ANALYTICAL STRATEGIES TO INVESTIGATE THE HAIR PROTEOME FOR HUMAN IDENTIFICATION

By

Fanny Chu

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Chemistry—Doctor of Philosophy

2020

ABSTRACT

ANALYTICAL STRATEGIES TO INVESTIGATE THE HAIR PROTEOME FOR HUMAN IDENTIFICATION

By

Fanny Chu

Analysis of genetic variation in DNA sequences serves as a powerful method for human identification owing to its exceptional discriminative power for distinguishing individuals. In cases where DNA is compromised in recovered forensic evidence, other approaches are needed to achieve a similar level of differentiative potential for human identification. Proteins offer a promising alternative, particularly in recovered hair evidence where minimal intact genomic DNA remains, as hair proteins often persist for long periods of time and their amino acid sequences derive from DNA. Detection of amino acid polymorphisms in hair proteins as genetically variant peptides (GVPs) permits inference of individualizing single nucleotide polymorphisms for identification. Expanding upon previous proof-of-concept work, this research interrogates the human hair proteome to address fundamental questions about how experimental variables affect GVP detection success rates, and to bridge the gap between laboratory-optimized studies and application of this protein-based approach in routine forensic analysis. Effects of intrinsic variation to hair protein chemistry, including differences among body locations and with increasing hair age, and external exposures, such as an explosive blast, on variant peptide marker detection are investigated using trypsin digestion and liquid chromatography-tandem mass spectrometry (LC-MS/MS). This multi-disciplinary work demonstrates success in GVP detection and advances in knowledge of protein chemistry in hair as a function of different body locations, in aged hairs, and in damaged hairs recovered after an explosive blast, providing greater

confidence in GVP analysis for forensic investigations. Not limited to forensic proteomics, these findings may be applicable to the wider bioanalytical sciences, including the medical, material, and agricultural sciences.

ACKNOWLEDGMENTS

I would first like to extend my thanks to my advisor, Professor A. Daniel Jones, who has been instrumental in my growth during my graduate school career, by challenging me to think critically, and unwavering in his support of me in navigating the M.S./Ph.D. dual degree. Many thanks to Dr. Deon S. Anex, my mentor at Lawrence Livermore National Laboratory, for his guidance in shaping my thesis work throughout my time at LLNL and his service as a thesis committee member. I also thank Professors Kevin Walker and Gary Blanchard for their service as thesis committee members.

Numerous individuals have supported me in a variety of ways throughout my graduate school career. From LLNL, M. Frank for co-mentorship, B. R. Hart and A. M. Williams for support of my work, the following scientists for assistance in the conduct of my experiments: K. E. Mason, P. H. Paul, Z. Dai, S. A. Malfatti, M. G. Lyman, T. M. Alfaro, B. Rubinfeld, and C. L. Strout, whose specific contributions are listed in the chapter forewords, and scientists at the Forensic Science Center for helping me acclimate to the lab. Additionally, special thanks to A. Alcaraz for his service in reviewing my work and associated documents for approved release. Individuals from MSU who supported me early on include R. W. Smith and V. L. McGuffin, who provided me with the foundations to navigate graduate school, and members of the Jones Lab and Forensic Chemistry group for their camaraderie, notably C. J. K. Tran, J. W. McIlroy, and T. E. Curtis. In particular, I thank K. L. Reese for productive discussions and the plethora of material in both editions of the C. A. K. R. & F. C. My thanks to family and friends for continued encouragement.

I acknowledge financial support from LLNL's Graduate Research Scholar Program (formerly the Livermore Graduate Scholar Program), for providing an exceptional opportunity for me to complete my graduate work at LLNL, and the Laboratory Directed Research and Development program (16-SI-002), and from the MSU Chemistry Department. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

PREFACE

Disclaimer. The Lawrence Livermore National Laboratory, Office of Scientific and Technical Information, Information Management (IM) number is: LLNL-TH-808258. This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: Introduction

Foreword

The material in Sections 1.2 – 1.4 has been previously presented in a published paper[1] and book chapter[2].

1.1 Conventional Methods for Human Identification

Human identification relies heavily on genomic DNA information, particularly variation in its nucleotide sequence, to serve many purposes, including mass casualty and missing persons situations, crime scene analysis, and study of human evolution, migration, and population genetics. However, in some cases, genomic DNA may not be intact or available in specific specimens.

The main advantage of genomic information, predominantly nuclear DNA, lies in its unparalleled discriminative power. Spanning more than 3 billion base pairs long across 23 chromosomes[3], genomic DNA encompasses 6 billion base pairs when both sets of chromosomes are considered for the diploid genome that is present in somatic cells. A typical human genome is estimated to differ from a reference sequenced genome by 0.1%, about 4.1 – 5.0 million sites[4], so the potential number of combinations of sequence variants is massive. Some of these differences are inherited across generations, yet additional variation is introduced by mutations. Given this enormous genetic variation between human genomes, analysis of variation in DNA sequences achieves high specificity, and phenomenal power to differentiate individuals.

As a conventional, yet powerful, method in forensic analyses, short tandem repeat (STR) profiling originally measured the number of repeated nucleotide units at 13 chromosomal positions, or loci, to yield random match probabilities on the order of 10^{-15}, or statistically, the chances of encountering the same profile in a population is 1 in a quadrillion[5]. Recent transition

to a panel of 20 STR loci further improves discriminative power[6]. The technique exploits loci containing repeated nucleotide units, of which the number of repeat units at each locus varies among individuals. The units are easily amplified by polymerase chain reaction (PCR) and are detected by capillary electrophoresis to determine their length variation[7-9]. For example, at locus CSF1PO in chromosome 5, the four-nucleotide unit TAGA repeats between 5 and 16 times[5, 9]. An example genotype that can be observed at this locus includes the unit repeated 7 and 10 times from the two copies of chromosome 5, respectively; detection of this unit, or STR marker, among others, which comprise an STR profile, serves to distinguish individuals. Further knowledge of genotype frequencies for each marker within a population allows quantification of the discriminative power of assembled STR profiles.

Another approach that utilizes genomic information for human identification involves analysis of single nucleotide polymorphisms (SNPs) via DNA sequencing, which offers advantages relative to STR profiling. SNPs make up > 99.9% of variation in the human genome and offer a more robust alternative to STRs due to a slower genomic mutation rate, estimated at $10^{-8} – 10^{-9}$ per base pair per year in contrast with a range of $10^{-3} – 10^{-5}$ per locus per year for STRs[10, 11]. STRs are more prone to mutation as their repeat units may be inserted or deleted from slippage errors in DNA-polymerase replication events over generations[12, 13], resulting in higher mutation rates than those observed for SNPs.

Applied to forensic analyses, SNPs can be used to differentiate individuals in a similar manner to STRs, though the nucleotide base varies instead of the number of repeat nucleotide units. To differentiate the variants at each locus, the nucleotide variant that occurs more frequently in a population is designated the major variant, which is often the reference allele, while minor variant refers to the less common nucleotide base, often the alternate allele, as

applicable to the most fundamental case: bi-allelic SNPs with only two observed nucleotide variants. At a SNP locus with two variants, 3 genotypes are possible; a heterozygous genotype indicates two different alleles, i.e., the major and minor variants, while homozygous genotypes reflect either two major variants or two minor variants. Identifications of these SNP genotypes at specific loci enable distinction of individuals, and associated genotype frequencies amassed for a large population and curated in public databases permit quantification of discriminative power.

SNP analysis attains a similar level of discrimination compared to STRs, though it requires a larger number of SNP loci, as corresponding allele frequencies vary drastically between 0 and 1 among populations, which affects random match probabilities[10]. The number of SNP markers needed to reach a certain level of discriminative power depends on the allele frequencies associated with the candidate markers. In contrast, genotype frequencies for STRs are low and similar among most populations, and as such, it is much easier to obtain low random match probabilities ($10^{-10} – 10^{-13}$)[14, 15]. However, by selecting SNP markers with similar, low allele frequencies among the variants, discriminative power similar to that for STR profiling using as few SNP markers as possible can be attained[10]. Utility of SNP detection also extends to other disciplines, such as human health. A number of these markers are correlated with diseases states[16, 17]; approximately 10,000 SNPs have been associated with human diseases[18], and thus, identifications of SNPs as biomarkers of disease states have received growing interest in clinical settings. For human identification purposes, most relevant to forensics, statistical probabilities of less than 10^{-15} have been achieved with a panel of 45 SNPs[15].

Their exceptional discriminative power notwithstanding, both approaches depend on the availability of nuclear DNA. Owing to the ubiquitous occurrence of nucleases in the environment[19, 20], DNA degrades with continuous external exposure[21]. In instances where nuclear

DNA no longer remains intact, or where DNA is not present, comparable discriminative power cannot be attained from analysis of other evidence types, thus demanding a more persistent specimen from which high levels of human differentiation can be achieved.

1.2 Hair as Forensic Evidence and Limitations of Conventional Analyses

Human hair, as one of the few biological specimen types that persist for long periods of time, is invaluable to forensic and archaeological investigations, yet limited identification information has often been obtained from hair using conventional approaches. Comprised primarily of keratins and keratin-associated proteins, hair exhibits high durability that contributes to its persistence. Packed into coiled-coils and localized to the cortex, cuticle, or medulla of the hair shaft[22, 23], hair keratins are stabilized and provide tensile strength via crosslinking between proteins by cystine disulfide bonds[24]. Differences in their amino acid composition further separate them into type I acidic keratins (K31, K32, K33A, K33B, K34 – K40) with low-sulfur content and type II neutral to basic keratins (K81 – K86) with high-sulfur content[23-27]. Keratin-associated proteins (KAPs), categorized by their amino acid composition as high-sulfur, ultra-high sulfur, or high glycine-tyrosine proteins, participate in crosslinking keratins to provide rigidity to the matrix[28, 29].

In contrast to its high protein content, hair contains minimal intact nuclear DNA, owing to the keratinization process and resulting from continued exposure to environments rich with nucleases. Found within keratinized corneocytes, or keratin-rich cells that have lost their nuclei and cellular organelles, nucleases including DNase1L2 degrade DNA to tiny fragments, depending on the abundances and catalytic activities of the nucleases[30]. DNA degradation begins immediately during hair formation as the keratinocyte, a keratin-producing cell, moves out of the hair follicle and terminal differentiation occurs, the process by which hair cortex, cuticle, and

medulla are formed[8]. It has also been found that nuclear DNA degradation rates in scalp hair exhibit high interindividual variation, with some having sufficient amounts for profiling and others whose nuclear DNA content is depleted[30]. With continued hair growth, DNA degrades to low and variable amounts, making forensic nuclear DNA profiling in hair unreliable.

Due to its minimal intact nuclear DNA content, hair evidence has been largely limited to analysis via comparative microscopy, or of sequence variation in the hypervariable regions within mitochondrial DNA (mtDNA), which is often protected from exposure to nucleases by sequestration in mitochondria. Although utilized when demonstrated to persist longer than nuclear DNA[31], exclusive inheritance of mtDNA through the maternal line limits the discriminative power that can be achieved with its analysis[32]. With comparative microscopy, often light microscopy, visual comparison of hair fibers permits differentiation of species[33], but claimed specificity in association to individuals using this technique alone has been criticized[34] owing to a lack of quantitative metrics in forensic analyses. When nuclear DNA, the gold standard, is compromised and techniques such as comparative microscopy and mtDNA analysis produce insufficient discriminative power, an alternative method for human identification becomes vital.

1.3 Protein-Based Human Identification

Protein-containing material offers an attractive specimen type for identification, as many proteins are robust and their amino acid sequences derive from DNA. Protein-based identification methods are largely supported by advancements in proteomics, such as faster-scanning high-resolution mass analyzers and bioinformatics tools. The proteome has emerged as a focus for biomarker indicators of various disease states and also for analysis of ancient specimens; proteins are often preserved where minimal intact DNA remains. High-resolution

mass spectrometry analysis of fossil bones revealed survival of proteins in the leucine-rich repeat family and serum proteins in samples dating back to 900 thousand years[35]. Peptide markers characterized in various keratinous tissues of agricultural and archaeological importance, including horn and baleen, enabled species identification within the mammalian kingdom based on variant amino acid sequences that distinguish genera[36]. Proteomics methods for biological fluid identification have also identified peptide markers specific to fluids including human saliva, urine, seminal fluid, and vaginal fluid, demonstrating the potential of protein detection and identification for utility in forensic investigations[37, 38]. While the hair proteome has been probed to characterize disorders including lamellar ichthyosis and trichothiodystrophy using protein expression differences[29, 39, 40], hair proteins have received limited attention for human identification.

The few reports of distinguishing individuals from hair proteomes include a 2014 paper by Laatsch and colleagues who analyzed protein composition and expression differences in hair from different ethnic populations and from various body locations and found specific protein expression differences to differentiate based upon ethnicity and body location[41]. Following in this vein, Wu et al. showed in 2017 that protein profiles differentiated monozygotic twins from unrelated individuals[42]. However, both studies focused on protein abundances to differentiate populations, which does not provide sufficient specificity to distinguish individuals because hair protein levels are evolutionarily conserved within ethnic populations.

Alternatively, genetically variant peptides (GVPs) in hair proteins possess great potential to differentiate individuals. Identification of single amino acid polymorphisms in GVPs permits inference of individualizing SNPs. Parker et al. first demonstrated the identification and use of single amino acid polymorphisms in GVPs from head hair to differentiate individuals[43]. Figure

1.1 illustrates the amino acid consequence of a SNP and the resulting GVP in the protein K83, gene name KRT83, as an example. The missense SNP, a type of nonsynonymous point mutation that alters the resultant three-nucleotide codon and changes the coding for a specific amino acid in the corresponding protein, occurs at position chr12:52319304 on chromosome 12 as mapped according to Genome Reference Consortium Human Build 38 (GRCh38) coordinates, and involves a G→A mutation at the nucleotide level. Though both alleles exist within the population, they can be distinguished based on their prevalence, or allele frequency, as described in Section 1.1; in this example, the nucleotide base G (guanine) is the reference allele and the major variant, and A (adenine) is the alternate allele and the minor variant. Conserved at the protein level, the change in codon from this SNP manifests as the single amino acid polymorphism (SAP) R149C, an Arg→Cys mutation at position 149 within K83. Using a bottom-up, or shotgun, proteomics approach in which proteins are enzymatically digested into peptides for detection via mass spectrometry, peptides carrying SAPs are termed genetically variant peptides (GVPs). To distinguish the two forms of the variant peptide, the terms major and minor GVP, respectively, are used, parallel to the major (more common) and minor (less common) variants of a SNP. Of note, while there are many more synonymous compared to nonsynonymous SNPs, the codon change for mutations in the former category does not change the amino acid sequence, and as such, synonymous SNPs are not useful for protein-based human identification. Therefore, all SNPs described herein refer to nonsynonymous, missense variants.

