from all participating individuals. Venous blood and mouthwash buccal cells (for genomic DNA) as well as epidermal samples (for proteomic analysis) were collected using IRB compliant protocols from nine unrelated adult (> 18 years old) individuals of European (n = 5) and South Asian (n = 4) origin (Table S1). Samples were collected as described previously [29]. Briefly, epidermal samples were collected using D110-D-squam stripping tapes (CuDerm, Dallas, USA) that were applied onto the skin area and pressed with a gloved thumb. The tapes were removed with the help of forceps and placed in a clean 15 mL polypropylene centrifuge tube such that the adhesive sides were facing inward and were not over-lapping. Each of the enrolled subjects was sampled at up to three different anatomic sites as described (Table S1). Five tapes were collected from each site. Samples that were used to study dithioerythritol-soluble (A_S) and insoluble (A_P) proteins were taken from the forearm (A, Table S1, Method 2). Subjects UCD006-008 were ichthyosis patients [30].

2.2. Sample processing

As described previously epidermal corneocytes were eluted from tape circles by soaking them overnight in 2% sodium dodecyl sulfate – 0.1 M sodium phosphate solution (SDS-$NaHPO_4$) (pH 7.8) [29,30]. The cell bodies settled at the bottom of the tubes, transferred to clean microfuge tubes by the help of Pasteur pipettes and were washed twice with 2% sodium dodecanoate (SD) – 0.05 M NH_4HCO_3 (SD - ABC), each wash being followed by centrifugation and discarding of the supernatants. The pellets were resuspended in 0.4 mL of the SD – ABC solution followed by the addition of 20 µL of 1 M dithioerythritol to make 50 mM and stirred at room temperature for 20 min. To enhance the reduction, the samples were incubated at 95 °C for 10 min. and then transferred to an incubator preset at 37 °C for one hour. Afterwards, iodoacetamide (to a concentration of 100 mM) was added and the samples were gently stirred for 45 min in the dark. The pH was adjusted to about 3 with trifluoroacetic acid and SD was extracted in the supernatants after the addition of 700 µL ethyl acetate, mixing, followed by centrifugation (16,100 *g* x 3 min) and removal of the upper organic phase. The pH was adjusted to about 8 with 2.5 µL of concentrated ammonium hydroxide and 20 µL of 1 M NH_4HCO_3. Reductively methylated trypsin was added for protein digestion [31]. For samples from UCD001 – 003, 87 µg of trypsin was added to the samples and incubated on a magnetic stirrer at room temperature for 6 h, at 3 h another addition of trypsin (87 µg) was added. For samples UCD004 – 009 and UCD002 – 003 A_S and A_P fractions, 25 µg of trypsin was added and the sample incubated for 3 days at room temperature with further 25 µg aliquots of trypsin added each day. The samples were then centrifuged and the supernatants, containing the digested peptides, were transferred to Lo-Bind microfuge tubes and kept frozen until analysis [30]. A subset of arm skin samples were collected and processed in 50 mM dithioerythritol as described above; however at the end of the reduction step the tubes were centrifuged, and the supernatant collected. The pellets were re-suspended and recentrifuged in 50 mM dithioerythritol in SDS-$NaHPO_4$ and the process repeated 4 times with the first two supernatants being pooled to comprise a supernatant (A_S), or DTE-soluble fraction [32]. The final pellet (A_P), or DTE-insoluble fraction, was resuspended using SDS-$NaHPO_4$ and both fractions processed identically to other samples by alkylation and subsequent treatments [32]. All resulting peptide samples were assayed for peptide concentration using a fluorescent peptide assay (Pierce, Thermo Fisher Scientific Inc.).

2.3. Mass spectrometry

A Thermo Fisher Scientific Q Exactive Plus Orbitrap mass spectrometer with inbuilt Proxeon nanospray and Proxeon Easy-nLC II HPLC was used for the mass spectrometric analysis. The digested samples (750 ng) were loaded on a 100 µm × 25 mm Magic C18 100 Å 5 U reverse phase trap, desalted online and separated over 120-min gradient of 5 to 80% acetonitrile in 0.05% formic acid via 75 µm × 150 mm Magic C18 200 Å 3 U reverse phase column at 300 nL/min flow rate. MS survey was conducted at the *m/z* range of 350–1600, and the 15 most abundant ions from the spectra were selected and subjected to higher-energy C-trap dissociation to fragment the precursor peptides and obtain MS/MS spectra [33]. Precursor ions selected in a 1.6 *m/z* isolation mass window and fragmented via 27% normalized collision energy. A 15 s duration was used for dynamic exclusion.

2.4. Proteomics analysis using peptide spectra matching software

Raw data files (.RAW) were converted to MzML format using the MSConvertGUI (Proteowizard 2.1, http://proteowizard.sourceforge.net) and applied to the desktop X!Tandem peptide spectra matching algorithm (Global Proteome Machine Fury, X!Tandem Alanine (2016.10.15.2)). Default search parameters were used except that no prokaryote and virus reference libraries were used and the point mutation function was activated. Peptide and protein log(e) scores were set to -1, and fragment mass errors of 20 ppm, and parent mass error of 100 ppm were used. The mass spectrometry proteomic datasets have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD012628 and MassIVE proteomics data repository at the UCSD Center of Computational Mass Spectrometry (MSV00008342) [34].

2.5. Label free quantitation and analysis of protein expression levels

To quantify the proteins from the LC–MS/MS datasets, a label-free quantification analysis was performed using PEAKS™ Studio 10.0 (Bioinformatics Solutions Inc., Waterloo, ON, Canada) [35]. A single representative proteome dataset from non-ichthyosis subjects (1 to 5, 9) from 3 skin body locations (P = palmar, A = arm, FH = forehead) for a total of 13 datasets was processed using a PEAKS™ validated UNIPROT

human reference proteome (UP000005640, www.uniprot.org/). Default settings were used with the exception of a 15 ppm mass error range for the precursor mass and 0.04 Da (Da) for fragmentation masses. A fixed amino acid carbamidomethylation (+57 Da) was assumed for cysteines and variable modifications assumed for methionines (oxidation, +15.99 Da; dioxidation, +29.99 Da), glutamines and asparagines (deamidation, +0.98 Da), tryptophan and histidine (oxidation, +15.99 Da) and N-terminal (pyroglutamation, -17.03 and -18.01 Da; formylation, +27.99). The resulting datasets were filtered with a 1% peptide false discovery rate and analyzed in the Q-Module to generate a heat map by label free quantitation using default settings.

2.6. Genetically variant peptide discovery and analysis

The peptides identified in the Global Proteome Machine to carry single amino acid variants (SAV) were selected based on the following inclusion criteria: a matching (log(e)) score of < -2, a matching dbSNP accession number with a minor allelic frequency greater than 0.5% in the 1000 Genomes Project (ensembl.org) [36,37]. Exclusion criteria included: the presence of unexpected chemical or genetic modifications (i.e., other than methionine oxidation, deamidation, N-terminal acetylation, and cysteine carboxymethylation), the presence of fragmentation masses consistent with the alternative reference allele and non-specific cleavage of the peptide backbone. Polymorphisms that resulted in the same masses (such as I to L, or Q to K) or in mass shifts that were identical to common modifications such as deamidation (N/Q to D/E), or oxidation (M to F) were also excluded. To prevent the inclusion of peptide with more than one genomic address, all peptide sequences were submitted to PROWL (prowl.rockefeller.edu/prowl/proteininfo) and searched against the IPI human (2010-02-01) database. Peptides with no match or represented by a single point in a gene were considered unique and included in the study. Peptides identified in this manner that met all proteomic criteria were defined as candidate GVPs and peptides that are demonstrated to accurately infer DNA genotype were defined as validated GVPs in this study.

