Forensic Science International 310 (2020) 110200

Contents lists available at ScienceDirect

Forensic Science International

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

Technical Note

Proteomic genotyping: Using mass spectrometry to infer SNP genotypes in pigmented and non-pigmented hair

Rachel N. Franklin[a], Noreen Karim[a], Zachary C. Goecker[a], Blythe P. Durbin-Johnson[b], Robert H. Rice[a], Glendon J. Parker[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

Article history:
Received 15 November 2019
Received in revised form 23 January 2020
Accepted 13 February 2020
Available online 25 February 2020

Keywords:
Proteomic genotyping
Genetically variant peptides
Hair shafts
Hair pigmentation
Protein-based human identification

ABSTRACT

Proteomic genotyping uses genetically variant peptides that contain single amino acid polymorphisms to infer the genotype of corresponding non-synonymous SNP alleles. We have focused on hair proteins as a source of protein-based genetic information in a forensic context. An optimized sample processing protocol for hair shafts has been developed for use on a single hair that allows us to conduct validation protocols on real world samples. This includes whether the inferred SNP genotypes are robust and not systematically affected by biological or chemical variation in hair proteomes that might be obtained from a crime scene. To this end we analyzed the hair of 4 mature individuals with a mixture of pigmented and non-pigmented hair. We demonstrate significant changes in the proteomes of grey versus pigmented hair. Vesicle specific proteins and lipid catabolism proteins were enriched in pigmented hair, and housekeeping proteins and lipid anabolic enzymes were enriched in grey, non-pigmented hair. The resulting profiles of genetically variant peptides, however, were more correlated with profiles from the same individuals regardless of pigmentation status. Together with other published evidence, this finding indicates that profiles of genetically variant peptides are robust and more correlated with other genetically variant peptide profiles from the same individual irrespective of changes occurring in the hair protein profile. Based on this small sample, investigators using profiles of genetically variant peptides to infer random match probabilities should not expect to observe differences based on the pigmentation of the hair shaft.

1. Introduction

Proteomic genotyping is the use of genetically variant peptides (GVPs), detected in a forensic protein sample, to infer the genotype of corresponding non-synonymous SNP alleles in the donor's genome [1]. This process does not depend on the presence of accessible or intact DNA in a sample. This makes proteomic genotyping an attractive alternative for analysis of problematic forensic samples where DNA extraction can be challenging, such as hair shafts, degraded bones or teeth, or fingermarks [1–5]. To demonstrate the concept in hair shafts we developed an optimized sample processing protocol that could be used with high effectiveness on single hairs [6]. This allows us to determine if the detected profiles of genetically variant peptides are robust and result in a consistent profile of inferred SNP alleles regardless of the chemical or biological history of the sample [3,7].

Several real world scenarios have been evaluated [3,7]. Here we include a study of 4 European subjects that had both pigmented and non-pigmented (or gray) hair shafts. We tested whether (a) protein profiles change as a result of the loss of pigmentation and (b) these changes were reflected in the inferred genotype derived from detection of genetically variant peptides. Using this information we can determine whether the resulting GVP profiles are more dependent on the biological context, in this case pigmentation status, of the sample or the underlying genotype.

2. Methods

2.1. Tissue procurement and processing

Matching pigmented and non-pigmented cranial hair shafts from 4 unrelated self-identified European-Americans were collected UUC Davis IRB# 832726. Samples were manually separated into pigmented and non-pigmented 4 mg fractions, the equivalent

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

http://dx.doi.org/10.1016/j.forsciint.2020.110200
0379-0738/

of 80–120 cm of a cut hair shaft. No attempt was made to exclude any segment of the hairs shaft except for the uncut root. Each sample was processed using an optimized hair sample processing protocol [6,7]. Sample digests were analyzed using a Thermo-Scientific Q Exactive Plus Orbitrap mass spectrometer with inbuilt Proxeon nanospray and Proxeon Easy-nLC II HPLC, under previously outlined conditions [3].

2.2. Analysis of proteomic profiles

The resulting datasets (.RAW format) were converted to a standard format (.mzML) and processed using the X!Tandem peptide spectra matching algorithm (thegpm.org). The protein amounts were quantified by label free quantification using the iBAQ function in MaxQuant and Scaffold (version 4.8.7) that focused on the three most abundant peptides that were specific for each gene product. Protein expression levels of pigmented and non-pigmented samples were compared using the Limma-voom Bioconductor pipeline (limma version 3.38.2, edgeR version 3.24.0, in R 3.5.1) for gene expression analysis and were adjusted for within-subject correlations [8]. This package fits a linear model to individual proteins and then applies empirical Bayes shrinkage to the experimental variances to stabilize the variances and maximize power. A Benjamini-Hochberg false discovery rate adjusted p-value was used for the analysis [9].