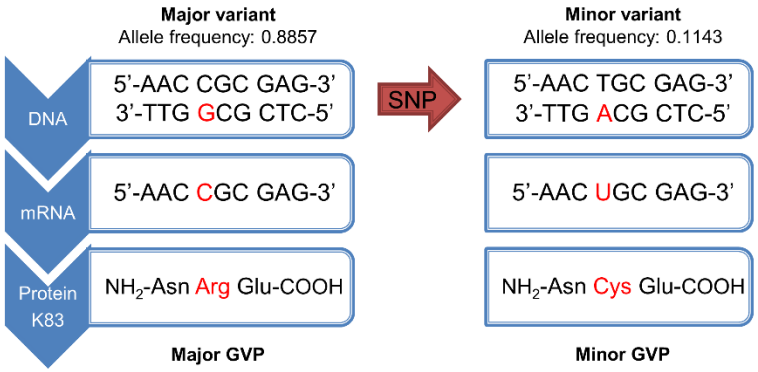


Figure 1.1. Example of conservation of a nonsynonymous, missense single nucleotide polymorphism (SNP) in chromosome 12 to an amino acid polymorphism within the protein K83. The text in red denotes the location of the mutation, from DNA to protein. The major variant is one whose allele is more prevalent in the population, with a higher allele frequency. In the resultant proteins, peptides carrying the amino acid polymorphism, or genetically variant peptides (GVPs), follow a similar designation to distinguish the two forms of the mutation present in the population.

Detection of one or more forms of the amino acid polymorphism in GVPs permits inference of SNP genotype. For instance, the presence of both the major and minor peptide variants in a sample implies a heterozygous genotype for the corresponding SNP, associated with a known genotype frequency at the locus from population frequencies curated in public databases. With accumulation of GVPs at the various SNP loci in a sample, e.g., hair fiber from an individual, the corresponding genotype frequencies can be used to not only build a profile that enables distinction of the individual within a population, but discriminative power can also be quantified, such as via random match probability, a common metric used in forensic science to evaluate DNA evidence that is described in Chapter 2. Analogous to a SNP panel developed by Pakstis et al.[15], Parker and co-workers compiled a panel of 33 SNPs identified from GVP markers detected in bulk quantities of scalp hair and verified by Sanger sequencing, and determined random match probabilities ranging up to 1 in 14,000 for a cohort of 60 subjects[43]. In this manner, the protein-based human identification approach enables differentiation of individuals.

1.4 Exome Sequence-Driven Approach to Protein-Based Human Identification

The limited number of GVP identifications presented in the initial proof-of-concept study has highlighted a need for an alternative approach to GVP discovery, such as with an exome sequence-driven process, which targets only SNPs in protein-coding regions within the DNA sequence (the exome) and strikes a balance between targeted and untargeted GVP identification. To investigate GVP markers for inclusion into a protein-based human identification panel and to improve discriminative power for hair evidence, as measured via RMPs, it is critical to maximize GVP discoveries in hair proteomes, but also equally important to detect GVPs without overtaxing computational resources. Initial demonstration of human identification with GVPs utilized a custom in-house protein sequence database for matching fragment ion masses predicted *in silico* for peptides derived from database proteins to MS/MS fragment ion masses from experimentally observed peptides[43]. However, this custom database primarily included keratins and a few intracellular proteins known to be present in hair, and thus, restricted GVP discovery to only a small number of proteins: 33 SNPs from 21 proteins. Identification of a small subset of SNPs limits discriminative power, and as such, GVP discovery needs to be widened to include those from a larger group of hair proteins, with a less targeted approach.

Although a completely untargeted GVP detection approach is expected to improve GVP identification yields, creation of a database to include all known SNPs, which is not a trivial task, and searching with this massive database introduce problems of overtaxing computational resources. Deviating from conventional approaches in proteomics, identification of peptides with SAPs necessitates matching MS/MS spectra to databases that contain protein sequences with amino acid mutations as opposed to canonical protein sequences found in curated databases (e.g., UniProt KnowledgeBase SwissProt database[44]). To include SAPs derived from SNPs detected in

the human population for GVP analysis in reference databases, reference protein sequences must be mutated to reflect SNPs. One method for generating relevant mutated protein sequences is to use known SNPs from databases such as the Single Nucleotide Polymorphism Database (dbSNP) (National Center for Biotechnology Information, U.S. National Library of Medicine)[45]. Roughly 12 million SNPs have been annotated, and though a large fraction of them do not result in SAPs (i.e., SNPs in non-coding regions and synonymous SNPs, both of which do not effect amino acid mutations, are annotated as well), exhaustive database searching to match tandem mass spectra with predicted peptides from all proteins affected by nonsynonymous SNPs for GVP identification becomes computationally expensive. To add to the computational challenges of mutated database creation, other mutations to DNA sequences that change the amino acid sequence of proteins (e.g., insertions, deletions, and proximal nonsynonymous SNPs acting on the same codon) need to be included in mutated protein sequences, as their exclusion may produce incorrect protein sequences and result in failure to detect GVPs. However, feasible combinations of these other mutations and nonsynonymous SNPs that are empirically observed in individuals' genotypes cannot be inferred directly from curated SNPs, which are presented as single, separate entries. Without knowledge of biologically observed variant combinations, inclusion of these other mutations requires generating permutations of mutated protein sequences, many of which may not be biologically relevant. This, in turn, would exponentially increase the number of entries in the database and tax computational resources during GVP identification. Furthermore, not all SNPs may be expressed in each individual and some may occur so rarely within a population that they would not be useful in linking an unknown profile to an individual within a subpopulation compared to more commonly occurring SNPs. As such, an entirely untargeted GVP discovery approach created from all curated SNPs also does not

represent an optimal method. The GVP search space needs to be simplified from an untargeted approach to reduce the amount of computational resources in creation of and searching with a more biologically relevant mutated protein database.

Instead, a focused search strategy using an individualized mutated protein database from exome sequence information balances GVP identification by including a larger pool of proteins but also by limiting the search to be more biologically relevant, and thus, more computationally economical. More specifically, this approach eliminates searching of SNPs not expressed within the individual, which conserves computational resources, while maximizing GVP identification to include SNPs from a larger group of proteins and SNPs that may not be curated in public databases. An individual's exome sequence only includes information from exons, protein-coding regions in genes[46], and can be obtained via high-throughput DNA sequencing, such as next-generation sequencing methods[47]. Further, searching with more accurately predicted mutated protein sequences, based on empirically-observed combinations of mutations that change amino acid sequences from exome sequence information, increases the chances of successful GVP detection in a computationally efficient manner. In essence, having prior knowledge of SNPs carried by individuals whose hair samples are analyzed guides GVP identification so that variant peptides detected by mass spectrometry reflect an individual's SNP genotypes. Such *a priori* knowledge supports evaluation of discriminating outcomes when the true result is known.

Transition to an exome sequence-driven process has substantially improved GVP discovery from hair protein digests. From an individual-matched DNA sample, exome sequencing enables detection of all SNPs relevant to the individual, which are then filtered for expression in genes of interest to the type of evidence, such as limiting to genes that produce

proteins found only in hair shaft, and are ultimately annotated in protein sequences within an individualized mutated database alongside their unmutated sequences[48]. Hair genes of interest, 691 in total, were selected from commonly identified hair proteins, as described in Mason et al.[48] Applied to single one-inch scalp hairs, individualized mutated databases derived from exome sequence information guided GVP identification from LC-MS/MS proteome data to provide identifications of an additional 20 SNPs from GVPs in 16 keratin-associated proteins (KAPs)[48]. This class of highly abundant hair structural proteins was not included in Parker et al. as a source of GVPs[43]. Inclusion of GVPs from this protein class with other intracellular hair proteins yielded RMPs up to 1 in 167,000,000 with a single inch of hair[48], which represents approximately 12,000-fold improvement in discriminative power with 100-fold less material relative to an RMP of 1 in 14,000 achieved with bulk amounts of scalp hair as described in Parker et al.[43]

While exome sequence-driven approaches present a number of advantages for GVP identification during this phase of GVP discovery and development of GVP analysis, when candidate markers are being evaluated for inclusion into a panel for protein-based human identification, it is expected that in practice, exome sequence information will neither be necessary nor expected to be recovered from forensic evidence where DNA quality is most likely compromised. As such, integration of exome sequence information and mass spectrometry proteome data serves primarily as a research and development tool. Following development, selection, and validation of GVP markers for a human identification panel, it is anticipated that GVP analysis will be performed in a targeted manner, via comparison of the presence or absence of panel GVPs from both recovered evidence and those from suspects or individuals in a database to determine the extent of profile matching.

1.5 Broad and Specific Aims

This research expands upon previous proof-of-concept work with detection of genetically variant peptides in hair to address fundamental questions about variables that may affect GVP detection success rates and to bridge the gap between laboratory-optimized studies and application of this approach in forensic analysis. In particular, the effects of intrinsic variation to hair protein chemistry and external exposures on variant peptide marker detection are investigated, which have not yet been explored. For example, can the same GVP information be found in a single inch of hair? What are best practices for working with mass-limited hair samples to maximize GVP detection? Can the same GVP information be extracted from single hairs from different body locations? How about in aged hairs that have encountered external damage since their formation? And what effect does exposure of single hairs to harsh environmental conditions have on GVP identification success rates? Understanding how the hair proteome changes in response to these variables and consequences for GVP detection are vital to developing this protein-based approach for routine operation in forensic analysis. Chapter 2 builds upon the findings in Mason et al.[48] to establish a framework for GVP identification from an optimized single hair analysis and integration of untargeted mass spectrometry and exome sequence analyses. Leveraging this workflow, Chapters 3 and 4 examine changes to hair protein chemistry with body location and hair age, respectively, and subsequent effects on GVP identification. Chapter 5 then characterizes damaged hairs that have been recovered after an explosive blast by both microscopy and untargeted mass spectrometry, representing an external exposure to harsh conditions at the extreme. Finally, Chapter 6 comments on the pathway forward for protein-based human identification and the broader implications beyond

advancement of forensic proteomics. Not limited to forensic science, results from this multi-disciplinary work may find application in the medical, agricultural, and material sciences.

REFERENCES

REFERENCES

1. Chu, F.; Mason, K. E.; Anex, D. S.; Jones, A. D.; Hart, B. R., Hair Proteome Variation at Different Body Locations on Genetically Variant Peptide Detection for Protein-Based Human Identification. *Scientific Reports* **2019,** *9* (1), 7641.

2. Chu, F.; Mason, K. E.; Anex, D. S.; Paul, P. H.; Hart, B. R., Human Identification Using Genetically Variant Peptides in Biological Forensic Evidence. In *Applications in Forensic Proteomics: Protein Identification and Profiling*, American Chemical Society: 2019; Vol. 1339, pp 107-123.

3. Jain, M.; Koren, S.; Miga, K. H.; Quick, J.; Rand, A. C.; Sasani, T. A.; Tyson, J. R.; Beggs, A. D.; Dilthey, A. T.; Fiddes, I. T.; Malla, S.; Marriott, H.; Nieto, T.; O'Grady, J.; Olsen, H. E.; Pedersen, B. S.; Rhie, A.; Richardson, H.; Quinlan, A. R.; Snutch, T. P.; Tee, L.; Paten, B.; Phillippy, A. M.; Simpson, J. T.; Loman, N. J.; Loose, M., Nanopore Sequencing and Assembly of a Human Genome with Ultra-Long Reads. *Nature Biotechnology* **2018,** *36* (4), 338-345.

4. The 1000 Genomes Project Consortium, A Global Reference for Human Genetic Variation. *Nature* **2015,** *526*, 68-74.

5. Butler, J. M., Genetics and Genomics of Core Short Tandem Repeat Loci Used in Human Identity Testing. *Journal of Forensic Sciences* **2006,** *51* (2), 253-265.

6. Hares, D. R., Selection and Implementation of Expanded CODIS Core Loci in the United States. *Forensic Science International: Genetics* **2015,** *17*, 33-34.

7. Butler, J. M.; Buel, E.; Crivellente, F.; McCord, B. R., Forensic DNA Typing by Capillary Electrophoresis Using the ABI Prism 310 and 3100 Genetic Analyzers for STR Analysis. *Electrophoresis* **2004,** *25* (10-11), 1397-1412.

8. McNevin, D.; Wilson-Wilde, L.; Robertson, J.; Kyd, J.; Lennard, C., Short tandem repeat (STR) genotyping of keratinised hair Part 2. An optimised genomic DNA extraction procedure reveals donor dependence of STR profiles. *Forensic Science International* **2005,** *153* (2), 247-259.

9. Butler, J. M.; Hill, C. R., Biology and Genetics of New Autosomal STR Loci Useful for Forensic DNA Analysis. *Forensic Science Review* **2012,** *24* (1), 15-26.

10. Kidd, K. K.; Pakstis, A. J.; Speed, W. C.; Grigorenko, E. L.; Kajuna, S. L. B.; Karoma, N. J.; Kungulilo, S.; Kim, J.-J.; Lu, R.-B.; Odunsi, A.; Okonofua, F.; Parnas, J.; Schulz, L. O.; Zhukova, O. V.; Kidd, J. R., Developing a SNP Panel for Forensic Identification of Individuals. *Forensic Science International* **2006,** *164* (1), 20-32.

11. Balanovsky, O., Toward a Consensus on SNP and STR Mutation Rates on the Human Y-Chromosome. *Human Genetics* **2017,** *136* (5), 575-590.

12. Ellegren, H., Microsatellites: Simple Sequences with Complex Evolution. *Nature Reviews Genetics* **2004,** *5* (6), 435-445.

13. Willems, T.; Gymrek, M.; Poznik, G. D.; Tyler-Smith, C.; Erlich, Y., Population-Scale Sequencing Data Enable Precise Estimates of Y-STR Mutation Rates. *The American Journal of Human Genetics* **2016,** *98* (5), 919-933.

14. Kim, J.-J.; Han, B.-G.; Lee, H.-I.; Yoo, H.-W.; Lee, J.-K., Development of SNP-Based Human Identification System. *International Journal of Legal Medicine* **2010,** *124* (2), 125-131.

15. Pakstis, A. J.; Speed, W. C.; Fang, R.; Hyland, F. C. L.; Furtado, M. R.; Kidd, J. R.; Kidd, K. K., SNPs for a Universal Individual Identification Panel. *Human Genetics* **2010,** *127* (3), 315-324.

16. Hampe, J.; Franke, A.; Rosenstiel, P.; Till, A.; Teuber, M.; Huse, K.; Albrecht, M.; Mayr, G.; De La Vega, F. M.; Briggs, J.; Günther, S.; Prescott, N. J.; Onnie, C. M.; Häsler, R.; Sipos, B.; Fölsch, U. R.; Lengauer, T.; Platzer, M.; Mathew, C. G.; Krawczak, M.; Schreiber, S., A Genome-Wide Association Scan of Nonsynonymous SNPs Identifies a Susceptibility Variant for Crohn Disease in ATG16L1. *Nature Genetics* **2007,** *39* (2), 207-211.

17. Edwards, T. L.; Scott, W. K.; Almonte, C.; Burt, A.; Powell, E. H.; Beecham, G. W.; Wang, L.; Züchner, S.; Konidari, I.; Wang, G.; Singer, C.; Nahab, F.; Scott, B.; Stajich, J. M.; Pericak-Vance, M.; Haines, J.; Vance, J. M.; Martin, E. R., Genome-Wide Association Study Confirms SNPs in SNCA and the MAPT Region as Common Risk Factors for Parkinson Disease. *Annals of Human Genetics* **2010,** *74* (2), 97-109.