An alternative discovery process was also incorporated into this study. Exome genotypes of South Asian subjects were examined for the presence of non-synonymous SNPs in gene products that were detected in the epidermal corneocyte proteome. Proteomic data were then analyzed to confirm if there were GVPs that corresponded to either allele of the SNP locus. No filtering for genotype frequency was applied in this approach. Peptides confirmed in this manner were defined as confirmed GVPs in this study.

2.7. Exome sequencing and bioinformatic analysis

Two exome sequencing and analysis protocols were used in this study. Samples UCD001 to UCD005 were processed at the DNA Technologies Core Facility at University of California - Davis. Barcode-indexed sequencing libraries were generated from genomic DNA samples (1000 ng) sheared on an E220 Focused Ultrasonicator (Covaris, Woburn, MA). The sonicated DNA was size selected with KAPA Pure beads to obtain fragments of about 300bp. Size selected DNA (30 ng) were used for library preparations with the KAPA Hyper DNA library kit, according to the manufacturer's instructions. Ten cycles of PCR were conducted to amplify the libraries. Each library (500 ng) was pooled for exome capture using the IDT xGen® hybridization capture protocol according to the manufacturer's instructions. Seven cycles of PCR were conducted to amplify the library that was analyzed with a Bioanalyzer 2100 instrument (Agilent, Santa Clara, CA), quantified by fluorometry on a Qubit instrument (LifeTechnologies, Carlsbad, CA), and combined in two pools at equimolar ratios. The pools were quantified by qPCR with a Kapa Library Quant kit (Kapa Biosystems-Roche) and each pool was sequenced on one lane of an Illumina Nova Seq (Illumina, San Diego, CA) with paired-end 150 bp reads. Raw Illumina paired-end 151 bp reads were first subjected to quality control. Adapters were removed from the sequencing reads using scythe (https://github.com/vsbuffalo/scythe, version 0.994 beta). Base quality was controlled using a window-based method, sickle (https://github.com/najoshi/sickle, version 1.33), with the cutoff set at 30. Reads less than 30 bp in length were discarded. Reads that passed the quality control were mapped to hg19 reference genome using parameter -M for downstream analysis compatibility [38]. PCR duplicates were removed using Picard-tools (http://broadinstitute.github.io/picard/, version 2.18.4). Variants were identified using HaplotypeCaller function in GATK (version 4.0.5.2), followed by variant recalibration using the recommendations from GATK developers [39].

Samples UCD006 to UCD009 were processed at Macrogen Inc. (Seoul, South Korea). Genomic DNA was extracted employing the standard protocol for Wizard genomic DNA purification kit (Promega, USA) and was subjected to HiSeq4000 sequencing systems (Illumina, CA) for whole exome sequencing. The exonic regions were captured and enriched using SureSelect V5-post capture kit (Agilent Technologies, SC, CA) and loaded on an Illumina HiSeq4000 sequencer that uses bi-directional bridge sequencing approach and yields reads with an average size of 101 bp. The reads were mapped to reference sequence (GrCh37) by Burrows-Wheeler Aligner (bwa-0.7.10) and the duplicate molecules were marked and removed by Picard (picard-tools-1.118). Variant (indels and SNPs) calling and annotation were performed using GATK (GATK3.v4) and SnpEff (SnpEff_v4.1).

2.8. Validation and performance of genetically variant peptides

Inferred SNP alleles from GVPs were compared to corresponding genotyping data from sequenced exomes for each individual. Both positive and negative inferences were demarcated into true and false positives, and true and false negatives. From these false discovery rates (FDR = FP/(FP + TP)), positive predictive values (PPV = TP/(TP + FP)), sensitivity (sens = TP/(TP + FN)), and accuracy (accuracy = (TP + TN)/(TP + FP + TN + FN) were calculated [40].

2.9. Estimation of random match probabilities

The cumulative profile of matching inferred SNP alleles was used to estimate the random match probability (RMP) for that individual [13]. The RMP was calculated using a frequentist estimation of allele distribution in a reference population [15]. When more then one allele occurred in an open reading frame the distribution of the allele combination within the open reading frame was counted in the population, treating each gene as a single locus (Pr(imputed nsSNP allele gene combination|population)). These were then incorporated in the product-rule assuming full independence between open reading frames [13,15,41]. The occurrence of allele, or allele combinations, was counted in all of the major population groups in the 1000 Genomes Project, (European, African, South Asian, East Asian and Indigenous American; www.1000genomes.org; Phase 3) [37]. The final probability of an individual SNP, or SNP combination, occurring within a gene reading frame was estimated as (x + ½)/(n + 1), where x is the number of individuals with a given SNP, or combination of SNPs, in the sample population [42]. The above expression represents the Bayesian posterior mean of a binomial probability using the Jeffreys Beta (½, ½) prior, which has the advantage of giving a non-zero estimate of the population probability even for x = 0 [42,43]. Full independence between loci from different genes was assumed [13]. Two of the inferred SNP loci (rs7308811, rs7212938), while common (MAF > 0.5%), were not represented in the 1000 Genomes Project, but the alleles were observed in the Genome Aggregation Databases of exomes (gnomADe) [44]. In the event that corresponding GVPs from these SNPs were detected in the proteomic dataset the estimated genotype frequency ($gf_p = p^2 + 2pq$) from this reference population was substituted. If other loci from the same reading frames were inferred, then those values were used instead. In this scenario no mixing of alleles was used, since the

observation of allele combinations could not be counted. All ranges are reported as standard deviations.

3. Results

3.1. The epidermal corneocyte proteome

Epidermal corneocytes were sampled from different body sites using dermal patches and isolated by the sloughing of squamous cells and subsequent washing to remove detergent-soluble biological material [29]. The resulting cellular preparations consisted primarily of the solid cornified cores of terminally differentiated epidermal corneocytes [28,29,45]. These remnants were processed by reduction in 50 mM dithioerythritol (DTE) prior to alkylation and trypsinization. A subset of forearm (A) samples were treated as DTE soluble (A_S) and insoluble (A_P) fractions (n = 4, Table S1). A total of 40 acquisition and two reagent blank runs were conducted on samples from 9 individuals. An average of 530 ± 440 µg of peptide was obtained across all samples, well above the amount required (750 ng) for each run. Overall this resulted in identification of an average of 480 ± 180 proteins, and 22,600 ± 6,400 total peptides, corresponding to 2,800 ± 1,000 different sequences, or 2,300 ± 270 per subject (average ± standard error of the mean, Fig. 1A) and 3,000 ± 280 per body site (average ± standard error of the mean, Fig. 1B).