The differences between the proteomic profiles of the two groups were also evaluated using the Q-Module function of PEAKS Studio 10.0 (Bioinformatics Solutions Inc., Waterloo, ON, Canada). The raw data files were searched against a validated UNIPROT human reference proteome (uniprot-proteome_UP000005640_Human) using default settings of the algorithm except setting the precursor mass error range and fragment ion to 10 ppm and 0.04 Da, respectively. The PTMs search included cysteine carbamidomethylation as fixed modification, whereas oxidation and dioxidation of methionine, deamidation of glutamines and asparagines, pyroglutamation at glutamines and glutamates, and formylation and acetylation of N-termini and lysines as variable modifications. The resulting datasets, filtered with a 1% FDR, were analyzed by Q-module and a heat map was generated by label free quantitation for proteins identified by two unique peptides and with at least 2-fold difference between the groups. Significance was defined as described in [10]. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD016156 [11].

2.3. Analysis of genetically variant peptide profiles

The GVP peptides in the sample dataset were detected using the single amino acid variant (sav) function in X!Tandem (thegpm.org) [1]. Analysis of the detected profiles of genetically variant peptides (GVPs) was conducted as described in the literature [3]. Briefly, cumulative datasets of observed GVPs for each biological condition were obtained from the four subjects [3]. 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 weighting on the less common more discriminating 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 distances in a matrix to identify to what extent the samples are similar. The distances were then used to create a hierarchical clustering dendrogram using the hclust function [3]. Random match probabilities were calculated as described elsewhere [1,3]. Briefly, each gene was statistically treated as a single locus. Complete linkage was assumed to occur within an open reading frame and complete equilibrium was assumed outside of the gene. The genotype frequencies in the European reference population in the 1000 Genomes Project for each inferred SNP or SNP combination were then measured and the product calculated to estimate the random match probability of the observed GVP profile [1].

3. Results and discussion

Matching pigmented and non-pigmented hair samples from 4 European subjects were processed in biological duplicates as described in the Methods. Absolute protein abundances were quantified using label free quantification and each dataset manually screened to ensure uniform annotation of each gene product, with complementary protein assignments combined into single entries. Pigmented protein levels (n = 8) were compared to corresponding non-pigmented values (n = 8) to obtain both the logarithm of fold change (FC) and negative logarithm of the adjusted P-value ($-\log_{10}$(Adjusted P-value)) (Fig. 1, Table 1A). The resulting volcano plot illustrates that a population of proteins have significantly increased abundance in pigmented compared to non-pigmented hair. These include cathepsin B (CTSB), a protein enriched in melanosomes, and SEC23B, a vesicular transport protein [12,13]. Likewise, in non-pigmented hair, primarily housekeeping gene products were more abundant, such as elongation factor 2 (EEF2) and glyceraldehyde-3-phosphate dehydrogenase (GAPDH). Enzymes involved in lipid catabolism were up regulated and lipid anabolism was down regulated.

An additional analysis of protein profile changes between pigmented and non-pigmented hair shafts was conducted using label free quantitation in PEAKs™ software. Proteins that were identified as having both an average 2-fold change in quantitation and an adjusted P-value of less than 0.05 were identified and clustered both in terms of protein abundance and of individual datasets. (Fig. 2, Table 1B, Supplemental Fig. 1). Ten proteins met the inclusion criteria. Included were cathepsin B (P07858, CTSB), and phospholipase D (Q8IV0S, PLD3) that were increased in pigmented hair and were also included in the volcano plot based on abundances quantified using the Scaffold algorithm (Fig. 1). In addition, transmembrane glycoprotein NMB (Q14956, GPNMB) was also increased in pigmented hair. These proteins are all endosomal proteins, two of which are documented to be associated with melanosomes (GPNMB and CTSB) [14,15]. Non-pigmented grey hair had relative increases in the quantity of fructose-bisphosphate aldolase A (P04075, ALDOA), a housekeeping enzyme in the same pathway, glycolysis, as glyceraldehyde-3-phosphate dehydrogenase. The increase in housekeeping proteins in non-pigmented hair may reflect an overall increase in the ratio of cytoplasm to intermediate filaments in fully differentiated hair shaft corneocytes. In this analysis some structural proteins were more abundant in grey non-pigmented hair including trichohyalin (Q07283, TCHH), Keratin type I cuticular Ha7 (Q76014, KRT37), and keratin-associated protein 4–9 (Q9BYQ8, KRTAP4–9). The proteomic changes associated with the loss of pigmentation were consistent enough that a dendrogram of protein profiles of individual datasets, generated when preparing a heatmap with the PEAKs™ Q-Module algorithm, partitioned pigmented (Dark) and grey non-pigmented (White) hair shaft samples (Fig. 2).