18. Buniello, A.; MacArthur, J. A. L.; Cerezo, M.; Harris, L. W.; Hayhurst, J.; Malangone, C.; McMahon, A.; Morales, J.; Mountjoy, E.; Sollis, E.; Suveges, D.; Vrousgou, O.; Whetzel, P. L.; Amode, R.; Guillen, J. A.; Riat, H. S.; Trevanion, S. J.; Hall, P.; Junkins, H.; Flicek, P.; Burdett, T.; Hindorff, L. A.; Cunningham, F.; Parkinson, H., The NHGRI-EBI GWAS Catalog of Published Genome-Wide Association Studies, Targeted Arrays and Summary Statistics 2019. *Nucleic Acids Research* **2019,** *47* (D1), D1005-D1012.

19. Greaves, M. P.; Wilson, M. J., The Degradation of Nucleic Acids and Montmorillonite-Nucleic-Acid Complexes by Soil Microorganisms. *Soil Biology and Biochemistry* **1970,** *2* (4), 257-268.

20. Ogram, A. V.; Mathot, M. L.; Harsh, J. B.; Boyle, J.; Pettigrew, C. A., Effects of DNA Polymer Length on Its Adsorption to Soils. *Applied and Environmental Microbiology* **1994,** *60* (2), 393-396.

21. Higgins, D.; Rohrlach, A. B.; Kaidonis, J.; Townsend, G.; Austin, J. J., Differential Nuclear and Mitochondrial DNA Preservation in Post-Mortem Teeth with Implications for Forensic and Ancient DNA Studies. *PloS ONE* **2015,** *10* (5), e0126935-e0126935.

22. Zhang, Y.; Alsop, R. J.; Soomro, A.; Yang, F.-C.; Rheinstädter, M. C., Effect of Shampoo, Conditioner and Permanent Waving on the Molecular Structure of Human Hair. *PeerJ* **2015,** *3*, e1296.

23. Schweizer, J.; Langbein, L.; Rogers, M. A.; Winter, H., Hair Follicle-Specific Keratins and Their Diseases. *Experimental Cell Research* **2007,** *313* (10), 2010-2020.

24. Wolfram, L. J., Human Hair: A Unique Physicochemical Composite. *Journal of the American Academy of Dermatology* **2003,** *48* (6, Supplement), S106-S114.

25. Langbein, L.; Rogers, M. A.; Winter, H.; Praetzel, S.; Beckhaus, U.; Rackwitz, H.-R.; Schweizer, J., The Catalog of Human Hair Keratins: I. Expression of the Nine Type I Members in the Hair Follicle. *Journal of Biological Chemistry* **1999,** *274* (28), 19874-19884.

26. Langbein, L.; Rogers, M. A.; Winter, H.; Praetzel, S.; Schweizer, J., The Catalog of Human Hair Keratins: II. Expression of the Six Type II Members in the Hair Follicle and the Combined Catalog of Human Type I and II Keratins. *Journal of Biological Chemistry* **2001,** *276* (37), 35123-35132.

27. Rouse, J. G.; Van Dyke, M. E., A Review of Keratin-Based Biomaterials for Biomedical Applications. *Materials* **2010,** *3* (2).

28. Shimomura, Y.; Ito, M., Human Hair Keratin-Associated Proteins. *Journal of Investigative Dermatology Symposium Proceedings* **2005,** *10* (3), 230-233.

29. Lee, Y. J.; Rice, R. H.; Lee, Y. M., Proteome Analysis of Human Hair Shaft: From Protein Identification to Posttranslational Modification. *Molecular & Cellular Proteomics* **2006,** *5* (5), 789-800.

30. Szabo, S.; Jaeger, K.; Fischer, H.; Tschachler, E.; Parson, W.; Eckhart, L., In Situ Labeling of DNA Reveals Interindividual Variation in Nuclear DNA Breakdown in Hair and May Be Useful to Predict Success of Forensic Genotyping of Hair. *International Journal of Legal Medicine* **2012,** *126* (1), 63-70.

31. Melton, T.; Dimick, G.; Higgins, B.; Lindstrom, L.; Nelson, K., Forensic Mitochondrial DNA Analysis of 691 Casework Hairs. *Journal of Forensic Sciences* **2005,** *50* (1), JFS2004230-8.

32. Tridico, S. R.; Murray, D. C.; Addison, J.; Kirkbride, K. P.; Bunce, M., Metagenomic Analyses of Bacteria on Human Hairs: A Qualitative Assessment for Applications in Forensic Science. *Investigative Genetics* **2014,** *5* (1), 16.

33. Edson, J.; Brooks, E. M.; McLaren, C.; Robertson, J.; McNevin, D.; Cooper, A.; Austin, J. J., A Quantitative Assessment of a Reliable Screening Technique for the STR Analysis of Telogen Hair Roots. *Forensic Science International: Genetics* **2013,** *7* (1), 180-188.

34. National Research Council, U. S. A., *Strengthening Forensic Science in the United States: A Path Forward*. 2009.

35. Wadsworth, C.; Buckley, M., Proteome Degradation in Fossils: Investigating the Longevity of Protein Survival in Ancient Bone. *Rapid Communications in Mass Spectrometry* **2014,** *28* (6), 605-615.

36. Solazzo, C.; Wadsley, M.; Dyer, J. M.; Clerens, S.; Collins, M. J.; Plowman, J., Characterisation of Novel α-Keratin Peptide Markers for Species Identification in Keratinous Tissues Using Mass Spectrometry. *Rapid Communications in Mass Spectrometry* **2013,** *27* (23), 2685-2698.

37. Legg, K. M.; Powell, R.; Reisdorph, N.; Reisdorph, R.; Danielson, P. B., Discovery of Highly Specific Protein Markers for the Identification of Biological Stains. *Electrophoresis* **2014,** *35* (21-22), 3069-3078.

38. Legg, K. M.; Powell, R.; Reisdorph, N.; Reisdorph, R.; Danielson, P. B., Verification of Protein Biomarker Specificity for the Identification of Biological Stains by Quadrupole Time-of-Flight Mass Spectrometry. *Electrophoresis* **2016,** *38* (6), 833-845.

39. Rice, R. H.; Wong, V. J.; Price, V. H.; Hohl, D.; Pinkerton, K. E., Cuticle Cell Defects in Lamellar Ichthyosis Hair and Anomalous Hair Shaft Syndromes Visualized After Detergent Extraction. *The Anatomical Record* **1996,** *246* (4), 433-441.

40. Rice, R. H., Proteomic Analysis of Hair Shaft and Nail Plate. *Journal of Cosmetic Science* **2011,** *62* (2), 229-236.

41. Laatsch, C. N.; Durbin-Johnson, B. P.; Rocke, D. M.; Mukwana, S.; Newland, A. B.; Flagler, M. J.; Davis, M. G.; Eigenheer, R. A.; Phinney, B. S.; Rice, R. H., Human Hair Shaft Proteomic Profiling: Individual Differences, Site Specificity and Cuticle Analysis. *PeerJ* **2014,** *2*, e506.

42. Wu, P.-W.; Mason, K. E.; Durbin-Johnson, B. P.; Salemi, M.; Phinney, B. S.; Rocke, D. M.; Parker, G. J.; Rice, R. H., Proteomic Analysis of Hair Shafts from Monozygotic Twins: Expression Profiles and Genetically Variant Peptides. *Proteomics* **2017,** *17* (13-14), 1600462.

43. Parker, G. J.; Leppert, T.; Anex, D. S.; Hilmer, J. K.; Matsunami, N.; Baird, L.; Stevens, J.; Parsawar, K.; Durbin-Johnson, B. P.; Rocke, D. M.; Nelson, C.; Fairbanks, D. J.; Wilson, A. S.; Rice, R. H.; Woodward, S. R.; Bothner, B.; Hart, B. R.; Leppert, M., Demonstration of Protein-Based Human Identification Using the Hair Shaft Proteome. *PLoS ONE* **2016,** *11* (9), e0160653.

44. Boutet, E.; Liberherr, D.; Tognolli, M.; Schneider, M.; Bairoch, A., UniProtKB/Swiss-Prot. *Methods in Molecular Biology* **2007,** *406*, 89-112.

45. Kitts, A.; Sherry, S., The Single Nucleotide Polymorphism Database (dbSNP) of Nucleotide Sequence Variation. In *The NCBI Handbook*, McEntyre, J.; Ostell, J., Eds. National Center for Biotechnology Information (US): Bethesda, MD, 2002 [Updated 2011].

46. Bamshad, M. J.; Ng, S. B.; Bigham, A. W.; Tabor, H. K.; Emond, M. J.; Nickerson, D. A.; Shendure, J., Exome Sequencing as a Tool for Mendelian Disease Gene Discovery. *Nature Reviews Genetics* **2011,** *12* (11), 745-755.

47. Rabbani, B.; Mahdieh, N.; Hosomichi, K.; Nakaoka, H.; Inoue, I., Next-Generation Sequencing: Impact of Exome Sequencing in Characterizing Mendelian Disorders. *Journal of Human Genetics* **2012,** *57* (10), 621-632.

48. Mason, K. E.; Paul, P. H.; Chu, F.; Anex, D. S.; Hart, B. R., Development of a Protein-Based Human Identification Capability from a Single Hair. *Journal of Forensic Sciences* **2019,** *0* (0).

CHAPTER 2: Method Development for Single Hair Analysis and Genetically Variant Peptide Identification

Foreword

Contributions from others to the conduct of the experiments described in this chapter are as follows, in no particular order: P. H. Paul provided single nucleotide variant lists and individualized mutated protein FASTA files, M. G. Lyman acquired the mass spectrometry data from bulk quantities of hair, and D. S. Anex acquired the mass spectrometry data from single one-inch hair fibers. Procedures described in Sections 2.3.2 – 2.3.4 have been previously presented in Chu et al.[1]

2.1 Overview and Specific Aims

Although analysis of bulk amounts of hair demonstrated detection of genetically variant peptide (GVP) markers for identification purposes[2], application of this technology for forensic analysis requires successful marker identification from a substantially reduced amount of hair, as hair evidence recovered from crime scenes can be severely limited. The need to identify GVP markers in sample-limited evidence types provides the main driver for single hair analysis, and more specifically, for analysis of a single inch (2.54 cm) of hair.

Previous work with bulk hair fiber amounts also detected GVP markers in an untargeted manner, relying on comparison of amino acid polymorphisms in mutated peptides with a list of all SNPs annotated in a database and subsequent validation through Sanger sequencing from donor-matched DNA samples[2]. This is a computationally expensive approach not only for GVP marker discovery, but also for creating a search database from a public SNP database containing all known variants. Instead, with exome sequence information from DNA gathered concurrently with an individual's hair sample, marker discovery can be much more focused and simplified, as

the search space includes only mutations known to be within an individual and computational resources are not expended on searching for mutations not found in individuals. Implementation of this alternative approach for single hairs requires development of an informatic process for comparing mass spectral data and exome information post-data acquisition.

This chapter presents two sections: first, an optimization of hair sample preparation to analyze single hairs and second, the development of a process for GVP marker identification from a combination of mass spectral data and exome sequence information. To that end, Section 2.2 discusses sample preparation processes that can be adapted from bulk amounts to single hair fibers while also identifying key parameters in peptide identification software that maximize peptide identification, as GVP identification depends on successful peptide identification from a hair sample. Furthermore, Section 2.3 compares single hair preparation performance from utilizing the optimized parameters in Section 2.2 and describes a pipeline from mass spectral data acquisition to GVP marker identification that enables examination of chemical variation and environmental exposures in the single hair proteome.

2.2 Optimization of Parameters for Peptide Identification

Considered one of the most nebulous aspects of the analytical process for proteomics experiments, peptide identification from mass spectral data relies on algorithms built into search engines, which in turn, depend upon user input among the numerous options for each given parameter. Further, parameters may vary among different search engines, especially when the inherent algorithms for peptide identification differ, which may lead to identification of different peptide populations by search engine. This is particularly of concern because peptide identifications form the crux of bottom-up, or shotgun, proteomics experiments, especially

untargeted experiments designed to identify markers of interest, such as in this work for identification of robust GVPs.

Peptide identification approaches can be broadly classified into categories of *de novo* sequencing[3, 4], database searching[5-7], and spectral library searching[8, 9]. Fundamentally, in bottom-up proteomics, all methods seek to match tandem mass spectra to peptides, i.e., forming a peptide-spectrum match. The main differences lie in the representation of the peptide and the scoring for a valid peptide-spectrum match. Briefly, *de novo* sequencing relies on predictable peptide fragmentation rules to annotate experimental mass spectra[4], based on matches between fragment ion masses observed and those predicted for an array of computer-generated sequences. Database searching is more commonly used and is also more computationally economical than *de novo* sequencing. This approach centers on lists of peptide masses and those of their fragments based on theoretical spectra generated from protein sequences in the database of choice (usually derived from sequences of genomic DNA) to represent peptides. These mass lists are then compared to lists of fragment ion masses derived from experimental spectra, and list similarities are then scored[5]. Finally, spectral library searching, not as widely used for untargeted proteomics, depends on matching experimental MS/MS spectra with curated spectra in a library. These curated spectra have been obtained experimentally, such as from analysis of a reference protein or peptide standard or a well-characterized marker. In addition to matching fragment masses between experimental and library spectra, ion abundances are also incorporated to generate a similarity score[8]. Each method has its benefits and disadvantages; for this untargeted proteomics work, hybrid search engines are investigated.

This section examines two search engines for maximizing peptide identification, the Global Proteome Machine (GPM) and PEAKS, with the intent of selecting one for application to

GVP identification in subsequent experiments. GPM, through the combination of open source software X! Tandem and X! Hunter, uses a database-searching algorithm and spectral library searching[10], while PEAKS presents a fusion of *de novo* sequencing and database searching[11]. Such approaches require selection of a set of search engine parameters defined by the analyst, and these should undergo critical evaluation to assess their influence on the number of identified peptides and proteins. Not only does maximizing peptide identification provide a larger pool of annotated peptides from which GVPs can be identified, but the pool size of annotated peptides also reflects how well the results describe proteomes. In addition, sample preparation methods influence the number of peptides detected, and various approaches were examined using bulk volumes of hair to guide method development for single hair analysis. Because analysis of single hairs involves processing of mass-limited specimens, optimization of protein extraction and protein digestion is key.