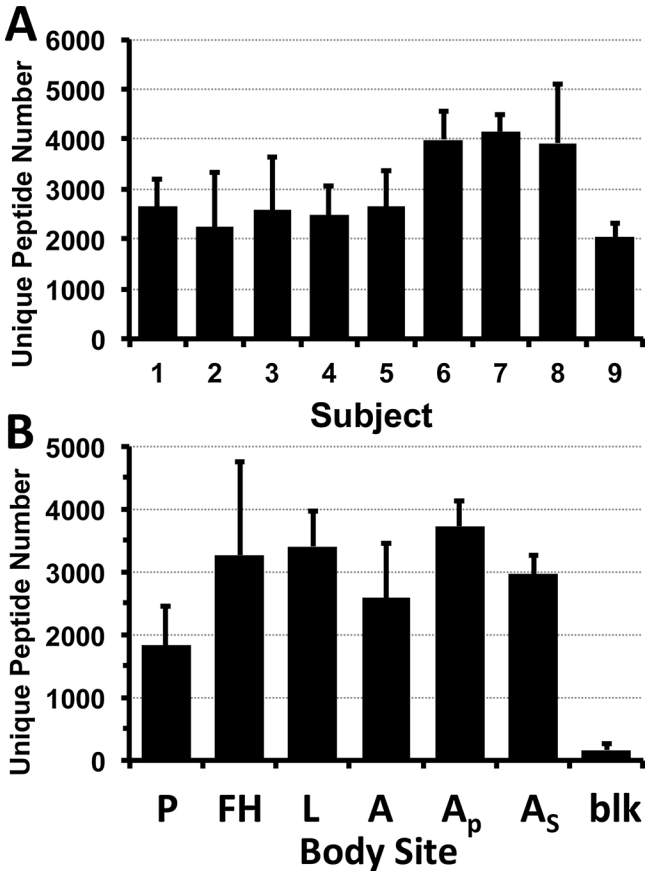

Fig. 1. The Epidermal Corneocyte Proteome.
Epidermal corneocyte preparations were obtained from each subject (UCD001 to 009). After obtaining and analyzing proteomic datasets using X!Tandem, the number of unique peptide sequences identified was collated for each subject (A), or from each body site location B). Mature corneocytes preparations were obtained from the palmar (P), forehead (FH), leg (L), and arm (A). A subset of arm preparations was further treated to isolate the protein fraction that was insoluble (A_p) or soluble (A_s) in the presence of 50 mM dithioerythritol. The indicated range is the standard deviation.

Differences in the protein profiles from different body locations were also observed. Using the Q-Module from PEAKs™, a single replicate dataset of each sample was placed into 3 groups corresponding to body location, quantified using label free quantitation and clustered to generate a heat map based on correlated protein abundance patterns (Figure S1). The ratio of increased quantification relative to the average dataset, in this case the forehead samples, was calculated based on the ion signal from the 3 most sensitive peptides. Based on this analysis a small subset of gene products was relatively over represented in palmar skin and under represented in other skin types. These high abundance proteins were keratin type I cytoskeletal 9 (Uniprot# P35527, ratio = 33.71) and hornerin (Uniprot # Q86Y23, ratio = 21.59). Variation was also observed in mature corneocyte preparations from other body sites. These were more complex and variable than palmer skin although there are distinctive abundance patterns specific for arm skin. These were consistent with abundance patterns observed previously [29,45]. The three ichthyosis subjects (UCD006 – 008) were removed from the analysis [30].

3.2. Discovery of candidate genetically variant peptides

Raw datasets were reformatted and applied to X!Tandem using the Global Proteome Machine [36]. Identified single amino acid variants (SAVs) were screened and variant peptides with matching common (> 1% genotype frequency) non-synonymous SNP alleles were identified and filtered based on the inclusionary and exclusionary characteristics described in the methods. Discovery of genetically variant peptides using analysis of exome genotypes was also employed on the South Asian subject cohort. The proteomes of these subjects were analyzed for gene products in the corneocyte preparations after non-synonymous SNP alleles in corresponding genes were identified in exome data that might be represented in the proteomic datasets. The matching proteomic datasets were then analyzed to confirm the presence of the candidate GVPs. Non-synonymous SNP loci (rs#) were identified that were expressed in the genes A2ML1 (rs7308811), SPRR1B (rs3795382), KRT13 (rs149773722), EPPK1 (rs6558399) and DSP (rs149773722), which is described in more detail below. In total, using both discovery approaches, a total of 74 candidate GVPs were identified, corresponding to 37 non-synonymous SNP loci from 31 genes. Of these, one is not unique (rs2239710_G, KRT34 280 T) and 13 were not observed in the proteomic data, leaving a total of 60 confirmed or validated biomarker peptides. GVPs are listed, along with chemical and genetic characteristics (Table S2). A catalog of potential epidermal corneocyte GVPs was generated based on a consensus proteome. A skin proteome compiled from 12 subjects was analyzed using PEAKs™ software version 10.0. Percent coverage was averaged among these samples. Proteins were excluded from the proteome if they had a percent coverage lower than 1%. The resulting proteins were used to search for missense variants using the Ensembl Biomart tool (ensembl.org/biomart/martview/). Variants with a global minor allele frequency lower than 0.005 were excluded. No other filtering was done on these putative GVPs (Table S3). A total of 1753 non-synonymous SNPs from 425 genes were identified.

3.3. Validation of epidermal corneocyte genetically variant peptides

Proteomic datasets from epidermal corneocyte preparations were screened for the presence of candidate GVPs (Fig. 2, PEPTIDE) and the resulting cumulative profiles used to infer the status of matching non-synonymous SNP alleles (Fig. 2, GN = Gene Name, rs# = dbSNP accession number-nucleotide allele). These predictions were cumulated for each subject (SUBJECTS, 1 to 9) and compared to the genotypes derived from matching exome datasets (Fig. 2, Table S4). The performance of proteomic genotyping for each candidate GVP was evaluated (true positive, blue; false positive, red; true negative, white; false negative, green). Of the 264 inferences, 260 were true positives and 4

