Forensic Science International: Genetics Supplement Series 7 (2019) 664–666

Contents lists available at ScienceDirect

Forensic Science International: Genetics Supplement Series

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

Proteomic genotyping: Using mass spectrometry to infer SNP genotypes in a forensic context

Glendon Parker[a,*], Zachary Goecker[a], Rachel Franklin[a], Blythe Durbin-Johnson[b], Jennifer Milan[a], Noreen Karim[a], Christina De Leon[a], Ashleigh Matzoll[a], Trevor Borja[a], Bob Rice[a]

[a] Department of Environmental Toxicology, University of California, Davis United States
[b] Department of Public Health Sciences, University of California, Davis United States

ARTICLE INFO

Keywords:
Proteomic genotyping
Hair
Mass spectrometry
Hair pigmentation
Hair proteomics
Genetically variant peptides
Protein-based human identification

ABSTRACT

Proteins are an intrinsic part of biological evidence. In the last two decades mass spectrometry has revolutionized the analysis of proteins, allowing for thousands of peptides to be detected in a single analysis, including peptides containing single amino acid polymorphisms. Identification of these genetically variant peptides (GVP) allows for inference of the underlying non-synonymous SNP genotype. We have focused on hair proteins as a source of genetic information, discovering and validating 240 GVPs. Using optimized processing, a single hair shaft can obtain random match probabilities (RMPs) of up to 1 in 100 million. We show that GVP-inferred genotypes are not affected by the anatomical origin of the hair shaft or pigmentation. Harsh oxidation with peroxide does not affect RMPs. Importantly, GVP-inferred SNP genotypes are statistically compatible with STR-typing and *Alu* retroelements, with 90% of GVP-inferred SNPs being located greater than 10 million and 75 million bp from the nearest STR or *Alu* element respectively. We have shown separation of DNA and peptide workflows from the same samples. Recent work has shifted to the use of a targeted peptide assay. When data acquisition of GVPs was dynamically triggered by the detection of a standard peptide, this approach resulted in a 2.5 fold increase in sensitivity and increased confidence in peptide detection. Based on Monte Carlo modeling, we predict this increase in sensitivity will result in RMPs in excess of 1 in 1 trillion from a single hair shaft. Proteomic genotyping can also be applied to any protein matrix, including fingermarks and bone. Proteomic genotyping therefore has potential to complement partial PCR-based DNA typing and to provide other options for investigators.

1. Introduction

Proteomic genotyping is the use of proteomic data to infer genotype. Non-synonymous SNP alleles change the codon assignment of a protein resulting in a single amino acid polymorphism. When genetically variant peptides (GVPs) containing these polymorphisms are detected in a proteomic dataset, then the status of the corresponding SNP allele can be inferred (Fig. 1) [1]. This process does not depend on the presence of accessible or useable DNA in a sample, making it an attractive alternative for problematic forensic samples, such as hair shafts, degraded bones or teeth, fingermarks, or sexual assault evidence.

When proteins are processed for mass spectrometry they are digested in a predictable manner using the protease trypsin. This results in a complex mixture of peptides that is resolved using liquid chromatography, and the mass of each peptide measured using mass spectrometry. Selected peptides are then fragmented and spectra of resulting masses obtained for each peptide. These spectra are unique and can be used to infer the presence of a protein in a sample. Spectra corresponding to a genetically variant peptide can be used to infer the presence of the corresponding non-synonymous SNP allele. In aggregate a profile of inferred SNPs can be used to develop a random match probability and likelihood of ancestry.

2. Methods

2.1. Tissue procurement and processing

Cranial hair shafts were collected from subjects with informed consent using a University of California IRB protocol (IRB# 832726). Each sample was processed as either single hair shafts or in 4 mg batches, using an optimized hair sample processing protocol and applied to a ThermoScientific Q Exactive Plus Orbitrap mass spectrometer with

* Corresponding author at: 4251B Meyer Hall, One Shields Ave, 95616, Davis, CA United States.
E-mail address: gjparker@ucdavis.edu (G. Parker).

https://doi.org/10.1016/j.fsigss.2019.10.130
Received 9 October 2019; Accepted 10 October 2019
Available online 18 October 2019
1875-1768/