The chemical structure of hair that contributes to its rigidity and persistence in the environment also presents a challenge for hair sample preparation. Notoriously robust, keratins and keratin-associated proteins (KAPs), as the dominant components of hair, are supported by extensive crosslinking via disulfide and isopeptide bonds[12], which reduces solubility and makes protein extraction from the matrix and protein digestion less efficient. Effective protein extraction from the hair matrix requires cleavage of these bonds, as some crosslinked proteins comprise the component of hair considered insoluble[13]. To improve protein extraction, use of ultrasonication at elevated temperatures and detergent were considered. Protein extraction of recalcitrant matrices conventionally has utilized the detergent sodium dodecyl sulfate (SDS)[13, 14], but this reagent suppresses electrospray ionization and may inhibit proteolysis, thus making it incompatible with mass spectrometry analysis without further procedures to remove SDS from

digests that may result in losses of some peptides. Alternatively, the reagent sodium dodecanoate exhibits similar activity to SDS, as it also has a 12-carbon aliphatic chain, but unlike SDS, it can be converted to a neutral form by acidification and can be removed from solution via an acidified liquid-liquid extraction process[15] prior to mass spectrometry analysis. Coupled with detergent, ultrasonication at elevated temperatures aids protein denaturation and unfolding, facilitating extraction, with the expectation that in single one-inch hairs, digestion will yield peptides from the entire hair segment. Precipitation of solubilized proteins with acetone precipitation, a conventional protein concentration step[14, 16] offers advantages for single hair analysis in the form of sample volume reduction and removal of acetone-soluble non-proteinaceous substances including residual dodecanoate detergent. Subsequent ultrasonication in aqueous buffer is anticipated to aid resolubilization of the protein pellet that remains after precipitation. These strategies were investigated to determine the combination that maximizes protein and peptide identifications for implementation in single hair analysis.

2.2.1 Bulk Hair Preparation Methods

Following the sample preparation method described in Parker et al.[2], 10.0 ± 0.2 mg of scalp hair were weighed on an analytical balance and transferred to a milling vial containing 2.8 mm ceramic beads. To each vial, 200 µL of aqueous denaturation buffer, containing 150 µL of 8 M urea, 20 µL of 1 M dithiothreitol (DTT), 4 µL of 1 M ammonium bicarbonate, and 2 µL of 1% (w/v) ProteaseMAX™ Surfactant (Promega, Madison, WI), was added. Hair samples were homogenized for 3 min via milling at a rate of 4.5 m/s (Bead Ruptor 12, Omni International; Kennesaw, GA), followed by incubation overnight at room temperature (RT) on a revolving turntable at 30 rpm to facilitate mixing during protein extraction. Extracts were then alkylated to prevent disulfide bonds from re-forming via addition of 80 µL of 0.5 M iodoacetamide, briefly

homogenized as above, and incubated in the dark for 30 min at RT. Again, protein extracts were homogenized using the bead mill homogenizer. 1.0 mL of aqueous protein digestion solution (containing 25 µL of 1 µg/µL TPCK-treated trypsin (sequencing grade; Worthington Biochemical, Lakewood, NJ), 50 µL of 1 M DTT, 50 µL of 1 M ammonium bicarbonate, and 10 µL of 1% (w/v) ProteaseMAX™ Surfactant) was added to each extract. Extracts were then homogenized using the bead mill homogenizer and allowed to incubate overnight at RT on the turntable. Protein digests were centrifuged at 15,000 × *g* for 15 min at RT to separate the undigested hair pellet from digested material. Digest supernatant was transferred to an Eppendorf LoBind tube and centrifuged at 9,000 × *g* for 15 min at RT. Supernatant was transferred to a centrifugal filter tube (PVDF, 0.1 µm) and filtered for particulates. Filtrate was transferred to an autosampler vial for LC-MS/MS analysis.

The above preparation method is hereafter referred to as the High Volume Protocol (HVP). Five other variations to this method were evaluated for maximizing protein extraction from bulk volumes of hair and are described in Table 2.1 below. Additional concentration step acetone precipitation is denoted as Acetone PPT. Detergent sodium dodecanoate (SDD) was examined as an alternative reagent to the combination of urea and ProteaseMAX™ for protein extraction. Ultrasonication in a water bath was performed at 70 °C, at a frequency of 37 kHz at 100% power (Elma, Singen, Germany). Acetone precipitation was performed to concentrate proteins by adding chilled acetone in a 4:1 organic to aqueous phase ratio. The protein extract, after brief mixing, was allowed to incubate at -20 °C overnight, followed by centrifugation at 15,000 × *g* for 15 min at RT to separate the protein pellet from supernatant. After supernatant was removed, the pellet was washed with the same volume of chilled acetone, with supernatant discarded. For resolubilization of the protein pellet, a 0.975-mL aqueous buffer containing 50

mM each of DTT and ammonium bicarbonate, and 0.01% (w/v) ProteaseMAX™ Surfactant was added and the suspension was ultrasonicated in a water bath as described above. For each sample preparation method, a 10-mg quantity of scalp hair from each of 4 individuals was extracted and digested to evaluate effects of biological variation in hair origin.

Table 2.1. Parameters for variations of the High Volume Protocol for preparation of bulk volumes of hair samples.

Sample Preparation Method	Protein Extraction Reagent	Protein Extraction Incubation Method	Protein Concentration	Resolubilization Method	Protein Digestion Buffer
HVP	150 µL of 8 M urea	Turntable rotation, overnight, RT, 30 rpm	None	None	1.0 mL buffer containing trypsin
HVP Acetone PPT	150 µL of 8 M urea	Turntable rotation, overnight, RT, 30 rpm	Acetone precipitation, overnight, -20 °C	None	1.0 mL buffer containing trypsin
HVP Acetone PPT-Sonication	150 µL of 8 M urea	Turntable rotation, overnight, RT, 30 rpm	Acetone precipitation, overnight, -20 °C	0.975 mL buffer, ultrasonication, 70 °C, 12 h	25 µL of 1 µg/µL trypsin
HVP SDD-Acetone PPT	80 µL of 5% (w/v) SDD	Turntable rotation, overnight, RT, 30 rpm	Acetone precipitation, overnight, -20 °C	None	1.0 mL buffer containing trypsin
HVP SDD-Acetone PPT-Sonication	80 µL of 5% (w/v) SDD	Turntable rotation, overnight, RT, 30 rpm	Acetone precipitation, overnight, -20 °C	0.975 mL buffer, ultrasonication, 70 °C, 12 h	25 µL of 1 µg/µL trypsin
HVP SDD-Acetone PPT-Sonication2X	80 µL of 5% (w/v) SDD	Ultrasonication, 70 °C, 12 h	Acetone precipitation, overnight, -20 °C	0.975 mL buffer, ultrasonication, 70 °C, 12 h	25 µL of 1 µg/µL trypsin

2.2.2 Liquid Chromatography-Tandem Mass Spectrometry Analysis on an LC-QTOF-MS

Filtered protein digests were analyzed on Agilent 1290 Infinity liquid chromatograph coupled to a 6550 iFunnel quadrupole time-of-flight mass spectrometer (Agilent Technologies, Santa Clara, CA). Injection volumes of 10 µL were separated on an AdvanceBio Peptide Mapping C18 analytical column (2.1 mm × 150 mm, 2.7 µm particle size, 120 Å pores; Agilent

Technologies, Santa Clara, CA). Separations were performed at a flow rate of 0.2 mL/min using mobile phases A (0.1% formic acid in water) and B (0.1% formic acid in acetonitrile) over a 90-min gradient: hold at 3% B for 5 min, 3 to 33% B in 75 min, 33 to 50% B in 5 min, and ramped to 95% B in 5 min. Flow was diverted to waste during the initial 5-min hold. Positive mode electrospray ionization was achieved using a capillary voltage of 3.5 kV, with gas and sheath gas temperatures of 250 °C, drying gas flow rate of 14 L/min, nebulizer gas pressure of 35 psig, fragmentor voltage of 150 V, and an octopole RF of 750 V. Full MS survey scans were acquired over a scan range between *m/z* 100 and 1700, at a scan rate of 8 spectra/s. Data-dependent MS/MS scans were triggered for the 10 most abundant survey scan ions at an intensity threshold of 5.0×10^4, scan rate of 3 spectra/s, dynamic exclusion of 30 s, and an isolation window of 4 Da. CID fragmentation was performed using collision energy ramps, with a slope of 3.5 and offsets of 4.7, 0.7, -3.3, and 6 for charge states of 2, 3, > 3, and 1, respectively, in descending order of priority. Singly-charged species were least prioritized and ions with unassigned charge states were excluded from MS/MS.

2.2.3 Parameters that Affect Peptide Identification

This section sought to identify the parameters in two different search engines, GPM and PEAKS, that maximize peptide identification, and to further select a search engine for application to non-targeted GVP identification. As parameters differ between the two search engines, local parameter optimization within each search engine was performed before comparing GPM and PEAKS performance.

Table 2.2 below lists the parameters that were varied to assess their effects on protein and peptide identification. Two different file formats, the Mascot Generic Format (mgf) and the mass spectrometry-based eXtensible Markup Language (mzXML) file, were examined, with the

former containing only the tandem mass spectral data, whereas the latter includes both the MS and MS/MS spectral information from raw spectral files[17, 18]. Different sequence databases including ENSEMBL's genome database[19], a joint endeavor by the European Bioinformatics Institute and the Wellcome Trust Sanger Institute, and SwissProt, a manually reviewed database within UniProtKB[20], were also evaluated. Performance from two different versions of GPM were compared, as the peptide identification algorithm may differ across different iterations of the search engine. Protein and peptide expectation values[21] and fragment mass error tolerance aid filtering of peptide identifications; application of more conservative values restrict identification of peptides that may not be commonly observed while widening these windows may introduce a large number of false positives. Performance differences arising from changing the other parameters, such as inclusion of various post-translational modifications, selection of a fixed or variable cysteine carbamidomethylation modification, and permittance of missed cleavage sites within peptides, all reflect the nature of the sample or indicate efficacy of various steps in the sample preparation process. For example, a substantial difference in peptide yields from selecting a variable carbamidomethylation modification as opposed to a fixed modification would indicate incomplete alkylation. Similarly, an increase in the number of peptides with missed cleavage sites may suggest a less efficient protein digestion process. Optimizing parameters for maximal peptide identification not only serves to benefit GVP identification, but is also necessary for selecting parameters that adequately describe the proteomics sample.

Table 2.2. List of GPM parameters examined for effects on peptide identification. "G" precedes each GPM dataset number.

Variable	GPM Datasets									
	G1	**G2**	**G3**	**G4**	**G5**	**G6**	**G7**	**G8**	**G9**	**G10**
File Format & Conversion Method*	mgf-MH	mgf-PW	mgf-PW	mgf-PW	mgf-PW	mgf-PW	mgf-PW	mgf-PW	mgf-PW	mzXML-PW
Software Version**	GPM Cyclone	GPM Fury	GPM Fury	GPM Fury	GPM Fury	GPM Fury	GPM Cyclone	GPM Cyclone	GPM Cyclone	GPM Fury
Protein Database	ENSEMBL	ENSEMBL	ENSEMBL	ENSEMBL	ENSEMBL	ENSEMBL	SwissProt	SwissProt	SwissProt	ENSEMBL
Protein and Peptide Expectation Values	Protein: < 0.0001 Peptide: < 0.01	Protein: < 0.0001 Peptide: < 0.01	Protein: < 0.1 Peptide: < 0.1	Protein: < 0.1 Peptide: < 0.1	Protein: < 0.1 Peptide: < 0.1	Protein: < 0.1 Peptide: < 0.1	Protein: < 0.1 Peptide: < 0.1	Protein: < 0.1 Peptide: < 0.1	Protein: < 0.1 Peptide: < 0.1	Protein: < 0.1 Peptide: < 0.1
Fragment Mass Error	0.4 Da	0.4 Da	0.4 Da	0.05 Da	0.05 Da	0.05 Da	0.05 Da	0.05 Da	0.05 Da	0.05 Da
Potential Modifications†	Variable C; M	Variable C; M	Variable C; M	Variable C; M; N & Q	Variable C; M; N & Q	Fixed C; M; N & Q	Variable C; M; N & Q	Variable C; M; N & Q	Fixed C; M; N & Q	Variable C; M; N & Q
Other Potential PTMs	No	No	No	No	No	Yes	No	Yes	Yes	No
Refinement Modifications†	N & Q; M & W	N & Q; M & W	N & Q; M & W	N & Q; M & W	N & Q; M & W	N & Q; M & W	N & Q; M & W	N & Q; M & W	N & Q; M & W	N & Q; M & W
Refinement Other Potential PTMs	No	No	No	No	No	Yes	No	Yes	Yes	No
Refinement Missed cleavage	No	No	No	No	Yes	Yes	No	Yes	Yes	No
Trypsin Missed Cleavage	No	No	No	Yes; max 1	Yes; max 1	Yes; max 3	Yes; max 1	Yes; max 1	Yes; max 3	Yes; max 1

*mgf = Mascot Generic Format file; mzXML = mass spectrometry-based eXtensible Markup Language file; MH = Agilent MassHunter software for raw mass spectral file conversion; PW = ProteoWizard open-source software for file conversion

**GPM Cyclone is a newer version (2017.2.1.4) of the search engine compared to GPM Fury (2017.2.1.3)

†C = cysteine carbamidomethylation (+57.0215 Da); M = methionine oxidation (+15.9949 Da); N & Q = asparagine and glutamine deamidation (+0.9840 Da); M & W = methionine and tryptophan oxidation (+15.9949 Da) and di-oxidation (+31.9898 Da)

To assess effects of the aggregated variables in Table 2.2 on peptide identification, a factorial multivariate analysis of variance (MANOVA) containing 12 independent variables and two dependent variables (numbers of identified proteins and peptides) was performed using the *manova* function in the *stats v3.5.1* package in R (x64 version 3.4.4), followed by univariate analysis of variance (ANOVA) post-hoc tests for each dependent variable, through the *aov* function in the same package. Bonferroni corrections to p-values were applied for univariate ANOVAs. Though listed in Table 2.2, the variable Refinement Modifications was not included in this analysis as the same criteria were applied to all datasets. Refinement refers to a second peptide identification process in which tandem mass spectra are matched to a list of protein sequences with slightly different parameters than in the first process; in this case, the second identification process matched tandem mass spectra to protein sequences, constraining the modifications to methionine and tryptophan oxidation and asparagine and glutamine deamidation. Table 2.3 details the results of statistical comparisons within the 12 independent variables, which includes sample preparation method and biological variation in addition to GPM peptide identification parameters. Statistical significance was assigned at $\alpha = 0.05$ for comparisons yielding statistically greater numbers of identified proteins and peptides. Interactions between independent variables are not reported.

Table 2.3. Statistical significance of GPM parameters on numbers of identified proteins and peptides. After the variables Sample Preparation Method and Biological Variation, parameters are listed in decreasing importance based on MANOVA p-values. Df = degrees of freedom.