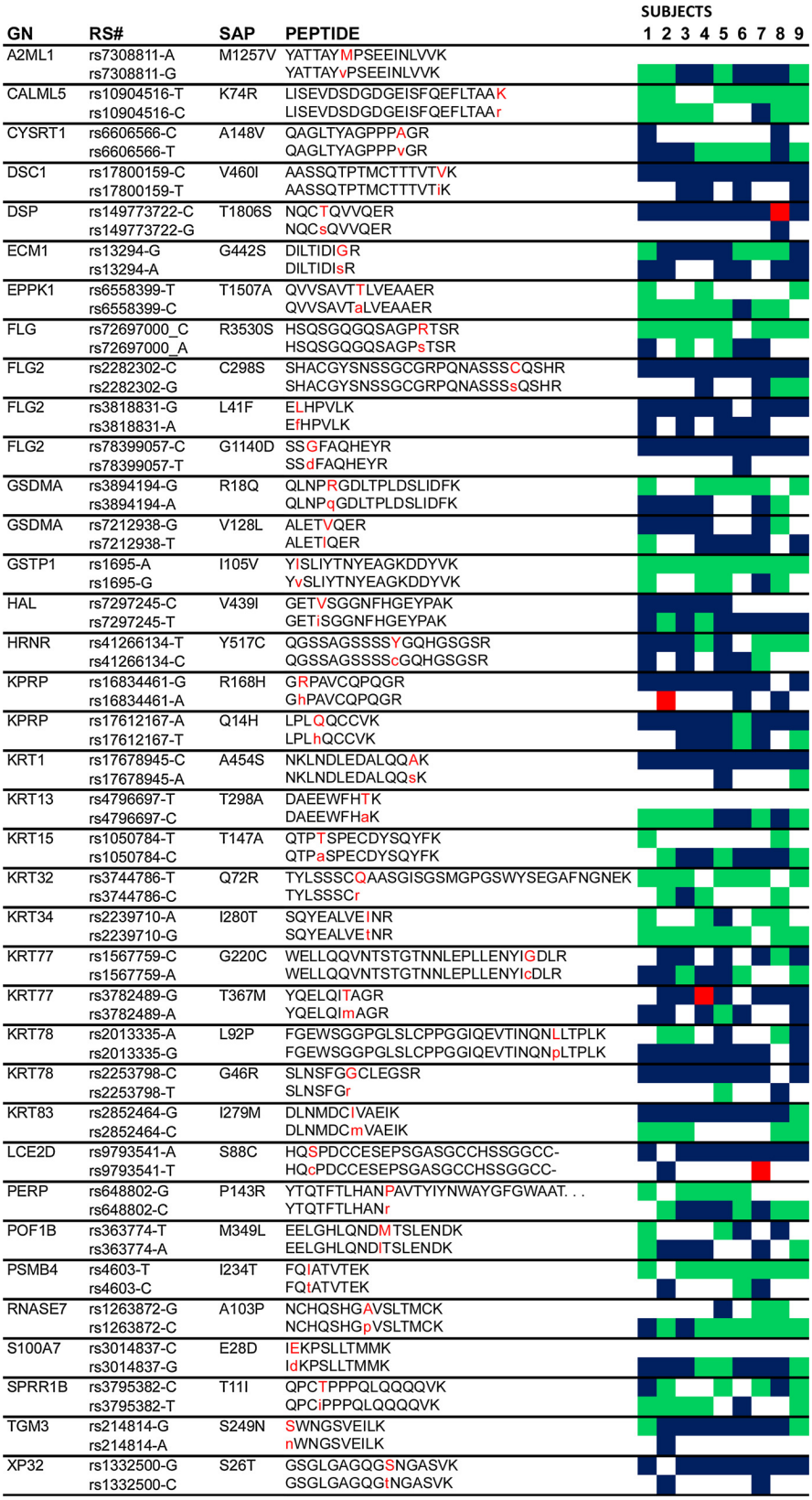

Fig. 2. Validation of Genetically Variant Peptides.
Genetically variant peptides (PEPTIDE) were detected in each dataset and cumulatively collated for each subject (SUBJECTS, 1 to 9). Inferred SNP allele genotypes (GN, gene name; RS#, dbSNP accession number - nucleotide allele) were compared to genotypes derived from matching exomic datasets. True and false positive predictions are indicated by blue and red squares, and true and false negative predictions indicated by white and green squares respectively. Single amino acid polymorphisms (SAP) are indicated in red with the variant non-reference allele being in lower case.

were false positives. There were 173 false negative and 229 true negative inferences. This results in an overall performance of a 1.5% false discovery rate (FDR = FP/(FP + TP)), 98.5% positive predictive value (PPV = TP/(TP + FP)), a sensitivity of 60.0% (TP/(TP + FN)), and accuracy of 73.4% ((TP + TN)/(TP + FP + TN + FN)). At the level of individual GVPs, 19 peptide-based inferences were 100% accurate. At the level of individual datasets (n = 40) a total of 725 positive predictions were made with 719 true and 6 false positives, for a false discovery rate of 0.8 ± 0.2%, overall sensitivity of 37.2 ± 11.6% and accuracy of 58.8 ± 7.8% (Figure S2). No GVPs appeared in the reagent blank datasets.

The variation in protein abundance patterns for different skin types has the potential to affect proteomic genotyping when using palmar compared to other skin types. We plotted the number of GVPs detected as a function of unique peptide number detected in each individual dataset (Figure S3, n = 40). Using linear regression (y = ax + b), and setting the y-intercept at b = 0, the yields of detected GVPs correlate significantly with the number of unique peptide sequences detected (y = 0.0061x, p = 0.02), a co-efficient of one GVP per 160 unique peptide sequences (www.graphpad.com/quickcalcs). The proteomic and GVP yields from palmar skin, 13 ± 4.1 GVPs, are 31% lower than the other skin types combined, 18.9 ± 5.4 GVPs (p = 0.02).

3.4. Statistical estimation of random match probabilities

The utility of corneocyte GVPs to provide forensically usable information depends on whether random match probabilities (RMPs) are sufficiently discriminating. As indicated by the dashed black line in Fig. 3A, the number of detected GVPs in the cumulative pooled datasets (Fig. 2) ranged from 21 to 34 with an average of 28.8 ± 4.4 across all subjects. The RMP was calculated for each individual (Fig. 3B). When using genotype frequencies from the 1000 Genomes Project or Genome Aggregation Database population consistent with the genetic background of each subject (solid black lines), estimated RMPs ranged from 1 in 2.8×10^4 to 1 in 1.7×10^8, with a median value of 1 in 2.4×10^6. European samples averaged a RMP of 1 in 3.7×10^7 and South Asian samples averaged 1 in 1.8×10^6. If RMPs were calculated using only loci represented in the 1000 Genomes Project, the detected number of GVPs was reduced to an average of 25.7 ± 4.1 and average probabilities were reduced to an average of 1 in 2.8×10^7 and 1 in 1.4×10^6 in the European and South Asian subjects, respectively.

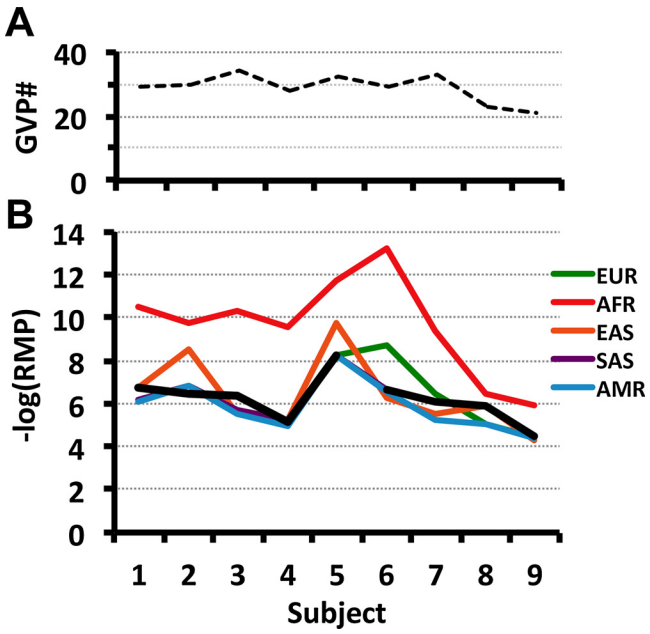

Fig. 3. Statistical Analysis of Proteomically-Inferred Genotypes.
A) Genetically variant peptides were detected and counted cumulatively for each subject (GVP#). B) Random match probabilities of each subjects profile of inferred SNP allele genotypes were estimated as described in the Methods using the genotype frequency values derived from the corresponding major population group in the 1000 Genomes Project (black line; Subjects: 1–5, European; 6–9, South Asian). Random match probability values were recalculated using the genotype frequency values from each of the major population groups in the 1000 Genomes Project (European, EUR green; African, AFR, red: East Asian, EAS, orange; South Asian, SAS, purple; American, AMR, blue).

To determine the effect of reference sample population on RMP values, probabilities were also calculated using the non-synonymous SNP, or SNP combination, genotype frequencies from all major populations in the 1000 Genomes Project [37]. These included European (EUR, green), African (AFR, red), East Asian (EAS, orange), South Asian (SAS, purple) and Native American (AMR, blue). As illustrated (Fig. 3B), the RMPs calculated from African values were considerably more discriminating, or less conservative, than the other four populations. The likelihood ratios (LR = Pr(GVP profile|matching population)/Pr(GVP profile|African)) ranged from 3.3×10^0 to 3.9×10^6 and averaged 4.4×10^5 with a median of 3.3×10^3. Estimated RMPs using frequency values obtained from other reference populations overlapped and were all within an order of magnitude, the maximum difference was 7.9 fold for UCD008, with a median difference of 2.6 fold across all subjects. Values generated using the genotype frequencies from the matching 1000 Genomes project populations (European and South Asian) resulted in the most conservative estimates for only 2 of the 9 subjects, indicating that more inferred SNP information would be required to further resolve ancestral contribution from non-African populations.