An analysis of the GVP profiles from each individual dataset allowed us to test the hypothesis that demonstrated proteomic changes in hair composition did not systematically introduce bias into the inferred SNP genotype. The profiles were extracted from individual datasets and weighted by the inverse of allelic frequency. Euclidean distance was measured based on the

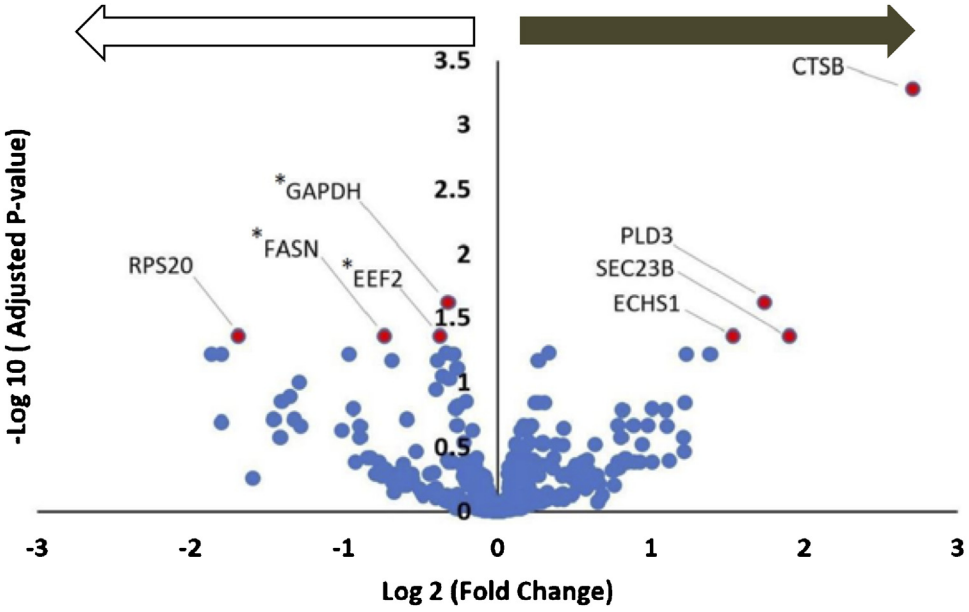

Fig. 1. Volcano Plot of Proteomic Changes in Pigmented and Non-Pigmented Hair. Matching pigmented and non-pigmented hair was processed and abundance measured. The fold change (Log_2(pigmented/non-pigmented)) and adjusted P-value (Log_{10}(Adjusted P-Value)) were calculated for each protein consistently present in the analysis and values plotted on logarithmic scales. Proteins, or protein clusters (*), with significant abundance changes are labeled (red). Proteins more abundant in pigmented hair are on the right side of the plot (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

presence or absence of equivalent GVP markers in other datasets. It should be noted that since exome sequencing was not used to validate the SNP genotypes inferred in this study, inferred genotypes could not be independently validated. The resulting hierarchical clustering dendrogram illustrated that profiles of hair shaft genetically variant peptides from the same individual were more similar to other profiles from the same individual, whereas profiles from other individuals were less similar (Fig. 3). Thus, for the sub-population of peptides with genotype information, the individual genotype was more dominant in determining distances between samples than the significant changes occurring in protein abundance due to the loss of hair pigmentation.