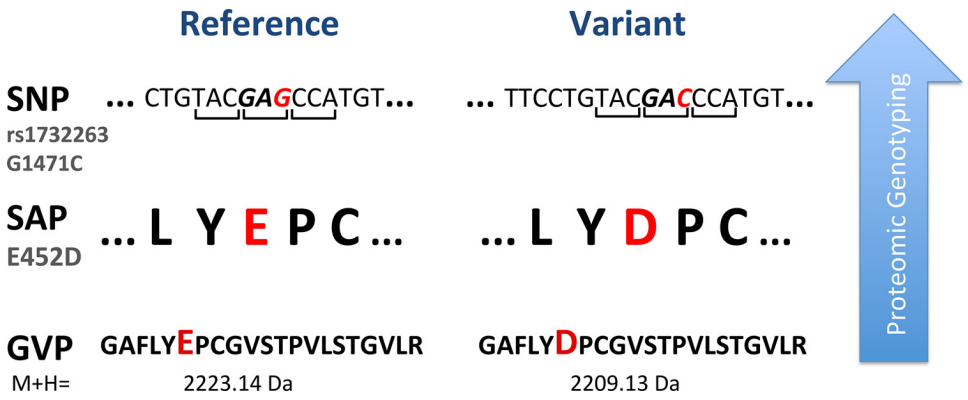

Fig. 1. Proteomic genotyping is the use of proteomic data to infer SNP genotype. A non-synonymous SNP results in a change in codon assignment, in this case G1471C in the gene KRT82 (rs1732263). This variant results in a change in codon assignment from a glutamate (E) to an aspartate (D) at amino acid in position 452. These changes can be detected in proteomic mass spectrometry, since the trypsin digested peptide results in a shift in peptide mass and changes in the resulting fragmentation spectrum.

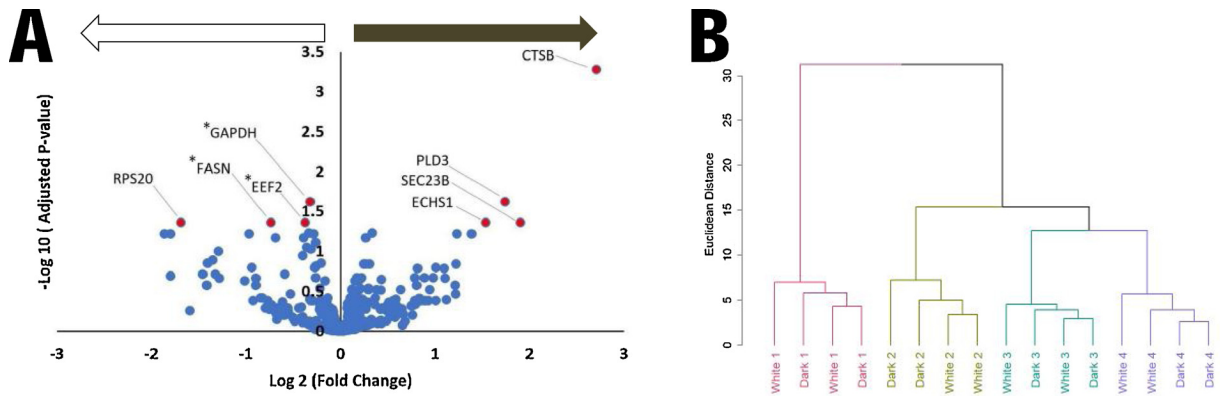

Fig. 2. Proteomic and GVP Profile Changes in Pigmented and Non-Pigmented Hair. A) Matching pigmented and non-pigmented hair was processed and abundance was measured. The fold change (pigmented/non-pigmented) and adjusted p-value were calculated for each protein or protein group (*). Significantly changed proteins are labelled and indicated in red. Proteins more abundant in pigmented hair are on the right hand side of the plot. **B)** Profiles of genetically variant peptides were extracted in four subjects using matching pigmented and non-pigmented samples. Euclidian distance was measured and plotted as a hierarchical dendrogram (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

inbuilt Proxeon nanospray and Proxeon Easy-nLC II HPLC [2,3].

2.2. Bioinformatic analysis and GVP detection

The resulting datasets (.RAW format) were formatted using MSConvert (.mzML format) and processed using the X!Tandem peptide spectra matching algorithm (thegpm.org). GVPs were detected using the single amino acid variant (SAV) function in theGPM algorithm [2]. The protein amounts were quantified by the Global Proteome Machine (www.thegpm.org) as described elsewhere [3]. Detected GVP peptides were collated and used to compile a genotype of inferred SNP alleles [1]. The positive detection of a GVP was recorded in a binary format where a “1” represented detection and a “0” was displayed if the GVP was not detected. The GVP assignments were weighted by the inverse of their corresponding allele frequency to place more value on the less common GVPs. Each individual’s GVP profiles were combined into a matrix and exported into R, version 3.2.1 for further analysis. The R dist function was used to calculate Euclidean distance coefficients in a matrix to identify to what extent the samples are similar. The distance coefficients were then used to create a hierarchical clustering dendrogram using the hclust function.