3.5. Use of exome-driven GVP-discovery

In this study an alternative approach of GVP discovery was also used. Exome data from South Asian subjects were reviewed for the presence of non-synonymous variants in gene products known to be in given proteomic datasets. These candidate genetically variant markers were then confirmed by scanning for the presence of corresponding peptides in proteomic datasets. The advantage of this approach is that low frequency is no longer an exclusionary criterion. When screening for candidate GVPs based on proteomic data alone, as conducted above, rare candidate GVPs are excluded due to the high likelihood of a chemical modification accounting for the mass shift as opposed to genetic polymorphism. Accordingly, a non-synonymous SNP (rs149773722) was detected in the desmoplakin gene (*DSP*) corresponding to an amino acid polymorphism of threonine to serine at position 1806 (T1806S) in an exome from the subject UCD008. The peptide NQCSQVVQER was incorporated into the reference protein database and then detected in the three proteomic datasets from subject UCD008 using both X!Tandem and PEAKs software (version 8.5) (Fig. 4). The fragmentation masses corresponding to the serine residue for both the y- (red) and b-series (blue) are indicated. When the peptide corresponding to the minor allele of the rare non-synonymous SNP rs149773722 was incorporated into the calculation of the RMP of subject UCD008, the value increased from 1 in 8.5×10^5 to 7.4×10^8 [42]. The SNP rs149773722 does not occur in the 1000 Genomes Project (phase 3), although it does occur at a low rate in the Genome Aggregation Database (4/246032 chromosomes), in the Trans-Omics for Precision Medicine (4/125568) and NHLBI exome sequencing project (1/13006) [44,46]. To be statistically consistent, we used values solely from the 1000 Genomes Project (0/489 subjects) and incorporated Jefferys prior into all calculations [42]. These result in an estimated genotypic probability (Pr(rs149773722|South Asian Population)) of 1.15×10^{-3}. This is a more conservative value than that obtained using the larger alternative databases above by orders of magnitude. As a point of interest, the genotype for subject UCD008 was homozygous for this rare allele, pointing perhaps to a degree of consanguinity in this subject's

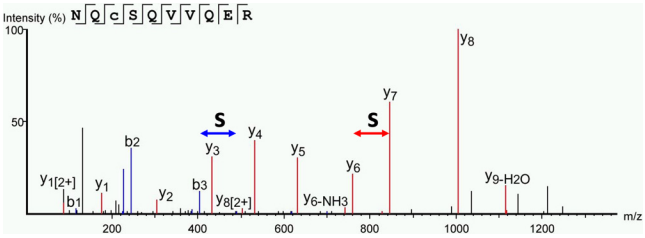

Fig. 4. Example of a Personal Genetically Variant Peptide Biomarker.
The exome of subject UCD008 was analyzed and a non-synonymous SNP (rs#149,773,722) that was not represented in the 1000 Genome Project was detected in the gene product of the desmoplakin gene (GN = *DSP*). The SNP changes the codon assignment of the threonine at position 1806 to serine (S). The polymorphism was incorporated into a reference protein database and the resulting peptide was detected using PEAKs™ software (version 8.5). All flanking fragmentation masses for serine (S) in both the b- (b_3 and b_4, blue) and y- series (y_6 and y_7, red) were detected and are indicated.

ancestry.

3.6. Analysis of linkage disequilibrium between GVP-Inferred SNPs and STR loci

Proteomic genotyping results in an inferred genotype of non-synonymous SNP alleles that was used above to estimate RMP. Because these SNPs are autosomal this genotype has the potential to also be incorporated into the product rule with autosomal STR genotypes, resulting in a single estimate of RMP. Combination of these probabilities into a single value assumes statistical independence between STR and inferred non-synonymous SNP loci and the absence of linkage disequilibrium effects. Linkage disequilibrium has been measured to decay by roughly 50% for each 60 kbp in nucleotide distance in European populations [47,48]. The nucleotide location (GRCh38 build) of each GVP-inferred SNP and the STR loci from the Illumina ForenSeq™ DNA Signature Prep Kit were identified and each nucleotide distance measured (bp) on the same chromosome (Fig. 5) [49]. The closest distance was 2.2 Mb between rs214814 and D20S482 on chromosome 20, 37-fold greater than the point of 50% linkage disequilibrium. Ninety per cent of the measured nucleotides distances were greater than 32 Mb.

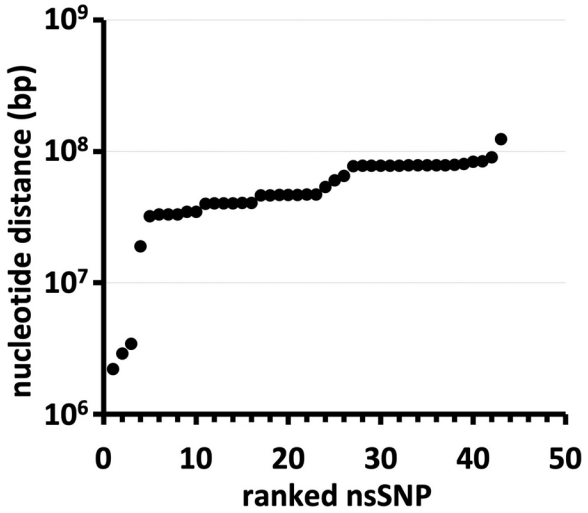

Fig. 5. Nucleotide Distance between Peptide-Inferred SNPs and STRs.
The nucleotide distance (bp) was measured between all GVP-inferred non-synonymous SNPs (nsSNPs) and STR loci on the same chromosome (Illumina ForenSeq™ DNA Signature Prep Kit). The distances for each SNP were ranked and plotted. Ninety percent of GVP-inferred SNPs were greater than 32 Mb from the nearest STR with the closest distance occurring at 2.2 Mb.

4. Discussion

This project describes the discovery, characterization and validation of 60 genetically variant peptides (GVPs) from the epidermal corneocyte proteome. The study was conducted using 40 proteomic datasets and 9 exome datasets from 9 subjects of European and South Asian origin. The genetic peptide biomarkers were primarily discovered using an analysis of single amino acid variants based on peptide fragmentation spectra [36]. The identifying potential of these peptides is significant, individual random match probabilities (RMPs) range up to values of over 1 in 100 million. While continental resolution of genetic background remains problematic, the data do show consistently lower relative likelihood of African ancestry as a source of proteomic material. The identification and use of private GVPs was also explored. These rare biomarkers are discovered through an analysis of non-synonymous SNPs in subjects exomes that were predicted to be represented in proteomic datasets and were confirmed by identifying the presence of the predicted spectra in the respective datasets. The compatibility of the GVP-inferred genotypes with RMPs derived from STR-based genotypes was also investigated. The nucleotide distance between GVP-inferred SNPs and STRs was well beyond the point where linkage disequilibrium effects would raise statistical dependence beyond background levels.

This project focused on the proteome of fully differentiated squamous epidermal corneocytes [8,27,28] for two reasons. First, this cellular population is easy to isolate and process. Second, continuously shed from the stratum corneum, these cells are more likely to be associated with the donor of the fingermark. The population of GVPs identified from these cells, therefore, is likely to reflect the genotype of the donor. Other protein populations will be present in fingermarks. Some of these may provide forensic intelligence or probative context, such as the presence of blood proteins or bodily fluids, but may not necessarily originate with the donor [50–52]. The proteome of eccrine sweat is also more likely to contain endogenous peptide markers [1,8,53,54].