Euclidian distance is a measure of relative similarity, in this case of peptides containing single amino acid polymorphisms. Euclidian distance is not a measure of identity, or association between a reference sample and trace evidence, per se. However, the profile of proteomically-inferred genotypes could also be used to estimate random match probabilities, and provide a measure of whether a hair shaft could be accounted for by the DNA genotype of a randomly selected individual from a reference population. This application of the product-rule was estimated using the approach described in the literature [1,3,5]. In this study the cumulative profile of inferred SNP genotypes resulted in estimated random match probabilities in the four subjects that ranged from

Table 1
Proteins with Significant Changes in Abundance in Pigmented or Non-Pigmented Hair Shafts. Proteins (UniProt identifier, gene name) with significant changes in abundance ($p < 0.05$ adjusted probability) were identified and ranked by fold change (pigmented/non-pigmented) hair. Adjusted P-value [9] or significance [10] is indicated. Proteins were processed by label-free quantitation using either MaxQuant (iBAQ) and Scaffold (A) or PEAKs™ (B). Proteins were filtered based on a 5% FDR (Adjusted P-Value < 0.05).

A

Protein	UniProt ID#	Gene Name	logFC	Adj. P Value
Cathepsin B	CATB_HUMAN	CTSB	2.70	0.001
Protein transport protein Sec23B	SC23B_HUMAN	SEC23B	1.90	0.043
Phospholipase D3	PLD3_HUMAN	PLD3	1.74	0.024
Enoyl-CoA hydratase, mt	ECHM_HUMAN	ECHS1	1.53	0.043
Glyceraldehyde-3-P dH	G3P_HUMAN	GAPDH	−0.32	0.024
Elongation factor 2	EF2_HUMAN	EEF2	−0.37	0.043
Fatty acid synthase	FASN_HUMAN	FASN	−0.74	0.043
40S ribosomal protein S20	RS20_HUMAN	RPS20	−1.69	0.043

B

Protein	UniProt ID#	Gene Name	FC	Significance
Transmembrane glycoprotein NMB	GPNMB_HUMAN	GPNMB	5.17	200
Phospholipase D3	PLD3_HUMAN	PLD3	2.85	151.76
Cathepsin B	CATB_HUMAN	CTSB	2.45	48.95
Apolipoprotein D	APOD_HUMAN	APOD	2.36	33.92
Aldolase A	ALDOA_HUMAN	ALDOA	0.38	35.73
Histone H2A/2-C	H2A2A_HUMAN	HIST2H2AC	0.36	117.83
Trichohyalin	TRHY_HUMAN	TCHH	0.29	200
Keratin type I cuticular Ha7	KRT37_HUMAN	KRT37	0.27	200
Keratin-associated protein 4–9	KRA49_HUMAN	KRTAP4–9	0.21	44.55

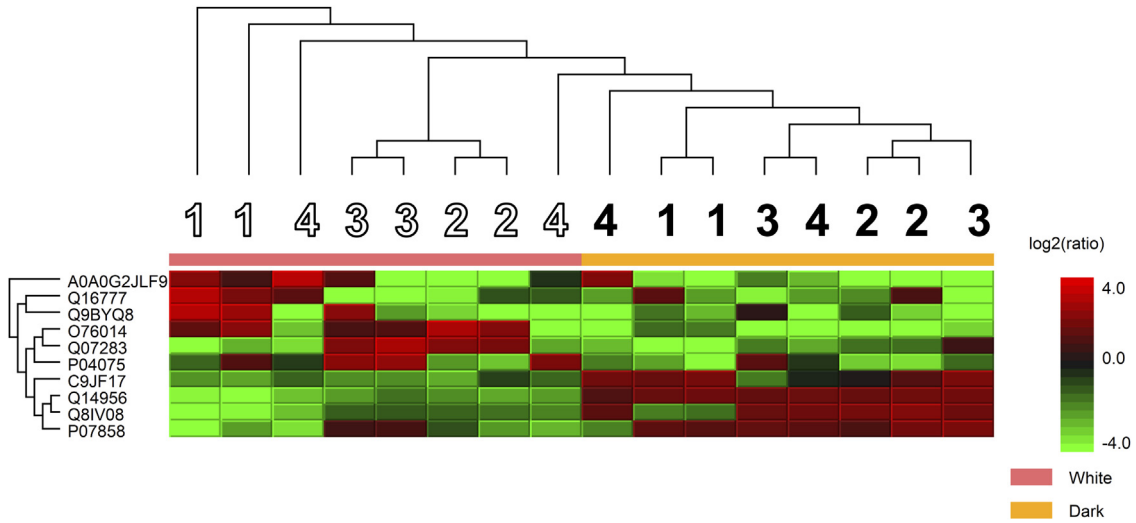

Fig. 2. Heat Map of Protein Levels in Matched Pigmented and Non-Pigmented Hair. Matching pigmented (Dark, bold numbers) and non-pigmented (White, open numbers) hair shafts from four subjects (1–4) were processed and protein levels quantified using PEAKs™ software (version 10.0). Proteins (Uniprot accession #) that had an average of at least a 2-fold change in quantitation and adjusted FDR-based P-values of less than 0.05 were filtered. Both protein levels and individual profiles were clustered based on correlated abundance levels.