Variable	Df	Pillai's trace	MANOVA p-value	Univariate ANOVA p-value	
				Proteins	Peptides
Sample Preparation Method	5	1.33	8.05×10^{-78}	7.71×10^{-82}	6.77×10^{-39}
Biological Variation	3	0.63	2.22×10^{-26}	6.71×10^{-23}	2.73×10^{-26}
Protein and Peptide Expectation Values	1	0.79	3.04×10^{-61}	2.90×10^{-60}	3.83×10^{-6}
File Format & Conversion Method	2	0.77	5.18×10^{-36}	9.68×10^{-54}	2.93×10^{-3}
Software Version	1	0.19	7.33×10^{-9}	2.76×10^{-7}	1
Fragment Mass Error	1	0.19	8.61×10^{-9}	2.18×10^{-8}	8.43×10^{-1}
Trypsin Missed Cleavage	1	0.002	8.42×10^{-1}	1	1
Refinement Missed Cleavage	1	0.002	8.63×10^{-1}	1	1
Other Potential PTMs	1	0.0001	9.90×10^{-1}	1	1

As expected, sample preparation method explains the greatest amount of variance in the numbers of identified proteins ($p = 7.71 \times 10^{-82}$) and peptides ($p = 6.77 \times 10^{-39}$), though biological variation among individuals also accounts for a substantial portion of the differences in the two metrics ($p = 6.71 \times 10^{-23}$ and $p = 2.73 \times 10^{-26}$, respectively). Effects of these two variables are examined later in this section.

Parameters in GPM that critically impact peptide identification consist of protein and peptide expectation values and to lesser extents, protein database and fragment mass error. Protein and peptide expectation values are used to evaluate peptide-spectrum match scores, where higher confidence of peptide-spectrum matches is associated with smaller expectation values, or the probability that a peptide-spectrum match was obtained by chance[21]. Given a

peptide-spectrum match score of *x* among all scores for peptide-spectrum matches from a protein database, expectation values represent the product of the number of protein or peptide sequences scored and the probability that a score higher than *x* could be obtained by randomly matching tandem mass spectra with database sequences[21]. For example, for a peptide-spectrum match score of *x* with a protein expectation value of 0.01, the protein sequence database would need to contain 100 times as many sequences or the experiment would need to be repeated 100 times to obtain a score of *x* by chance[21]. Not surprisingly, increasing the threshold expectation values from < 0.0001 and < 0.01 to < 0.1 for proteins and peptides, respectively, enabled a statistically greater number of protein and peptide identifications ($p = 2.90 \times 10^{-60}$ and $p = 3.83 \times 10^{-6}$ for proteins and peptides, respectively; comparison between Sets G2 and G3 in Figure 2.1a, d). File conversion through open-source software ProteoWizard[22], as opposed to Agilent's MassHunter software, appears to benefit protein ($p = 9.68 \times 10^{-54}$; Figure 2.1b) and peptide identification ($p = 2.93 \times 10^{-3}$; Figure 2.1e), though the datasets G2 – G10 for comparison to Set G1 used larger protein and peptide expectation values, which as discussed above, likely has a greater effect than the file conversion method itself.

Other variables were deemed important for protein identification, namely GPM version and fragment mass error. While the older version Fury yields a slightly larger number of proteins compared to Cyclone ($p = 2.76 \times 10^{-7}$; Figure 2.1c), again, additional parameters confound this comparison, as all analyses using Fury involved the ENSEMBL database, whereas searches using Cyclone were almost always performed with the UniProtKB/SwissProt database. It is much more likely that using different protein databases for peptide identification influences the number of identified proteins, particularly since protein populations differ between ENSEMBL and SwissProt databases. SwissProt contains proteins whose annotations have been manually

reviewed[20] while ENSEMBL contains a similar set of proteins in addition to proteins predicted from transcripts of novel genes[19]. As such, searching ENSEMBL (45,906 transcripts from human genes in build GRCh38 downloaded on November 8, 2017 compared to 42,202 protein and isoform sequences in the SwissProt Human database downloaded on September 21, 2017) yielded a slightly greater number of identified proteins.

Finally, use of a larger fragment mass error window, 0.4 Da, yielded a greater number of proteins ($p = 2.18 \times 10^{-8}$; comparison between Sets G3 and G4 in Figure 2.1a). This effect is expected, as a narrower mass error tolerance (i.e., 0.05 Da) restricts more variable fragment masses from being included in peptide mass lists when matching to those from theoretical spectra. With a QTOF mass analyzer, mass errors of 20 ppm (0.016 Da at *m/z* 800, the average *m/z* value in the range of interest) or less are considered common owing to limitations of ion statistics and drift in laboratory temperatures, although in practice, errors within 100 ppm (0.08 Da at *m/z* 800) may be encountered. Though use of the narrower mass window (0.05 Da) yields reports of fewer peptides, this conservative measure ensures a higher quality of peptide-spectrum matches, from which a more confident set of proteins can then be inferred.

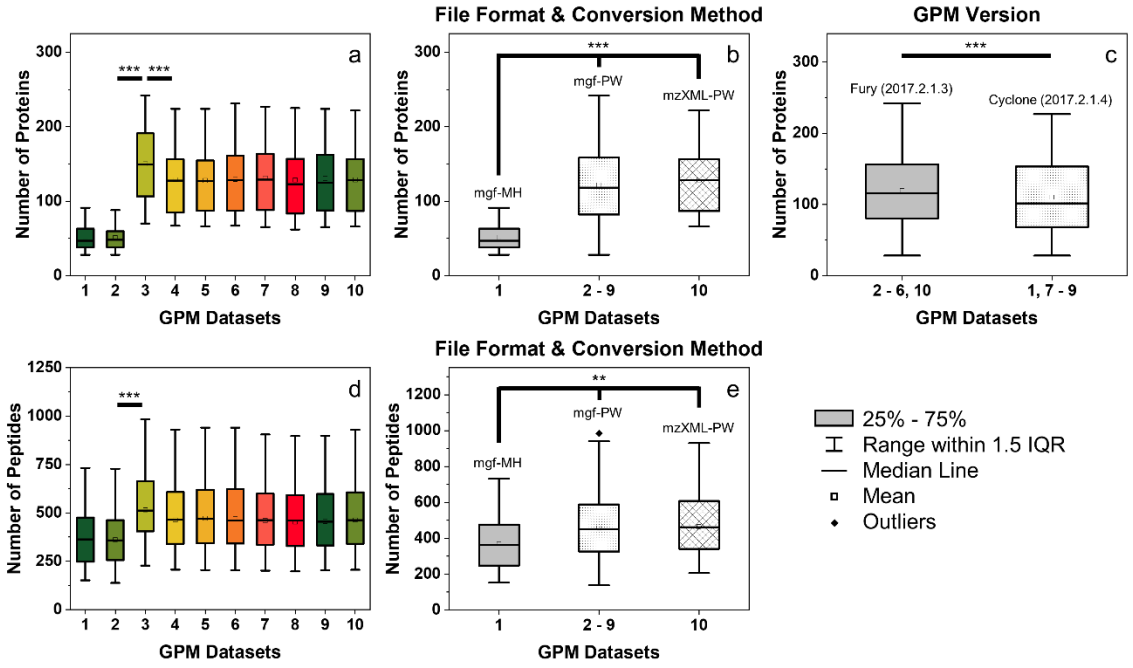


Figure 2.1. Comparison of the effects of GPM parameters on the numbers of identified proteins and peptides, using factorial MANOVA and univariate ANOVAs as post-hoc tests. Number of identified proteins for (a) each dataset and comparison based on aggregate datasets by (b) file format and conversion method, and (c) GPM version. Number of identified peptides for (d) each dataset and comparison based on aggregate datasets by (e) file format and conversion method. Each dataset has been averaged across sample preparation methods and individuals. Black lines represent statistically significant comparisons and significance levels are represented as $p \leq 0.05$ (*), $p \leq 0.01$ (**), and $p \leq 0.001$ (***). (a) shows statistically significant comparisons for Protein and Peptide Expectation Values (Set G2 vs. G3) and Fragment Mass Error (Set G3 vs. G4). (d) shows a statistically significant comparison for Protein and Peptide Expectation Values. Larger expectation values yield greater numbers of identified proteins and peptides. A larger fragment mass error tolerance yields more proteins. Conversion via ProteoWizard (PW) appears to yield more proteins and peptides, though other variables contribute to the differences in these aggregated datasets, chiefly larger protein and peptide expectation values. Slightly more proteins were identified using the older GPM version Fury, though Fury was often coupled to ENSEMBL as the protein database, as opposed to the UniProtKB/SwissProt database used in Cyclone, which likely exerts a greater influence on the metric.

Hair proteins are subjected to an assortment of chemical modifications, with some occurring during their biosynthesis, others resulting from aging and environmental exposures, and a third set may occur during sample processing for proteome analysis. Successful protein and peptide detection requires that matches to database sequences allow for such modifications. Notably, inclusion of the deamidation modification to asparagine and glutamine residues, a

commonly encountered chemical modification in proteins[23], and alternating between a fixed and variable carbamidomethylation modification (comparison of Sets G5 and G6) had negligible effect on protein and peptide identification. In Set G5, 128 ± 45 (mean ± s.d.) proteins and 472 ± 180 peptides were identified, not statistically different from 129 ± 46 proteins and 471 ± 180 peptides in Set G6 (Figure 2.1a, d). It was expected that the addition of a common post-translational modification would yield a greater number of identified proteins and peptides, as it enables peptide-spectrum matching for a set of previously unmatched experimental spectra if the modification is indeed prevalent within the protein digest samples. However, it is likely that the effect of the former was compounded in a comparison of fragment mass error (between Sets G3 and G4), as inclusion of the deamidation modification was introduced concomitantly with a narrower fragment mass error tolerance. Additionally, the modification was included during the refinement, a second peptide identification process, which may minimize any effects from excluding it during the initial cycle of peptide-spectrum matching.

Alternating between a variable or fixed carbamidomethylation modification induced minimal effect on the number of identified peptides (472 ± 180 and 471 ± 180 for Sets G5 and G6, respectively). This observation suggests that the alkylation step during sample preparation reached near-completion and that only one cysteine alkylation site remained unmodified after incubation with iodoacetamide. A greater number of peptides would have been identified by searching with a variable carbamidomethylation modification if alkylation had been far from complete. Numbers of identified protein and peptide also remained unchanged when widening the tolerance on the maximum number of missed cleavage sites during protein digestion with trypsin from 1 to 3 missed cleavage sites within each peptide (comparison of Sets G5 and G6). Although this parameter change was introduced concurrently with a fixed carbamidomethylation

modification, protein and peptide identification invariance between the two datasets indicates that in addition to a complete alkylation, protein digestion with the enzyme trypsin was consistent and effective. A substantial increase in the number of identified peptides with a widening maximum number of missed cleavage sites would have suggested incomplete protein digestion efficacy. In sum, changes in post-translational modification and protein digestion parameters minimally influence protein and peptide identification in hair samples; protein and peptide expectation values, protein database selection, and fragment mass error tolerance exerted the largest impacts on database-searching outcomes using GPM.

Parameter selection for maximal peptide identification in GPM also necessitates optimizing for a low false discovery rate, that is, minimal misidentification of peptides, which is an important part of ensuring that search engine results reflect the peptide populations within the samples themselves. As such, conservative measures may need to be adopted for greater confidence in the set of identified peptides. Relaxing protein and peptide expectation values but narrowing the fragment mass error tolerance (as in Set G4) achieves this balance. Discussed above, protein database selection also influenced peptide identification during database searching in GPM. SwissProt, as the manually curated database, represents a biologically relevant protein population whereas ENSEMBL contains a set of putative and predicted proteins that may not be expressed. Because this work inherently contains an additional level of complexity to identify mutated peptides, i.e., noncanonical peptides that carry amino acid polymorphisms, use of a database where proteins whose biological relevance has not been verified may result in misidentifications of GVPs. As the database searching approach attempts to maximize matches between observed peptides and *in silico* peptides derived from database protein sequences, minor GVPs may be incorrectly identified as having matched to regions of predicted protein sequences

that are never expressed rather than as non-detects from databases containing only biologically relevant proteins, resulting in false positives. Thus, despite a larger number of proteins encoded within ENSEMBL, the SwissProt database was preferentially selected; the GPM dataset that embodies these parameters is Set G8.

Although the PEAKS search engine employs *de novo* sequencing, the software has a built-in peptide-spectrum matching process analogous to GPM, with similar parameters that require optimization. In addition to supporting analysis from processed file formats, the software allows raw mass spectral files to be deposited directly for analysis. Also different from GPM is the ability for users to deposit custom protein databases for database searching, whereas only those protein databases contained within the server were provided as options in GPM. Further, while GPM uses protein and peptide expectation values to filter peptide-spectrum matches, PEAKS exploits a target-decoy method to determine a false discovery rate for peptide misidentification[11]. Conventional target-decoy relies on the introduction of decoy, or biologically nonsensical, proteins, that is, sequences that are not in the database, for each database sequence as separate entries during the database searching phase and determines the false discovery rate based on the percentage of decoy peptides that form a peptide-spectrum match[24], whereas the decoy fusion approach used in PEAKS inserts the decoy proteins, sequences derived from randomly shuffling amino acids within each protein sequence entry in the database, into the same sequence as those in the user-defined protein database[11]. This modification affords a more conservative approach than the target-decoy method when incorporated into the PEAKS database searching algorithm, as more decoy matches can be formed, thus preventing an underestimation of the false discovery rate[11]. Table 2.4 below details the parameters in PEAKS that were examined for maximal peptide identification. Software version is not expected to affect

peptide identification, as in the results described above, though is listed for completeness. A fragment mass error tolerance of 0.05 Da was selected here based on the discussion above. Furthermore, similar to the parameters adopted in GPM searches, methionine oxidation and asparagine and glutamine deamidation were applied as variable modifications, and missed cleavages were allowed, as these parameters reflected the peptide populations within the digested hair samples. Again, variable and fixed carbamidomethylation was examined, although it is not expected to affect peptide identification as observed above.

Table 2.4. List of PEAKS parameters examined for effects on peptide identification. "P" precedes each PEAKS dataset number.

Variable	PEAKS Datasets							
	P11	**P12**	**P13**	**P14**	**P15**	**P16**	**P17**	**P18**
File Format & Conversion Method*	Agilent raw-none	Agilent raw-none	mgf-PW	mgf-PW	Agilent raw-none	Agilent raw-none	mgf-PW	mgf-PW
Software Version	PEAKS v7.5	PEAKS v8	PEAKS v7.5	PEAKS v8	PEAKS v8	PEAKS v8	PEAKS v8	PEAKS v8
Protein Database	SwissProt	SwissProt‡	SwissProt	SwissProt‡	SwissProt	SwissProt‡	SwissProt‡	SwissProt‡
False Discovery Rate	1.0%	1.0%	1.0%	1.0%	1.0%	0.5%	0.5%	0.5%
Fragment Mass Error	0.05 Da	0.05 Da	0.05 Da	0.05 Da	0.05 Da	0.05 Da	0.05 Da	0.05 Da
Potential Modifications†	Variable C; M; N & Q	Variable C; M; N & Q	Variable C; M; N & Q	Variable C; M; N & Q	Variable C; M; N & Q	Fixed C; M; N & Q	Variable C; M; N & Q	Fixed C; M; N & Q
Other Potential PTMs	Yes; max 3	Yes; max 3	Yes; max 3	Yes; max 3	Yes; max 3	Yes; max 3	Yes; max 3	Yes; max 3
Trypsin Missed Cleavage	Yes; max 3	Yes; max 3	Yes; max 3	Yes; max 3	Yes; max 3	Yes; max 3	Yes; max 3	Yes; max 3

*Agilent raw = unprocessed mass spectral file from Agilent LC-QTOF-MS; mgf = Mascot Generic Format file; PW = ProteoWizard open-source software for raw mass spectral file conversion

‡SwissProt database was downloaded from GPM instead of UniProtKB

†C = cysteine carbamidomethylation (+57.0215 Da); M = methionine oxidation (+15.9949 Da); N & Q = asparagine and glutamine deamidation (+0.9840 Da)

Factorial MANOVA was performed with seven independent variables and three dependent variables as detailed above to identify parameters in PEAKS that impact peptide identification. The variables Fragment Mass Error, Other Potential PTMs, and Trypsin Missed

Cleavage were not included in the MANOVA analysis as they did not vary among the datasets. In addition to the numbers of identified proteins and peptides, the percentage of peptide-spectrum matches (% PSMs) was used as a metric for evaluating PEAKS parameters. % PSMs was determined as the fraction of peptide-spectrum matches out of a total number of acquired tandem mass spectra. Similar to the analysis of GPM parameters above, univariate ANOVAs with Bonferroni corrections to p-values were performed as post-hoc tests, and Table 2.5 lists the statistical significance of each parameter by decreasing importance based on MANOVA p-values.