The study analyzed the corneocyte from four different body regions only one of which, palmar skin, is responsible for the production of fingerprints. Palmar skin has a distinctively thicker stratum corneum, friction ridges, no hair follicles and differences in the distribution and type of sweat glands [1,5]. These anatomical differences are reflected by differences at the proteome level (Figure S1), which is distinctive and dominated by a relative small set of proteins. Other proteins, including those containing the GVPs identified in this study, are relatively under-represented when compared to other corneocyte preparations. Therefore the yield of GVPs from palmar skin was less than other skin types (Figure S3). The range of detection covers several orders of magnitude and even low abundance proteins may provide good quality unambiguous fragmentation spectra. Nevertheless there is an increased potential for GVPs to drop below the lower limit of detection when the gene product is less abundant. Targeted data acquisition of specific peptides, or more powerful instrumentation, can mitigate this by increasing the sensitivity and quality of peptide and fragmentation data [55–57]. Fingermark aging is also a factor to consider. A recent study suggested that a reduction in complexity of the fingermark proteome might be a means to estimate the time since deposition [12]. One feature of that particular study was that only low levels of reductant (about 6 mM dithiothreitol) were used. This would not solubilize all proteins given the high levels of disulphide bonds in the cornified envelope [29]. The focus on the less cross-linked, less stable elements of the fingermark proteome may introduce bias and potentially reduce the apparent half-life of corneocyte proteins.

The GVPs identified in this study were primarily identified through empirical detection and screening to eliminate peptides that had multiple addresses in the genome [13]. The selection of potential GVPs should be stringent. Poor quality spectra with low expectation scores (log(e)) were eliminated [13]. Genetically variant peptides that match the mass shifts present in chemically variant peptides, such as

methionine oxidation (M to F), deamidation (N/Q to D/E) and variants with the same mass (Q to K and I to L), would also be poorly predictive and were also removed from the list of useable biomarkers (Table S2) [13]. Miss-assignment of a chemical variant peptide as a genetically variant peptide is common and these incorrect peptide sequences would still score highly since many of the fragmentation masses are the same. In this study one GVP pair (KRT78, rs#2,013,325), contained a leucine to proline polymorphism (Fig. 2 & S2). This has the same mass shift as proline oxidation, a common post-translational modification in extracellular matrix collagens, but is included since it was predictive in this study and proline oxidation is not common among keratinized tissue [58]. As another point of interest, the *POF1B* gene is located on the X-chromosome. This points to the theoretical possibility of the genetically variant peptides associated with the SNP alleles for rs363774 being used to demonstrate female sex in a fingermark donor. While one of these alleles was inferred from 6 of the 9 cumulative proteomic datasets, only one subject was heterozygote for that locus and no subject had both corresponding alleles in their cumulative proteomic dataset. One method to increase the likelihood of identifying an actual genetically variant peptide is to focus on peptides that have a common matching SNP allele (MAF > 0.5%) [13]. Generally it is not possible to quantify the probability of a chemical modification on a peptide, but it is possible to determine the probability of a genetic variation due to the inclusion of matching SNP alleles in extant SNP databases. Increased allele frequency therefore increases the likelihood of correct peptide sequence assignment.

The estimation of RMP in this study assumes complete correlation of SNP variants within an open reading frame and complete independence outside of it [13]. The initial calculations used the genotype frequencies that were derived from the matching major reference population in the 1000 Genomes Project, the European and South Asian population groups. When the estimates were recalculated using all of the major reference populations the estimate of RMP obtained from the African population were substantially more discriminating, reflecting a deeper evolutionary history of the African population and increased biodistance relative to other population groups [13,37,47,59–61]. While this shows that an African source for the GVP profiles is less likely, conclusions about the non-African continental origin of the subjects proteome could not be made based on the data obtained. The estimates of RMP based on other continental genotype frequencies resulted in values that were within an order of magnitude. The discovery and detection of additional GVPs has the potential to resolve the continental origin of subject's corneocyte proteomes [14,62,63]. A recent paper examining GVPs in East Asian hair was able to show that RMPs calculated using East Asian frequencies from the 1000 Genomes Project were consistently more conservative than those obtained using European or African values, potentially statistically separating two non-African populations [64]. The use of mass spectrometry techniques such as parallel reaction monitoring, data independent acquisition, and more sensitive mass spectrometry instrumentation, will increase the number of detected GVPs in a fingermark and add data points for RMP calculations and likelihood values for the differing genetic backgrounds.

Fingermark analysis is often further complicated by the presence of multiple contributors. Deconvolution of the various contributors, particularly in conditions of high sample limitation, can be challenging. Proteomic analysis provides some tools for the forensic investigator in separating out individual proteomic signals [2,17]. One approach when using GVPs is the use of rare peptide biomarkers that correspond to rare, private non-synonymous SNPs. Each individual has a number of novel non-synonymous SNPs, that are either not represented in aggregation databases or occur very infrequently [16,37]. The detection of these personal GVPs can resolve the various contributors of a proteomic signal. Identification of these rare variants is problematic however when depending on fragmentation data alone for identification of potential GVPs, because rare candidate peptides are typically false positive sequence assignments. An exome - dependent analysis of non-synonymous SNP alleles predicted to be present in proteomic datasets is necessary for identification of rare variant peptides, although it comes with its own set of assumptions, some of which are not amenable to standard forensic casework. It assumes prior knowledge of a subject's exome and requires a customized investigative workflow [65]. The private GVP described above from the desmoplakin gene product (GN = *DSP*, rs149773722, Fig. 4) is an example of proteomic information that can resolve complex contributor patterns by virtue of rarity. In a mixture only one individual is likely to be a contributor for that particular peptide or combination of rare GVPs. Care must be taken to exclude family members who would be likely to share rare alleles. This type of personalized genetic analysis is, for now, highly expensive, labor intensive and occurs in an uncertain ethical and legal framework [66]. These unique markers would be difficult to incorporate into a rigorous statistical framework, for example they would have high levels of linkage disequilibrium and standard deviations well in excess of the estimated genotype frequencies [48,67]. In spite of these factors the use of complete forensic genomic datasets, including rare variants, are beginning to be incorporated into casework [68].

Perhaps the most significant potential of proteomic genotyping is the ability to complement and enhance RMPs from incomplete or absent STR profiles, a common occurrence when relying on touch DNA [6,18,69]. However, combining the RMPs of SNP genotypes with STR profiles assumes statistical independence, a criterion that is not met when genetic loci are close together and are under linkage disequilibrium. Linkage disequilibrium decays as loci become more distant to each other in nucleotide space [48,67,70]. At 60 kb these effects average around 50% and decay to background levels around 200 kb [47,48]. The correlations between separate loci are highly variable however and should ideally be empirically measured instead of depending solely on assumptions based on nucleotide distance.

A multi-informative approach incorporates all forms of analysis into the study of a single fingermark [4,11,71]. This allows for the triangulation of physical data, impression patterns of friction ridge skin, genetic data from nucleotide and now proteomic sources. The contribution of proteomic data to the genetic information extracted from a fingermark will increase with identification of additional GVPs, and more sensitive proteomic data acquisition. Additional work is needed to develop sample workflows that are compatible with DNA-based modalities and recognize the differing chemistries of DNA and protein in the fingermark [21,22,72,73]. With further development external stable isotope standard peptide mixtures will be developed and characterized on a range of mass spectrometry platforms that are available to the forensic investigator. The GVPs identified, characterized and validated in this study provide a foundation for additional studies developing this approach to fingermark analysis. We predict that this additional point of triangulation will increase the probative and actionable information that can be derived from fingermark and other trace evidence.