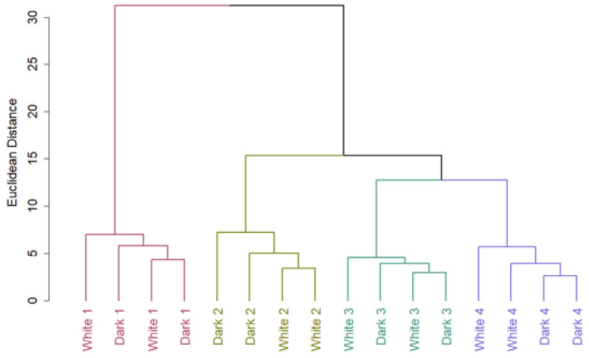

Fig. 3. Hierarchical Cluster Dendrogram of Genetically Variant Peptide Profiles. Profiles of inferred genotypes from individual datasets were extracted from 4 individuals (1–4) with matching pigmented (Dark) and non-pigmented (White) duplicate hair samples. Euclidean distance was measured and plotted as a hierarchical dendrogram.

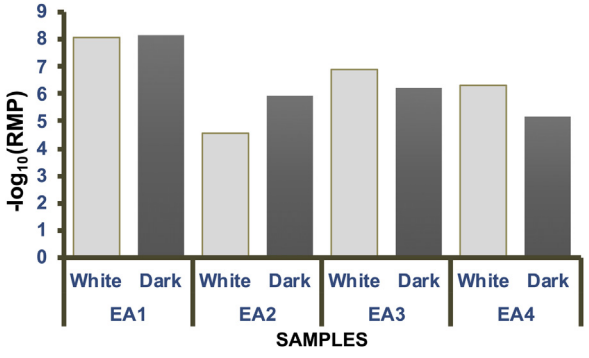

Fig. 4. Estimation of Random Match Probabilities Using Cumulative Profiles of Detected Genetically Variant Peptides. Cumulative profiles of genetically variant peptides from each biological condition (four European subjects, EA1 to EA4, and two conditions, non-pigmented and pigmented) were used to estimate the negative logarithm of random match probabilities ($-\log_{10}$(RMP)).

1 in 3.5×10^4 to 1 in 1.2×10^8 (Fig. 4). These values correlated with detected numbers of genetically variant peptides that ranged from 59 to 78 peptides (Supplemental Fig. 2). The random match probabilities were similar for matching pigmented and non-pigmented samples and varied by an average of 0.80 ± 0.57 orders of magnitude with no consistent trends with pigmentation status. This was consistent with the above finding that profiles of genetically variant peptides were not biased by the pigmentation status of the hair shaft.

4. Conclusion

Significant changes occur at the protein level when hair bulbs change phenotype and produce non-pigmented hair. However, in this cohort of four European subjects, these changes did not systematically introduce bias into the profile of the inferred SNP genotypes derived from genetically variant peptides. Therefore, the pigmentation status of a hair should not influence the resulting proteomic genotyping analyses used in a forensic investigation.

Role of the funding source

U.S. DOJ Award 2015-DN-BX-K065 supported purchase of mass spectrometry reagents, mass spectrometry data acquisition, and salary support of ZCG. National Institutes of Health, through grant number UL1 TR001860 supported the salary of BDJ.

CRediT authorship contribution statement

Rachel N. Franklin: Formal analysis, Investigation, Writing - review & editing, Visualization. **Noreen Karim:** Formal analysis, Data curation, Writing - review & editing, Visualization. **Zachary C. Goecker:** Methodology, Writing - review & editing. **Blythe P. Durbin-Johnson:** Formal analysis, Visualization. **Robert H. Rice:** Conceptualization, Methodology, Resources, Writing - review & editing, Supervision, Project administration, Funding acquisition. **Glendon J. Parker:** Conceptualization, Methodology, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

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 EP11759843.3).

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). BDJ was supported by the National Center for Advancing Translational Sciences, National Institutes of Health, through grant number UL1 TR001860.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at https://doi.org/10.1016/j.forsciint.2020.110200.