Table 2.5. Statistical significance of PEAKS parameters on numbers of identified proteins and peptides, and the percentage of peptide-spectrum matches. After the variables Sample Preparation Method and Biological Variation, parameters are listed in decreasing importance based on MANOVA p-values.

Variable	Df	Pillai's trace	MANOVA p-value	Univariate ANOVA p-value		
				Proteins	Peptides	%PSMs
Sample Preparation Method	5	1.72	6.20×10^{-69}	6.45×10^{-33}	1.53×10^{-27}	1.04×10^{-18}
Biological Variation	3	0.63	1.14×10^{-17}	4.61×10^{-6}	3.93×10^{-17}	2.30×10^{-8}
File Format & Conversion Method	1	0.27	1.86×10^{-9}	5.78×10^{-5}	7.82×10^{-1}	9.51×10^{-2}
Software Version	1	0.13	2.62×10^{-4}	7.29×10^{-2}	1	1
False Discovery Rate	1	0.13	2.72×10^{-4}	1	2.04×10^{-1}	7.89×10^{-2}
Potential Modifications	1	0.02	3.49×10^{-1}	1	1	1
Protein Database	1	0.01	5.98×10^{-1}	1	1	1

For the majority of parameters examined in PEAKS, modification within each variable yielded one statistical difference in the number of proteins and no statistical differences in peptide identification and the percentage of peptide-spectrum matches (Figure 2.2). Again, sample preparation method and biological variation among individuals dominate the variance

among dependent variables. Analysis from Agilent raw mass spectral files compared to converted mgf files (via ProteoWizard) resulted in a greater number of identified proteins in the former, i.e., 205 ± 80 proteins averaged from Sets P11 and P12 as opposed to 167 ± 70 proteins from Sets P13 and P14 ($p = 5.78 \times 10^{-5}$; Figure 2.2a), though this effect is likely marginal as peptide identification and the percentage of peptide-spectrum matches were not statistically different (855 ± 321 peptides averaged from Sets P11 and 12 compared to 816 ± 365 peptides from Sets P13 and P14). In general, the search engine reports all proteins that contain the sequence of the identified peptide (i.e., a peptide shared among multiple proteins), as PEAKS utilizes parsimony in protein inference[11, 25], which inflates the number of inferred proteins from identification of one peptide. Applied to the comparison of Sets P11 and P12 and Sets P13 and 14, a slight reduction in the number of identified peptides (5% decrease) with use of mgf files, which was not statistically different ($p = 0.782$; Figure 2.2b), resulted in a decrease of approximately the same number of proteins (19% decrease); this decrease in the number of proteins was considered statistically different owing to the larger percent change. As this may affect protein identification when applied to subsequent analyses, the protein inference process is addressed in Section 2.3.4. Regardless, usage of the most unprocessed data format when possible avoids file reading errors that may arise from file format conversions. Interestingly, requiring a lower false discovery rate for peptide-spectrum matches did not affect peptide identification (comparison between Sets P14 (820 ± 373 peptides) and P17 (779 ± 353) was not statistically significant), though it is likely due to an incremental difference between the two levels of this variable (from 1% to 0.5%). Further reduction in false discovery rates (e.g., from 1% to 0.1%) results in 13% fewer peptide identifications, though restricting this parameter may hinder discovery of noncanonical peptides, such as GVPs. The convention in proteomics is 1% FDR.

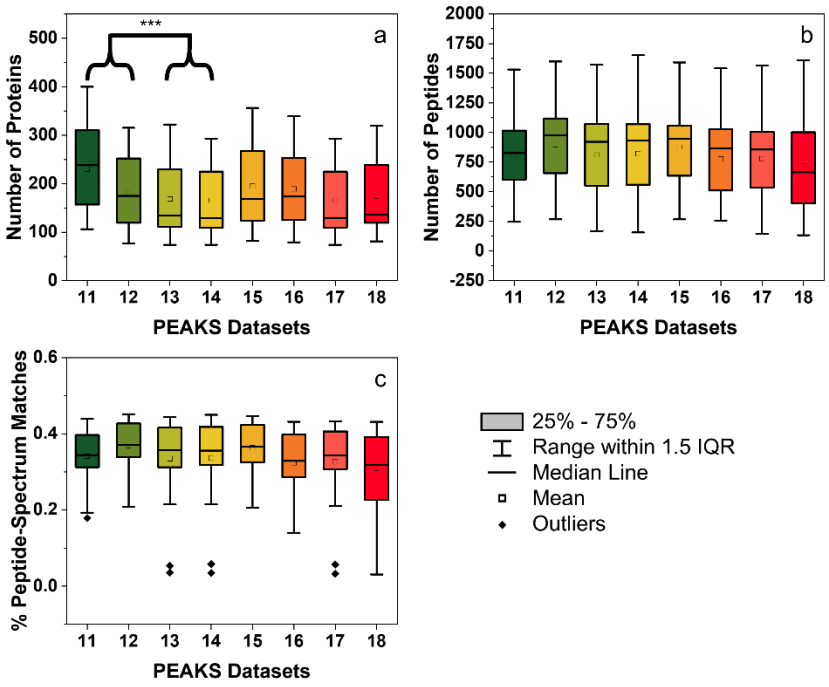


Figure 2.2. Comparison of the effects of PEAKS parameters on the numbers of identified (a) proteins and (b) peptides, and the (c) percentage of peptide-spectrum matches. Each dataset has been averaged across sample preparation methods and individuals. Black lines represent statistically significant comparisons and significance levels are represented as $p \leq 0.05$ (*), $p \leq 0.01$ (**), and $p \leq 0.001$ (***). (a) shows a statistically significant comparison of file format. Peptide identification from raw mass spectral files (unprocessed) yields a greater number of identified proteins.

To compare performance from GPM and PEAKS database searching, a factorial MANOVA with 3 independent variables (i.e., search engine, sample preparation method, and individual) and 2 dependent variables (numbers of identified proteins and peptides) was performed on Sets G8 and P12, followed by univariate ANOVAs for each dependent variable. G8 embodies search parameters optimized for GPM, discussed above, and P12 parameters are similar to those in G8, including protein sequence database, but includes a slightly larger, conventional FDR (1%) and variable carbamidomethylation modification to enable a less restrictive peptide identification, especially if alkylation does not achieve completion in hair samples. Table 2.6 tabulates the numbers of proteins and peptides (mean ± s.d.) among the six sample preparation methods for the two datasets. Appendix Table S-2.1 lists the values for each

of the 24 bulk hair samples analyzed. The numbers of peptides reported by PEAKS in Set P12 were truncated after generating a non-redundant list of identified peptide sequences for direct comparison to the number of peptides reported by GPM. PEAKS counts different combinations of peptide sequence and post-translational modification as separate peptides whereas GPM reports the number of peptides corresponding to unique peptide sequences, thus requiring de-duplication of peptide sequences when reporting the number of peptides identified using PEAKS; the number of PEAKS-identified peptides in Table 2.6 represents counts from non-redundant, de-duplicated peptide lists.

Table 2.6. Average numbers of proteins and peptides (mean ± s.d.) across sample preparation methods (n = 4 hair samples from 4 individuals per method) from datasets G8 and P12. Statistical significance is indicated for variables that achieved peak performance.

Sample Preparation Method	GPM Set G8		PEAKS Set P12[‡]	
	Proteins	**Peptides**	**Proteins**	**Peptides***
HVP	114 ± 18	522 ± 128	162 ± 35	657 ± 170
HVP Acetone PPT	101 ± 23	390 ± 76	124 ± 13	510 ± 78
HVP Acetone PPT-Sonication[†]	203 ± 26	552 ± 257	277 ± 29	911 ± 212
HVP SDD-Acetone PPT	71 ± 8	213 ± 16	99 ± 19	314 ± 95
HVP SDD-Acetone PPT-Sonication	159 ± 6	563 ± 70	235 ± 45	739 ± 103
HVP SDD-Acetone PPT-Sonication2X	123 ± 29	476 ± 112	185 ± 63	644 ± 158

*Number of peptides reported represents a non-redundant list of peptide sequences

[†]MANOVA $p = 2.91 \times 10^{-10}$; univariate ANOVA protein $p = 5.74 \times 10^{-12}$, peptide $p = 4.70 \times 10^{-10}$

[‡]MANOVA $p = 2.17 \times 10^{-8}$; univariate ANOVA protein $p = 3.77 \times 10^{-7}$, peptide $p = 4.45 \times 10^{-7}$

PEAKS outperformed GPM in both protein and peptide identification across the six different sample preparation methods ($p = 3.77 \times 10^{-7}$ and $p = 4.45 \times 10^{-7}$ for proteins and peptides, respectively). For example, peptide identification using PEAKS permitted identification of 277 ± 29 proteins and 911 ± 212 peptides from bulk hairs prepared using the HVP Acetone PPT-Sonication method, statistically greater than the 203 ± 26 proteins and 552 ± 257 peptides identified using the same sample preparation but GPM for peptide identification

(Table 2.6). This suggests that the hybrid *de novo* sequencing-database searching approach offers deeper profiling of hair proteomes, in this case, unmutated peptides from canonical protein sequences.

Compared to GPM, PEAKS offers features that are more appropriate for identification of genetically variant peptides. Spectral library searching, while advantageous for de-replication, does not facilitate identification of uncommon peptides such as GVPs, which are unlikely to be curated in GPM's database. On the other hand, initial *de novo* sequencing of experimental spectra in PEAKS facilitates peptide identification by focusing the database search to a shorter list of proteins while performing peptide sequence inferences without expectations of matching to protein database entries; this approach is expected to be beneficial for identifying uncommon peptides. Another advantage of utilizing PEAKS for peptide identification is the ability to search experimental spectra against custom protein databases. This is particularly attractive for identification of GVPs, as these mutated peptides cannot be identified from canonical sequences unless single amino acid polymorphisms (SAPs) are allowed and specified during the search process, as would be required by GPM. However, each individual carries a distinct set of SAPs, especially those derived from single nucleotide polymorphisms; permitting all SAPs during peptide identification may lead to many false positive GVP identifications. Furthermore, single nucleotide polymorphisms identified from exome sequences of individual subjects will serve to direct GVP discovery in this work; as such, incorporation of custom protein sequences already containing the amino acid consequences from identified SNPs will prevent false positive identifications of GVPs. Given these advantages and its performance in identifying canonical peptides above, PEAKS was selected as the software of choice for peptide identification. As an

aside, GPM ended its online data analysis service in 2018 and is no longer available as a public peptide identification tool.

Biological variation among the hair samples from four individuals accounted for a large portion of the variance in identified proteins and peptides. Protein digests from the hair samples of Individual 2 consistently resulted in the largest protein and peptide yields (209 ± 79 proteins and 784 ± 249 peptides averaged over the sample preparation methods, as reported in PEAKS; $p = 3.19 \times 10^{-2}$ and $p = 3.20 \times 10^{-6}$, respectively). In contrast, hair samples from Individual 1 yielded the fewest numbers of proteins and peptides (161 ± 68 proteins and 528 ± 190 peptides). Variable hair protein extraction efficiency among individuals may be attributed to interindividual differences in hair type, grooming, and conditioning. As such, biological variation in hair from interindividual differences is included as an additional variable when statistically comparing other independent variables in the experimental design.

In addition to identifying a search engine and optimizing parameters for peptide identification, this section also aimed to identify sample preparation methods to guide development of single hair analysis that would be effective for different types of hair among individuals, given the large biological variation observed above. Comparison of different sample preparation methods for bulk volumes of hair demonstrated maximal hair protein extraction efficacy using acetone precipitation for protein concentration and ultrasonication for resolubilization of the protein pellet prior to digestion (HVP Acetone PPT-Sonication) ($p = 5.74 \times 10^{-12}$ and $p = 4.70 \times 10^{-10}$ for proteins and peptides, respectively). Alone, acetone precipitation (HVP Acetone PPT) yielded fewer proteins and peptides when compared with the HVP method. These two results indicate that although acetone precipitation concentrates protein into a pellet, cleavage sites remain inaccessible to trypsin during protein digestion without adequate

resolubilization; protein concentration is in fact an effective strategy for hair samples but requires resolubilization, in this case, via ultrasonication, before digestion.

Similarly, introduction of ultrasonication steps facilitated extraction of hair proteins and protein digestion when chaotropic agent urea was replaced with detergent sodium dodecanoate. Replacement of urea with detergent alone for protein extraction (HVP SDD-Acetone PPT) showed lowest protein and peptide yields among the six methods. But when accompanied by ultrasonication, first during protein extraction and again during protein pellet resolubilization (HVP SDD-Acetone PPT-Sonication2X) or even only for resolubilization (HVP SDD-Acetone PPT-Sonication), performance improved substantially, with yields at least approximating those from the HVP approach. Overnight ultrasonication during protein extraction resulted in a slight decrease in the numbers of proteins and peptides over simple turntable rotation; it is likely that extracted proteins experienced some degradation in the course of an extensive ultrasonication. Though effective, an overnight protein extraction step such as the turntable rotation used in these experiments may not be the most efficient with regards to sample preparation time. As such, a shorter ultrasonication step to facilitate protein extraction with minimal protein degradation poses an attractive alternative for single hair analysis. Further, any use of heat during protein extraction automatically precludes use of urea, as heat accelerates its decomposition to isocyanic acid in solution, reducing its protein extraction efficiency and leading to preparation-induced carbamylation in proteins[26]. These collective results underscore the advantage of protein concentration in hair sample preparation and outlines potential alternative methods for protein extraction, which is investigated for single hair analysis in Section 2.3.

2.3 Optimization of Single Hair Analysis

Because of the more than 100-fold difference in amount of material between 10 mg of scalp hair and a single hair fiber, successful single hair analysis for GVP markers hinges on maximizing protein extraction and digestion during sample preparation. Mason et al. described a method utilizing ultrasonication with a detergent that enabled similar peptide and GVP identification performance in single hairs compared to bulk amounts[15]. Comparing this method to one where an additional protein concentration step is taken and to the canonical bulk hair preparation method, single hair preparation efficacy was quantified via measurement of protein and peptide identifications, protein sequence coverage, and identification of GVP markers from combining output from mass spectra and exome sequence information into an automated workflow.