Disclaimers

The authors have declared no conflict of interest, with the exception of GJP who has a patent based on the use of genetically variant peptides for human identification (US 8,877,455 B2, Australian Patent 2011229918, Canadian Patent CA 2794248, and European Patent EP11759843.3). The patent is owned by Parker Proteomics LLC. Protein-Based Identification Technologies LLC (PBIT) has an exclusive license to develop the intellectual property and is co-owned by Utah Valley University and GJP. This ownership of PBIT and associated intellectual property does not alter policies on sharing data and materials. These financial conflicts of interest are administered by the Research Integrity and Compliance Office, Office of Research at the University of California, Davis to ensure compliance with University of California Policy.

CRediT authorship contribution statement

Trevor Borja: Data curation, Formal analysis, Investigation, Visualization, Writing - review & editing. **Noreen Karim:** Data curation, Formal analysis, Investigation, Visualization, Writing - review & editing. **Zachary Goecker:** Data curation, Formal analysis, Methodology. **Michelle Salemi:** Investigation. **Brett Phinney:** Resources, Supervision, Methodology. **Muhammad Naeem:** Conceptualization, Funding acquisition, Supervision. **Robert Rice:** Conceptualization, Data curation, Funding acquisition, Supervision, Methodology, Visualization, Writing - review & editing. **Glendon Parker:** Conceptualization, Formal analysis, Funding acquisition, Supervision, Project administration, Methodology, Visualization, Writing - original draft, Writing - review & editing.

Acknowledgments

This work would not be possible without the support of the Department of Justice, National Institute of Justice (DX-BN-K065) and the International Research Support Initiative Program of the Higher Education Commission of Pakistan and the USDA(NIFA)/ The University of California Agricultural Research Station. The authors thank Dr Susan Walsh of Indiana University – Purdue University Institute, for her advice. This publication was made possible, in part, with support from the UC Davis Genome Center Bioinformatics Core Facility. The sequencing was carried out at the DNA Technologies and Expression Analysis Cores at the UC Davis Genome Center, supported by NIH Shared Instrumentation Grant1S10OD010786-01. We specifically acknowledge the assistance of Jie Li, Emily Kumimoto, Siranoosh Ashtari, Vanessa K Rashbrook, and Lutz Froenicke.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.fsigen.2019.05.005.

References

[1] R. Ramotowski, A.K. Datta (Ed.), Composition of Latent Print Residue, in Advances in Fingerprint Technology, CRC Press LLC, 2001, pp. 63–104.
[2] S. Francese, Techniques for fingermark analysis using MALDI MS: a practical overview, in: R. Cramer (Ed.), Advances in MALDI and Laser-Induced Soft Ionization Mass Spectrometry, Springer, 2016, pp. 93–128.
[3] J.E. Templeton, A. Linacre, DNA profiles from fingermarks, Biotechniques 57 (5) (2014) 259–266.
[4] A. van Dam, et al., Techniques that acquire donor profiling information from fingermarks - a review, Sci. Justice 56 (2) (2016) 143–154.
[5] Scientific Working Group on Friction Ridge Analysis, The Fingerprint Sourcebook, National Institute of Justice, Department of Justice, Washington D.C, 2011.
[6] I. Hefetz, et al., Touch DNA: the effect of the deposition pressure on the quality of latent fingermarks and STR profiles, Forensic Sci. Int. Genet. 38 (2019) 105–112.
[7] S. Fieldhouse, E. Oravcova, L. Walton-Williams, The effect of DNA recovery on the subsequent quality of latent fingermarks, Forensic Sci. Int. 267 (2016) 78–88.
[8] A. Girod, R. Ramotowski, C. Weyermann, Composition of fingermark residue: a qualitative and quantitative review, Forensic Sci. Int. 223 (1-3) (2012) 10–24.
[9] C. Huynh, J. Halámek, Trends in fingerprint analysis, Trends Anal. Chem. 82 (2016) 328–336.
[10] R.A.H. van Oorschot, et al., DNA transfer in forensic science: a review, Forensic Sci. Int. Genet. 38 (2019) 140–166.
[11] S. Francese, et al., Beyond the ridge pattern: multi-informative analysis of latent fingermarks by MALDI mass spectrometry, Analyst 138 (15) (2013) 4215–4228.
[12] S. Oonk, et al., Proteomics as a new tool to study fingermark ageing in forensics, Sci. Rep. 8 (1) (2018) 16425.
[13] G.J. Parker, et al., Demonstration of protein-based human identification using the hair shaft proteome, PLoS One 11 (9) (2016) e0160653.
[14] F. Oldoni, K.K. Kidd, D. Podini, Microhaplotypes in forensic genetics, Forensic Sci. Int. Genet. 38 (2019) 54–69.
[15] I.W. Evert, B.S. Weir, Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists, 1st ed., Sinauer Associates, 1998.
[16] K.E. Mason, et al., Protein-based forensic identification using genetically variant peptides in human bone, Forensic Sci. Int. 288 (2018) 89–96.
[17] L.S. Ferguson, et al., Direct detection of peptides and small proteins in fingermarks and determination of sex by MALDI mass spectrometry profiling, Analyst 137 (20) (2012) 4686–4692.
[18] B. Martin, et al., DNA profiles generated from a range of touched sample types, Forensic Sci. Int. Genet. 36 (2018) 13–19.
[19] A. van Dam, et al., Simultaneous labeling of multiple components in a single fingermark, Forensic Sci. Int. 232 (1-3) (2013) 173–179.
[20] E. Patel, et al., Alternative surfactants for improved efficiency of in situ tryptic proteolysis of fingermarks, J. Am. Soc. Mass Spectrom. 26 (6) (2015) 862–872.
[21] S.A. Sterling, et al., Combined DNA typing and protein identification from unfired brass cartridges, J. Forensic Sci. (2019) in press.
[22] S. Kranes, et al., Simultaneous DNA and protein extraction using trypsin, Forensic Sci. Int. Genet. Suppl. Ser. 6 (2017) e203–e204.
[23] T. Lindahl, The Croonian Lecture, 1996: endogenous damage to DNA, Philos. Trans. R. Soc. Lond. B Biol. Sci. 351 (1347) (1996) 1529–1538.
[24] C. Ottoni, et al., Preservation of ancient DNA in thermally damaged archaeological bone, Naturwissenschaften 96 (2) (2009) 267–278.
[25] H.N. Poinar, B.A. Stankiewicz, Protein preservation and DNA retrieval from ancient tissues, Proc. Natl. Acad. Sci. U. S. A. 96 (15) (1999) 8426–8431.
[26] E.K. Hanson, J. Ballantyne, "Getting blood from a stone": ultrasensitive forensic DNA profiling of microscopic bio-particles recovered from "touch DNA" evidence, Methods Mol. Biol. 1039 (2013) 3–17.
[27] T. Matsui, M. Amagai, Dissecting the formation, structure and barrier function of the stratum corneum, Int. Immunol. 27 (6) (2015) 269–280.
[28] L. Eckhart, et al., Cell death by cornification, Biochim. Biophys. Acta 1833 (12) (2013) 3471–3480.
[29] R.H. Rice, et al., Corneocyte proteomics: applications to skin biology and dermatology, Exp. Dermatol. 27 (8) (2018) 931–938.
[30] N. Karim, et al., Proteomic manifestations of genetic defects in autosomal recessive congenital ichthyosis, J. Proteomics 201 (2019) 104–109.
[31] R.H. Rice, G.E. Means, W.D. Brown, Stabilization of bovine trypsin by reductive methylation, Biochim. Biophys. Acta 492 (2) (1977) 316–321.
[32] R.H. Rice, Proteomic analysis of hair shaft and nail plate, J. Cosmet. Sci. 62 (2011) 229–236.
[33] E.L. de Graaf, et al., Improving SRM assay development: a global comparison between triple quadrupole, ion trap, and higher energy CID peptide fragmentation spectra, J. Proteome Res. 10 (9) (2011) 4334–4341.
[34] Y. Perez-Riverol, et al., PRIDE inspector toolsuite: moving toward a universal visualization tool for proteomics data standard formats and quality assessment of Proteome Xchange datasets, Mol. Cell Proteomics 15 (1) (2016) 305–317.
[35] N.H. Tran, et al., Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry, Nat. Methods 16 (1) (2019) 63–66.
[36] D. Fenyö, J. Eriksson, R. Beavis, Mass spectrometric protein identification using the global proteome machine, Methods Mol. Biol. 673 (2010) 189–202.
[37] Genomes Project Consortium, et al., A global reference for human genetic variation, Nature 526 (7571) (2015) 68–74.
[38] H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, arXiv 1303 (3997) (2013).
[39] R. Poplin, et al., Scaling accurate genetic variant discovery to tens of thousands of samples, bioRxiv 201178 (2018) 1–22.
[40] M.H. Zweig, G. Campbell, Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine, Clin. Chem. 39 (4) (1993) 561–577.
[41] J.M. Butler, Fundamentals of Forensic DNA Typing, Academic Press, 2010.
[42] H. Jeffreys, An invariant form for the prior probability in estimation problems, Proc. R. Soc. Lond. A Math. Phys. Sci. 186 (1007) (1946) 453–461.
[43] A. Gelman, et al., Bayesian data analysis, CRC Texts in Statistical Science, second edition, Chapman & Hall, 2003 Vol. Book 106.
[44] M. Lek, et al., Analysis of protein-coding genetic variation in 60,706 humans, Nature 536 (7616) (2016) 285–291.
[45] R.H. Rice, et al., Distinguishing ichthyoses by protein profiling, PLoS One 8 (10) (2013) e75355.
[46] J.A. Brody, et al., Analysis commons, a team approach to discovery in a big-data environment for genetic epidemiology, Nat. Genet. 49 (11) (2017) 1560–1563.
[47] D.E. Reich, et al., Linkage disequilibrium in the human genome, Nature 411 (6834) (2001) 199–204.
[48] M.A. Eberle, et al., Allele frequency matching between SNPs reveals an excess of linkage disequilibrium in genic regions of the human genome, PLoS Genet. 2 (9) (2006) e142.
[49] C. Phillips, et al., Global patterns of STR sequence variation: sequencing the CEPH human genome diversity panel for 58 forensic STRs using the Illumina ForenSeq DNA Signature Prep Kit, Electrophoresis 39 (21) (2018) 2708–2724.
[50] S. Kamanna, et al., A complementary forensic 'proteo-genomic' approach for the direct identification of biological fluid traces under fingernails, Anal. Bioanal. Chem. 410 (2018) 6165–6175.
[51] S. Kamanna, et al., Bottom-up" in situ proteomic differentiation of human and non-human haemoglobins for forensic purposes by matrix-assisted laser desorption/ionization time-of-flight tandem mass spectrometry, Rapid Commun. Mass Spectrom. 31 (22) (2017) 1927–1937.
[52] S. Kamanna, et al., A mass spectrometry-based forensic toolbox for imaging and detecting biological fluid evidence in finger marks and fingernail scrapings, Int. J. Legal Med. 131 (5) (2017) 1413–1422.
[53] P. Kanokwongnuwut, et al., Shedding light on shedders, Forensic Sci. Int. Genet. 36 (2018) 20–25.
[54] V. Drapel, et al., Identification of promising antigenic components in latent fingermark residues, Forensic Sci. Int. 184 (1-3) (2009) 47–53.
[55] N. Rauniyar, Parallel reaction monitoring: a targeted experiment performed using high resolution and high mass accuracy mass spectrometry, Int. J. Mol. Sci. 16 (12) (2015) 28566–28581.
[56] S.A. Carr, et al., Targeted peptide measurements in biology and medicine: best