3. Results

3.1. Optimization of hair shaft sample processing

Application of this technology with hair shafts in a forensic context requires several milestones to be met. The resulting random match probabilities need to be useable, even when obtained from a single hair shaft. We optimized the processing of hair shafts by increasing the concentration of reductant DTT to 100 mM, reducing the temperature, and the time of extraction and trypsin digestion. When tested in biological triplicates on a cohort of three European and three African hair samples, significant improvements were detected. The optimized conditions resulted in an increase of peptide identifications from an average of 1585 to 2703 peptides, and an increase in detection of genetically variant peptides from 20 ± 5 to 73 ± 5 peptides. Corresponding random match probabilities increased from an average of 1 in 158 to 1 in 45 million. When these probabilities were recalculated using genotype frequencies from both the European and African reference populations in the 1000 Genomes Project, the resulting likelihood values (LR = Pr(GVP profile|EUR)/Pr(GVP profile|AFR)) followed predicted outcomes [4]. African samples were positive and increased by 4.3 fold to a maximum of 5.7×10^4. European likelihoods were negative and reduced further by 1.5 orders of magnitude to a minimum of 1×10^{-4}.

3.2. Validation using real-world samples

Finally we need to demonstrate that the resulting profiles of inferred SNP alleles are robust and not overly dependent on environmental conditions or biological history of the sample. Investigators may reasonably obtain forensic hair evidence from different body sites, pigmented or grey hair, or hair that has been chemically treated. We conducted a series of validation studies using the optimized sample processing chemistry to evaluate the effect of biological or chemical history of the GVP profile. The initial study was focused on the effect of body site origin on GVP profiles [3]. Hair shafts from 4 different body sites were processed in 4 biological replicates in 5 individuals. The protein profiles were processed and significant differences in

abundance were quantified between each hair type with the maximum being present between beard and other hair types, the minimum differences occurring between pubic and axillary hair, with scalp hair being in between. In spite of these differences, a correlation analysis showed GVP profiles were quite stable with greater association with an individual genotype than with other hair types [3]. A study of pigmented and grey hair gave similar results (Fig. 2). Pigmented and non-pigmented hairs from four individuals were processed in biological duplicates. The significant protein differences in non-pigmented hair are indicated in a volcano plot of the logarithm of fold change compared to the logarithm of p-value (Fig. 2A). In spite of these differences in protein abundance, the GVP profile was, again, more dependent on individual genotype than the biological status of the hair sample.

4. Conclusion

Single hair shafts are a significant source of forensically useful genetic information in the form of genetically variant peptides. Using optimized chemical processing, we demonstrate that the resulting profiles correlate more with an individual's genotype than the body site origin, or pigmentation status of the hair shafts.

Disclaimer

GJP 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).

Acknowledgements

This study was supported by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice (Award 2015-DN-BX-K065).

References

[1] G.J. Parker, T. Leppert, D.S. Anex, J.K. Hilmer, N. Matsunami, L. Baird, J. Stevens, K. Parsawar, B.P. Durbin-Johnson, D.M. Rocke, C. Nelson, D.J. Fairbanks, A.S. Wilson, R.H. Rice, S.R. Woodward, B. Bothner, B.R. Hart, M. Leppert, Demonstration of protein-based human identification using the hair shaft proteome, PLoS One 11 (9) (2016) e0160653, , https://doi.org/10.1371/journal.pone.0160653.

[2] Z.C. Goecker, M.R. Salemi, B.S. Phinney, R.H. Rice, G.J. Parker, The Optimization of Human Hair Proteomic Processing for Single Hair and Ancestral Analysis, American Association of Forensic Science, Seattle, WA, 2018.

[3] J.A. Milan, P.-W. Wu, M.R. Salemi, B.P. Durbin-Johnson, D.M. Rocke, B.S. Phinney, R.H. Rice, G.J. Parker, Comparison of protein expression levels and proteomically-inferred genotypes using human hair from different body sites, Forensic Sci. Int. 41 (2019) 19–23, https://doi.org/10.1016/j.fsigen.2019.03.009.

[4] A. Genomes Project Consortium, L.D. Auton, R.M. Brooks, E.P. Durbin, H.M. Garrison, J.O. Kang, J.L. Korbel, S. Marchini, G.A. McCarthy, G.R. McVean, A. Abecasis, Global reference for human genetic variation, Nature 526 (7571) (2015) 68–74, https://doi.org/10.1038/nature15393.