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), doi:http://dx.doi.org/10.1371/journal.pone.0160653 e0160653.

[2] S.A. Sterling, K.E. Mason, D.S. Anex, G.J. Parker, B. Hart, M. Prinz, Combined DNA typing and protein identification from unfired brass cartridges, J. Forensic Sci. (2019), doi:http://dx.doi.org/10.1111/1556-4029.14042 in press.

[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. Genetics 41 (2019) 19–23, doi:http://dx.doi.org/10.1016/j.fsigen.2019.03.009.

[4] K.E. Mason, D. Anex, T. Grey, B. Hart, G. Parker, Protein-based forensic identification using genetically variant peptides in human bone, Forensic Sci. Int. 288 (2018) 89–96.

[5] T. Borja, N. Karim, Z. Goecker, M. Salemi, B. Phinney, M. Naeem, R. Rice, G. Parker, Proteomic genotyping of fingermark donors with genetically variant peptides, Forensic Sci. Int. Genetics 42 (2019) 21–30, doi:http://dx.doi.org/10.1016/j.fsigen.2019.05.005.

[6] 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.

[7] P.W. Wu, K.E. Mason, B.P. Durbin-Johnson, M. Salemi, B.S. Phinney, D.M. Rocke, G.J. Parker, R.H. Rice, Proteomic analysis of hair shafts from monozygotic twins: expression profiles and genetically variant peptides, Proteomics 17 (13–14) (2017), doi:http://dx.doi.org/10.1002/pmic.201600462.

[8] M.E. Ritchie, B. Phipson, D. Wu, Y. Hu, C.W. Law, W. Shi, G.K. Smyth, limma powers differential expression analyses for RNA-sequencing and microarray studies, Nucleic Acids Res. 43 (7) (2015), doi:http://dx.doi.org/10.1093/nar/gkv007 e47.

[9] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc. Ser. B 57 (1) (1995) 289–300.

[10] J.D. Storey, R. Tibshirani, Statistical significance for genomewide studies, Proc. Natl. Acad. Sci. U. S. A. 100 (16) (2003) 9440–9445, doi:http://dx.doi.org/10.1073/pnas.1530509100.

[11] Y. Perez-Riverol, A. Csordas, J. Bai, M. Bernal-Llinares, S. Hewapathirana, D.J. Kundu, A. Inuganti, J. Griss, G. Mayer, M. Eisenacher, E. Perez, J. Uszkoreit, J. Pfeuffer, T. Sachsenberg, S. Yilmaz, S. Tiwary, J. Cox, E. Audain, M. Walzer, A.F. Jarnuczak, T. Ternent, A. Brazma, J.A. Vizcaino, The PRIDE database and related tools and resources in 2019: improving support for quantification data, Nucleic Acids Res. 47 (D1) (2019) D442–D450, doi:http://dx.doi.org/10.1093/nar/gky1106.

[12] S. Diment, M. Eidelman, G.M. Rodriguez, S.J. Orlow, Lysosomal hydrolases are present in melanosomes and are elevated in melanizing cells, J. Biol. Chem. 270 (9) (1995) 4213–4215, doi:http://dx.doi.org/10.1074/jbc.270.9.4213.

[13] B.L. Tang, J. Kausalya, D.Y. Low, M.L. Lock, W. Hong, A family of mammalian proteins homologous to yeast Sec24p, Biochem. Biophys. Res. Commun. 258 (3) (1999) 679–684, doi:http://dx.doi.org/10.1006/bbrc.1999.0574.

[14] V. Basrur, F. Yang, T. Kushimoto, Y. Higashimoto, K. Yasumoto, J. Valencia, J. Muller, W.D. Vieira, H. Watabe, J. Shabanowitz, V.J. Hearing, D.F. Hunt, E. Appella, Proteomic analysis of early melanosomes: identification of novel melanosomal proteins, J. Proteome Res. 2 (1) (2003) 69–79.

[15] A. Chi, J.C. Valencia, Z.Z. Hu, H. Watabe, H. Yamaguchi, N.J. Mangini, H. Huang, V.A. Canfield, K.C. Cheng, F. Yang, R. Abe, S. Yamagishi, J. Shabanowitz, V.J. Hearing, C. Wu, E. Appella, D.F. Hunt, Proteomic and bioinformatic characterization of the biogenesis and function of melanosomes, J. Proteome Res. 5 (11) (2006) 3135–3144, doi:http://dx.doi.org/10.1021/pr060363j.