practices for mass spectrometry-based assay development using a fit-for-purpose approach, Mol. Cell Proteomics 13 (3) (2014) 907–917.
[57] F. Meier, et al., Online parallel accumulation-serial fragmentation (PASEF) with a novel trapped ion mobility mass spectrometer, Mol. Cell Proteomics 17 (12) (2018) 2534–2545.
[58] L.D. Lee, H.P. Baden, Chemistry and composition of the keratins, Int. J. Dermatol. 14 (3) (1975) 161–171.
[59] W. Fu, et al., Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants, Nature 493 (7431) (2013) 216–220.
[60] Q. Fu, et al., A revised timescale for human evolution based on ancient mitochondrial genomes, Curr. Biol. 23 (7) (2013) 553–559.
[61] K. Bryc, et al., Genome-wide patterns of population structure and admixture in West Africans and African Americans, Proc. Natl. Acad. Sci. U. S. A. 107 (2) (2010) 786–791.
[62] J.L. King, et al., Increasing the discrimination power of ancestry- and identity-informative SNP loci within the ForenSeq DNA Signature Prep Kit, Forensic Sci. Int. Genet. 36 (2018) 60–76.
[63] K.K. Kidd, et al., Evaluating 130 microhaplotypes across a global set of 83 populations, Forensic Sci. Int. Genet. 29 (2017) 29–37.
[64] F. Lei, et al., Development and validation of protein-based forensic ancestry inference method using hair shaft proteome, Prog. Biochem. Biophys. 46 (1) (2019) 81–88.
[65] K.E. Mason, et al., Development of a protein-based human identification capability from a single hair, J. Forensic Sci. (2019) p. Epub ahead of print.
[66] D. Syndercombe Court, Forensic genealogy: some serious concerns, Forensic Sci. Int. Genet. 36 (2018) 203–204.
[67] K.G. Ardlie, L. Kruglyak, M. Seielstad, Patterns of linkage disequilibrium in the human genome, Nat. Rev. Genet. 3 (4) (2002) 299–309.
[68] C. Phillips, The Golden State Killer investigation and the nascent field of forensic genealogy, Forensic Sci. Int. Genet. 36 (2018) 186–188.
[69] J. Burrill, B. Daniel, N. Frascione, A review of trace "Touch DNA" deposits: variability factors and an exploration of cellular composition, Forensic Sci. Int. Genet. 39 (2019) 8–18.
[70] J.K. Pritchard, M. Przeworski, Linkage disequilibrium in humans: models and data, Am. J. Hum. Genet. 69 (1) (2001) 1–14.
[71] E. Brunelle, et al., Fingerprint analysis: moving toward multiattribute determination via individual markers, Anal. Chem. 90 (1) (2018) 980–987.
[72] K. Falkena, et al., Prediction of DNA concentration in fingermarks using autofluorescence properties, Forensic Sci. Int. 295 (2018) 128–136.
[73] C.E. Stanciu, et al., Optical characterization of epidermal cells and their relationship to DNA recovery from touch samples, F1000Res 4 (2015) 1–1.