2.3.1 Single Hair Preparation Methods

Single hair fibers were prepared for mass spectrometry analysis using two different methods, an acidified liquid-liquid extraction (Single-LLE) and by acetone precipitation (Single-Acetone PPT), for comparison to the Bulk-HVP approach. Both single hair methods relied on protein extraction via ultrasonication followed by alkylation. To each one-inch single hair, four segments of approximately 6 mm were cut and placed into a Protein LoBind Eppendorf tube with 100 μL of denaturation buffer, which contained 2% (w/v) sodium dodecanoate, 50 mM ammonium bicarbonate, and 50 mM dithiothreitol; as discussed in the previous section, the detergent sodium dodecanoate was used to replace urea for protein extraction under heated conditions. Hair segments in buffer were placed into a water bath and ultrasonicated at 70 °C, 37 kHz, and 100% power until dissolution; on average, 2 h of ultrasonication ensured that the hair segments dissolved entirely. Protein extracts were then alkylated using iodoacetamide to prevent

reduced disulfide bonds from re-forming. Following alkylation, detergent was removed either via acidified liquid-liquid extraction or by acetone precipitation, as dodecanoate interferes with ionization during mass spectrometry acquisition.

To remove detergent, protein extracts in the Single-LLE sample set were acidified following alkylation with 0.75% (v/v) trifluoroacetic acid in 100 µL of ethyl acetate, mixed, and incubated for 15 min at RT at a 1:1 ratio of organic to aqueous solution volume. After phase separation via centrifugation at 15,000 × *g* for 10 min at RT, the organic top layer was removed. A thin layer of protein aggregate appeared at the interface of the top organic and bottom aqueous layers owing to phase separation and was not removed. Incubation and separation were repeated once to further remove detergent. The remaining extract was then made basic by addition of 10 µL of 1 M ammonium bicarbonate to achieve pH 8 for protein digestion. Alternatively, for detergent removal via acetone precipitation in the Single-Acetone PPT sample set, an aliquot of 400 µL of acetone chilled at -20 °C was added to each protein extract to allow formation of a protein pellet after overnight incubation at -20 °C. Supernatant was removed after phase separation via centrifugation at 15,000 × *g* for 15 min at RT, and protein pellets were washed with another aliquot of 400 µL chilled acetone. Protein pellets from acetone precipitation were resolubilized with a solution containing 0.01% (w/v) ProteaseMAX™ in 50 µL of 50 mM ammonium bicarbonate and placed on a shaker for 2 h prior to protein digestion.

Using 2 µL of 1 µg/µL trypsin (TPCK-treated) with magnetic stirring, protein digestion was allowed to incubate overnight at RT, with two additions of 2 µL of enzyme over a three-day incubation period for greater protein digestion efficiency. To inactivate the enzyme, formic acid was added for a final concentration of 0.1% in solution; supernatant was separated from any precipitate after acidification for filtration of protein digest. Protein digests were filtered through

centrifugal filter tubes (PVDF, 0.1 μm; MilliporeSigma, Burlington, MA) and filtrates were transferred to autosampler vials for mass spectrometry analysis.

2.3.2 Liquid Chromatography-Tandem Mass Spectrometry Analysis on a nano-LC-Orbitrap-MS

Protein digests were analyzed on an EASY-nLC 1200 system coupled to a Q Exactive Plus Orbitrap mass spectrometer (Thermo Fisher Scientific Inc., Waltham, MA). 1 μL injections were loaded onto an Acclaim™ PepMap™ 100 C18 trap (75 μm × 20 mm, 3 μm particle size), washed, and separated on an Easy-Spray™ C18 analytical column (50 μm × 150 mm, 2 μm particle size). Separations were performed at a flow rate of 300 nL/min using mobile phases A (0.1% formic acid in water) and B (0.1% formic acid in 90% acetonitrile/10% water) over a 107-min gradient: 2 to 3% B in 1 min, 3 to 11% B in 75 min, 11 to 39% B in 15 min, ramped to 100% B in 1 min, and held at 100% B for 15 min. Positive mode nano-electrospray ionization was achieved at a voltage of 1.9 kV. Full MS scans were acquired at a resolution of 70,000, with a maximum ion accumulation time of 30 ms, and a scan range between *m/z* 380 and 1800. Data-dependent MS/MS scans were triggered for the 10 most abundant ions at an intensity threshold of 3.3×10^4 and acquired at a resolution of 17,500, with a maximum ion accumulation time of 60 ms, dynamic exclusion of 24 s, and an isolation window of 2 Da. HCD fragmentation was performed at a collision energy setting of 27. Singly-charged species and ions with unassigned charge states were excluded from MS/MS.

2.3.3 Exome Sequencing and Genetically Variant Peptide Prediction

A detailed description of predicting GVPs from exome sequences is found in Mason et al.[15] Briefly, full exome sequencing was performed using DNA isolated from blood samples of individuals who provided a hair sample (ACE Research Exome with Secondary Analysis; Personalis Inc., Menlo Park, CA). Variant call format files from exome sequencing were then

filtered to include only missense variants (SNPs) from 691 genes commonly found in proteomic analyses. Sequence data quality of PASS or better (according to scoring system by the Genome Analysis Toolkit (Broad Institute)) was applied to filter the subset of SNPs. Conversion of sequence data to the current Genome Reference Consortium Human Build 38 (GRCh38) from GRCh37.5 coordinates was performed using the Bioconductor package Variant Annotation in R. Variant lists were further annotated using ENSEMBL's Variant Effects Predictor to include transcript, genetic mutation location, and corresponding amino acid substitution for each SNP. Human Genome Variation Society (HGVS) notation was used to specify SNPs in the absence of Reference SNP IDs. ENSEMBL Genome Browser transcripts associated with the subset of SNPs were downloaded using the R package BiomaRt and altered to reflect the genetic mutation. Original and altered transcripts were then used to create protein sequences, both with and without amino acid variants, in R. Mutated and their non-mutated counterpart protein sequences were combined and converted into FASTA files to be used as individualized protein databases for each subject.

2.3.4 Protein and Peptide Identification

Mass spectral data were imported into PEAKS Studio 8.5 (Bioinformatics Solutions Inc., Waterloo, ON, Canada) for peptide identification via *de novo* sequencing and subsequent database searching. Precursor ion mass tolerance was selected as ±20 ppm, while a mass error of 0.05 Da was allowed for fragment ions. A list of 313 potential post-translational modifications, including cysteine carbamidomethylation, methionine oxidation, and asparagine and glutamine deamidation, was allowed as variable modifications for peptide identification. The maximum number of PTMs allowed per peptide was three, and 3 tryptic missed cleavages on either end of the peptide were permitted. All *de novo*-sequenced peptides with a confidence score (-10lgP)

greater than 15% were matched to protein sequences in a reference database. To capture non-mutated proteins in the samples, the UniProtKB SwissProt Human protein database (downloaded September 21, 2017) was used for protein inference from identified peptides[20]. A second protein and peptide identification using the same raw mass spectral files was performed in PEAKS, where *de novo*-sequenced peptides were searched against individualized protein databases created from exome sequence data described in Section 2.3.3 above. The second PEAKS analysis enabled a focused search for proteins with expected mutations to identify GVPs in each sample. Each individualized protein database contains protein sequences from a list of 691 common gene products found in hair with the appropriate mutations expected in an individual based on their exome sequence. GVPs identified from each hair specimen were matched to mutated protein sequences in individualized protein databases in the second peptide identification analysis.

Identified proteins and peptides were further filtered with a 1% false discovery rate threshold for peptide-spectrum matches and then exported from PEAKS. An in-house Python-based script was applied to the output files to merge results from the two PEAKS analyses and generate a non-redundant protein profile for each sample. In particular, only peptide sequences attributed to a single gene product, or unique peptides, were retained and protein inferences were established using only unique peptides, thus preventing any inflation of protein inference by shared peptides. Protein profile metrics include the number of proteins, unique peptide sequences, amino acids, and SNPs identified from both major and minor GVPs, which are variant peptides from canonical, i.e., non-mutated, protein sequences found in conventional protein sequence databases such as UniProtKB SwissProt, and mutated protein sequences carrying amino acid polymorphisms, respectively.

2.3.5 Comparison of Hair Proteome Coverage

Single hair sample preparation methods demonstrated comparable protein extraction efficacy to the bulk volume method (Bulk-HVP). One-way ANOVAs, performed to compare protein, unique peptide, and amino acid yields among the three preparation methods, showed that these metrics did not statistically differ ($p \geq 0.127$; Table 2.7), indicating that either single hair preparation method provides similar extent of hair proteome coverage to that of Bulk-HVP (10.0 ± 2.0 mg of scalp hair), even with over 100-fold less material (on average, 84.4 ± 27.7 μg for a single inch of scalp hair). Though not statistically different, a protein concentration step (Single-Acetone PPT) allows slightly greater yields in protein and peptide identifications and amino acid coverage over the acidified ethyl acetate liquid-liquid extraction approach (Single-LLE).

Table 2.7. Numbers of identified proteins, unique peptides, and amino acids from three hair sample preparation methods (mean ± s.d.), with associated statistical significance from one-way ANOVAs. Both single hair preparation methods permit identification of similar numbers of proteins, peptides, and amino acids to those from bulk hair amounts, though acetone precipitation enables slightly greater yields overall.

Metric	Bulk-HVP	Single-LLE	Single-Acetone PPT	One-way ANOVA p-value
Proteins	238 ± 110	117 ± 46	179 ± 53	0.127
Unique Peptides	1,586 ± 686	827 ± 299	1,339 ± 375	0.131
Amino Acids	22,521 ± 10,021	12,063 ± 4,438	17,944 ± 5,595	0.169

Sequence coverage for each of the 470 detected proteins within the dataset was also examined to compare the composition of hair proteome coverage, which may vary owing to protein extraction and digestion efficacy differences among sample preparation methods. Shared peptides were included in the calculation of protein sequence coverage. Statistical comparisons among the three preparation methods found that of the 38 proteins that exhibit different extents of sequence coverage, 6 (16%) were from keratins, 3 (8%) from keratin-associated proteins (KAPs), and the majority, 29 (76%), from intracellular proteins (One-way ANOVA and Tukey

HSD; Appendix Table S-2.2). Greater sequence coverage was obtained in 88% and 85% of proteins when prepared using Bulk-HVP and Single-Acetone PPT methods, respectively, over the Single-LLE approach, with intracellular proteins comprising 82% and 59% of proteins, respectively. Single hair analysis with acetone precipitation enabled greater coverage of KAPs, up to 47% sequence coverage, and enabled coverage exceeding 90% in the majority of keratins, comparable in coverage attained with the Bulk-HVP method. Not only did the Single-Acetone PPT approach facilitate identification of intracellular proteins, but also improved sequence coverage of keratins and KAPs over the Single-LLE method, demonstrating similar performance to that of the bulk hair preparation method.

Although considered statistically comparable in performance to both the Bulk-HVP and Single-Acetone PPT methods, Single-LLE lags in overall peptide yields and protein sequence coverage, likely due to a variable efficacy in the liquid-liquid extraction process, particularly with manual removal of detergent present in the organic layer. While both single hair sample preparation methods incorporate a detergent removal step, either through liquid-liquid extraction or supernatant removal from the precipitated protein pellet, the primary advantage to the latter method is the ease of detergent removal, i.e., extraction into acetone and near-complete removal of supernatant distinct from the protein pellet, compared to the first method, in which separation of the two liquid phases in the presence of protein aggregate at the interface of the organic and aqueous layers was sometimes nebulous. Furthermore, acetone precipitation serves dual purposes since proteins are concentrated in the same process as detergent removal, which is an additional advantage when working with minute amounts of sample as in a single one-inch hair. As such, incorporating acetone precipitation in preparation of single hair fibers benefits peptide

identification and is expected to be advantageous for GVP identification, which is discussed in Section 2.3.6.

2.3.6 GVP Identification from Untargeted Mass Spectrometry Analyses

To identify GVPs in each hair sample, the list of identified peptides was queried with a list of known missense SNPs. This list of missense variants was produced from exome sequencing of individuals' DNA and served as the basis for generating individualized mutated databases described in Section 2.3.3. Only two genotypes are reported in the variant list: the heterozygous genotype and the homozygous genotype for the alternate, or minor, allele; the homozygous genotype for the reference, or major, allele is not included. An in-house Python script was written to compare and match the location and amino acid consequence of the polymorphisms with the list of identified peptides for GVP identification. For example, the peptide DLNMDC**M**VAEIK from K83 (located at positions 273 – 284) successfully matched as a minor GVP corresponding to the SNP rs2852464 from gene KRT83 (mutation site in peptide denoted in larger, bold red text), which manifests as the SAP I279M. This process was performed for both the reference and alternate alleles regardless of variant genotype; false positives are removed in a later process.

However, because the homozygous genotype for reference alleles, hereafter referred to as the homozygous-major genotype, is not included in the missense variant list, a second GVP identification process specifically for the homozygous-major variants is needed. The list of homozygous-major variants differs from individual to individual given the possible genotypes. Without this component, non-detection of major GVPs corresponding to the relevant SNP could be misconstrued as a false negative rather than an incomplete analysis for variant detection. This second analysis utilizes individualized variant lists based on the list of SNPs inferred from the

entire dataset, tailored to include only SNPs associated with homozygous-major genotypes, and matched to identified peptide lists as described above. For example, Individual 2 exhibits the homozygous-major genotype for SNP rs34302939 in protein KAP10-12 whereas Individuals 1 and 3 are heterozygotes based on their exome sequences; therefore, this SNP is not annotated in the variant list for Individual 2. Consistent with this definition of the genotypes, the major GVPs for this SNP were identified, after querying the peptide lists with the respective variant list of missense SNPs, in the protein digests from only Individuals 1 and 3, but not Individual 2. Once detected in at least one sample, in this case, the protein extracts from Individuals 1 and 3, the location and amino acid consequence corresponding to the homozygous-major genotype for this SNP was added to the variant list for Individual 2 for the second phase of GVP identification. The second targeted query using a variant list for specific SNPs successfully identified major GVPs corresponding to the SNP rs34302939 from KRTAP10-12 in all three protein digests from Individual 2. Through this process, the number of SNPs inferred from detection of major GVPs doubled, from 5 ± 3 to 10 ± 5 (repeated measures t-test; $p = 4.08 \times 10^{-4}$; Figure 2.3). Not only does inclusion of homozygous-major variants in GVP identification improve the number of inferred SNPs, but this approach also completes GVP phenotype observations for each hair sample necessary for quantifying discriminative potential, which is discussed in Section 2.3.7.

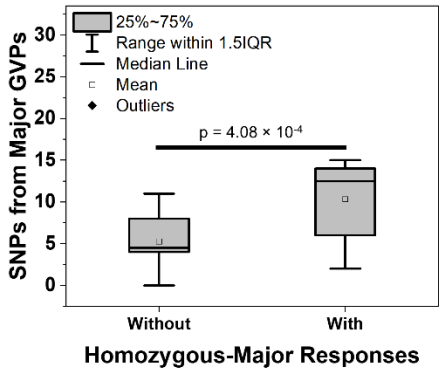


Figure 2.3. Comparison of the numbers of SNPs inferred from major GVPs without and with addition of homozygous-major responses. Inclusion of homozygous-major responses permits identification of statistically greater numbers of SNPs from major GVPs (repeated measures t-test; n = 12 per condition).

Comparison of the numbers of SNPs identified from major and minor GVPs via an untargeted mass spectrometry approach showed similar performance in hair samples prepared with either the Bulk-HVP and Single-Acetone PPT methods, but also strikingly large variability in successful GVP detection for SNP inference among the three sample preparation methods (Table 2.8). However, this variability, represented by coefficient of variation, ranged between 26% and 46%, which is comparable to the variability in numbers of identified proteins and peptides among the three preparation methods (Table 2.7), except for the number of SNPs identified from major GVPs using the Single-LLE method, which yielded 75% variability. It is likely that the large ranges of identified SNPs within each preparation method derive from a combination of biological variation among the 4 individuals and some irreproducibility in GVP identification during data-dependent mass spectrometry analysis owing to peptide ion competition for MS/MS fragmentation. Protein content and ease of protein extraction with slightly different hair physicochemical properties among individuals can contribute to varying GVP detection success rates observed here. Further, the data-dependent approach limits the number of peptide ions that undergo fragmentation to the 10 most abundant per survey scan; the hair matrix is sufficiently complex that the same peptide ions from protein digests may not be

selected for MS/MS fragmentation. The route by which peptides are identified in a bottom-up proteomics approach varies, from run to run, even among hair samples from the same individual or technical replicate analyses of the same sample. As an alternative, data-independent mass spectrometry methods, which can be more sensitive with higher signal-to-noise ratios, are expected to minimize variability in GVP identification from peptide ion competition, as this approach can provide targeted analysis of pre-selected precursor peptide ions, either as specific *m/z* values or ranges, for fragmentation. And though not statistically significant, SNP yields from hair samples prepared by the Single-LLE method (7 ± 5 and 5 ± 2 SNPs from major and minor GVPs, respectively) were lower than those prepared by the Bulk-HVP and Single-Acetone PPT approaches, demonstrating that a protein concentration step in single hair analysis is beneficial to attain comparable GVP detection success when hair samples are mass-limited. With development of this workflow for GVP marker identification, GVP profiles can subsequently be generated from these aggregate SNP numbers to quantify and compare discriminative potential from each single hair sample.

Table 2.8. Numbers of SNPs from major and minor GVPs (mean ± s.d.) annotated for untargeted proteome analysis using three different sample preparation protocols, with associated statistical significance from one-way ANOVAs (n = 4 individuals per preparation method). Large variability in SNP identification within each sample preparation method is attributed to biological variation among individuals and variation in mass spectrometry analysis. All sample preparation methods yield statistically similar numbers of identified SNPs, though acetone precipitation results in more comparable yields to bulk amounts than does the liquid-liquid extraction method.

SNPs	Bulk-HVP	Single-LLE	Single-Acetone PPT	One-way ANOVA p-value
from Major GVPs	12 ± 3	7 ± 5	12 ± 4	0.172
from Minor GVPs	11 ± 5	5 ± 2	13 ± 5	0.078

2.3.7 GVP Profile Generation and Evaluation of Discriminative Potential

Within the dataset comprising 12 hair samples from 4 individuals prepared by three different sample preparation methods, 40 SNPs were identified for generating a GVP profile for each hair sample that allows comparison of discriminative potential among the four individuals, after removing SNPs with false positive responses. False positive responses (i.e., detection of GVPs in an individual's hair sample when the corresponding SNP is not detected in the exome sequence from an individual's DNA) arose primarily from false detection of the major GVP when the individual exhibited a homozygous genotype for the SNP; for this SNP genotype, only the minor GVP should be detected. These SNPs were removed from the analysis and not considered further as potential markers. Two SNPs that were identified by exome sequence analysis but are not documented in the dbSNP database[27] were also removed; it is likely that these SNPs occur so infrequently in the population that they have not yet been added to the dbSNP database. As they are not well-represented in the population and are not associated with any reported allele frequencies, rare SNPs do not enhance distinction of GVP profiles for the vast majority of individuals, and thus, were excluded from the GVP panel.

To generate GVP profiles and quantify discriminative potential from the presence and absence of major and minor GVPs, population frequencies corresponding to the SNP genotypes must be known or estimated. Such data are accessible within the Genome Aggregation Database (gnomAD)[28, 29], which are shared by investigators who contribute genomic or exomic data from experimental cohorts. Publicly available data include allele frequencies and the number of homozygotes delineated by ancestry and biological sex. Genomic and exomic sequence data quality are also reported, based upon whether the data pass a random forest test. SNPs whose population frequency data do not pass a quality check with a random forest model were not

considered further since failure indicates that the variant is an artifact and not a true genetic variant[28]. Using allele frequency data from gnomAD, population frequencies at each SNP locus can be calculated in two manners: through genotype observations from allele frequency and homozygote data or estimated from allele frequencies by assuming Hardy-Weinberg equilibrium.

Hardy-Weinberg equilibrium is sometimes assumed in forensic analyses for quantifying discriminative power for DNA evidence. However, the following must hold true for a population to conform to Hardy-Weinberg equilibrium[30]:

1. The population is infinitely large,
2. Absence of migration into and out of the population,
3. Absence of mutation,
4. Absence of natural selection, and
5. Random mating occurs.

These assumptions permit calculation of population frequencies based solely on allele frequencies, without homozygote data, which may be viewed as a simpler approach to determine population frequencies, and perhaps a necessary method for calculating population frequencies if homozygote data are not available. The two methods for determining population frequencies are compared here for a set of SNPs to determine whether Hardy-Weinberg equilibrium can be assumed for this population without further evaluation of population structure, when there is the option to use either method. For bi-allelic systems with one reference allele, p, and one alternate allele, q, population frequencies f for each genotype were determined given n_{total}, the total number of alleles at the locus sampled over a population with size $\frac{n_{total}}{2}$, n_q, the number of alternate alleles at the locus, and h_{qq}, the number of homozygotes with genotype qq, using Equations 2.1 – 2.6:

$$f_{obs,pp} = \frac{n_{total} - 2n_q + 2h_{qq}}{n_{total}}, \quad \text{Eq. 2.1}$$

$$f_{obs,pq} = \frac{2n_q - 4h_{qq}}{n_{total}}, \quad \text{Eq. 2.2}$$

$$f_{obs,qq} = \frac{2h_{qq}}{n_{total}}, \quad \text{Eq. 2.3}$$

$$f_{HWE,pp} = \left(\frac{n_{total} - n_q}{n_{total}}\right)^2, \quad \text{Eq. 2.4}$$

$$f_{HWE,pq} = \frac{2n_q n_{total} - 2n_q^2}{n_{total}^2}, \quad \text{Eq. 2.5}$$

and

$$f_{HWE,qq} = \left(\frac{n_q}{n_{total}}\right)^2, \quad \text{Eq. 2.6}$$

where $f_{obs,pp}$, $f_{obs,pq}$, and $f_{obs,qq}$ represent the observed population frequencies for genotypes pp, pq, and qq, respectively, and $f_{HWE,pp}$, $f_{HWE,pq}$, and $f_{HWE,qq}$ are the calculated population frequencies for genotypes pp, pq, and qq, respectively, under Hardy-Weinberg equilibrium conditions.

To determine whether genotypes for the 26 bi-allelic SNPs conform to Hardy-Weinberg equilibrium (HWE) conditions, a X^2 goodness-of-fit test was performed using the *HWChisq* function in the *HardyWeinberg* package[31] (R x64 version 3.4.4) for each SNP and evaluated for the global population frequencies and the non-Finnish European (NFE) population frequencies. The global population in gnomAD v2.1.1 consists of cohorts among the non-Finnish European, Finnish European, East Asian, South Asian, Latino, African, Ashkenazi Jewish, and Other ancestries. The non-Finnish European population was singled out for comparison to the global population as the hair samples in this dataset are known to originate from non-Finnish Europeans. Figure 2.4 displays the results of goodness-of-fit statistical testing for each bi-allelic

SNP against two metrics: HWE deviation for heterozygotes and inbreeding coefficient. Briefly, HWE deviation for heterozygotes (*D*) was determined using the formula:

$$D = \frac{n_{total}}{4}\left(f_{obs,pq} - f_{HWE,pq}\right), \qquad \text{Eq. 2.7}$$

which is adapted from Graffelman's work[31] using the variables presented herein (Eqs. 2.2 and 2.6), and inbreeding coefficient is the probability that a pair of alleles at a locus is identical (homozygous genotype) as both alleles are inherited from one ancestor[31]. A negative HWE deviation indicates low heterozygosity, associated with little genetic variability and often attributed to inbreeding; this is similarly shown with positive inbreeding coefficients[32]. X^2 values for most global population frequencies (96%) for bi-allelic SNPs are above the critical value (Figure 2.4a), indicating that the vast majority of global population frequencies for bi-allelic SNPs deviate significantly from HWE, whereas their NFE counterparts align well with HWE conditions (27% deviation). The positive inbreeding coefficients obtained by using global population frequencies for these bi-allelic SNPs, which exhibit X^2 values greater than the critical value (Figure 2.4b), also reflect a departure from HWE. Aggregated homozygotes and/or homozygotes-major outnumber heterozygotes when considering global population frequencies, likely because there are populations among the 8 ancestries that deviate from Hardy-Weinberg equilibrium. Lack of diversity in populations not well sampled may contribute to an overall departure from HWE. On the other hand, population frequencies for a few SNPs showed significantly positive deviation from HWE and negative inbreeding coefficients, which may be attributed to bad mapping of genetic coordinates during alignment of DNA sequence reads, suggesting that the affected SNPs do not belong in the designated chromosomal regions. Nevertheless, a substantial number of bi-allelic SNPs show departure from HWE when considering global population frequencies. When applied to forensic evidence where the

ancestral source may not be known, use of global population frequencies allows calculation of discriminative potential that better generalizes to the overall population and avoids incorrect assumptions to any ancestral subset. As such, global population frequencies should be used without assumption of HWE conditions whenever possible.

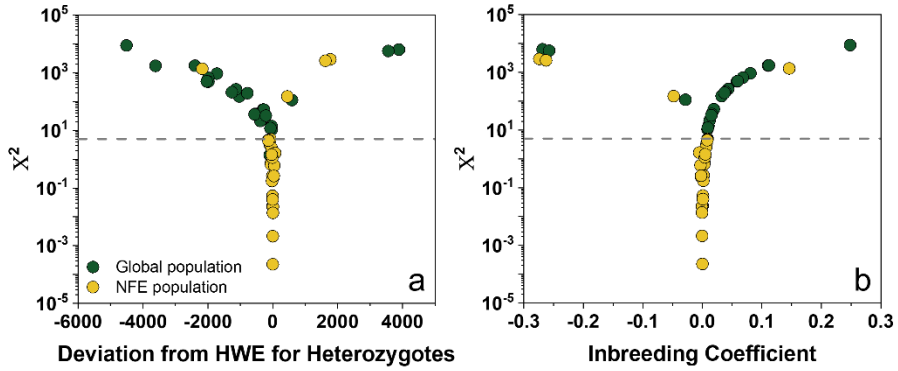


Figure 2.4. Chi-square (X^2) statistic as a function of the (a) deviation from Hardy-Weinberg equilibrium for the heterozygous genotype and (b) inbreeding coefficient for each bi-allelic SNP identified in this dataset (n = 26 SNPs, X^2 goodness-of-fit test), derived from allele frequency data in gnomAD. The dashed line denotes the Chi-square critical value at $\alpha = 0.05$. Statistical significance was evaluated for each SNP using both the global (green circle), or total, population frequencies observed across all measured ancestral populations and the population frequencies tabulated for non-Finnish Europeans (yellow circle). Global population frequencies deviate significantly from Hardy-Weinberg equilibrium, resulting from fewer than expected heterozygotes when considering the total sampled population among the various ancestries. Inbreeding coefficients also illustrate low heterozygosity, inversely from (a), for the majority of global population frequencies. As such, Hardy-Weinberg equilibrium cannot be assumed when using global population frequencies at each SNP locus.

Although global population frequencies can be readily calculated for bi-allelic SNPs, data provided in gnomAD for multi-allelic SNPs are often incomplete for similar calculations and thus require assumption of HWE conditions. While some of the minor alleles occur with low frequency in the current sampled population in comparison to the reference allele and the more common alternate allele, these frequencies are likely to change and may not be negligible with more sampling of genetic data and better representation of the global population. As such, it is useful to include these minor alleles when determining population frequencies for multi-allelic

SNPs. For example, with three alleles, p, q, r, where p is the reference, or major, allele and q and r are the alternate alleles, that yield six genotypes pp, pq, qq, pr, qr, and rr, and given n_{total}, n_q, and n_r, the number of alleles in total and for the q and r alleles, respectively, and h_{qq} and h_{rr}, the number of homozygotes for genotypes qq and rr, respectively, the following Equations 2.8 – 2.11 hold true:

$$n_q = 2h_{qq} + h_{pq} + h_{qr}, \quad \text{Eq. 2.8}$$

$$n_r = 2h_{rr} + h_{pr} + h_{qr}, \quad \text{Eq. 2.9}$$

$$n_{total} - n_q - n_r = 2h_{pp} + h_{pq} + h_{pr}, \quad \text{Eq. 2.10}$$

and

$$h_{total} = \frac{n_{total}}{2}, \quad \text{Eq. 2.11}$$

where h_{pp}, h_{pq}, h_{pr}, and h_{qr} are the number of individuals with genotypes pp, pq, pr, and qr, respectively, and h_{total} represents the total number of individuals. However, there is insufficient information regarding heterozygotes when determining h_{pp} and h_{pq}, as shown in Equations 2.12 and 2.13:

$$h_{pp} = \frac{n_{total} - 2n_q + 2h_{qq} - 2h_{rr} - 2h_{pr}}{2} \quad \text{Eq. 2.12}$$

and

$$h_{pq} = n_q - n_r - 2h_{qq} + 2h_{rr} + h_{pr}, \quad \text{Eq. 2.13}$$

because h_{pr} remains an unknown variable in both equations. As such, $f_{obs,pp}$ and $f_{obs,pq}$ cannot be determined; to estimate f_{pp} and f_{pq}, Hardy-Weinberg equilibrium must be assumed using:

$$f_{HWE,multi,pp} = \left(\frac{n_{total} - n_q - \sum_{i=1}^{x} n_i}{n_{total}}\right)^2 \quad \text{Eq. 2.14}$$

and

$$f_{HWE,multi,pq} = \frac{2n_q(n_{total} - n_q - \sum_{i=1}^{x} n_i)}{n_{total}^2}, \quad \text{Eq. 2.15}$$

where x is the total number of alternate alleles excluding the q allele. f_{qq} is estimated from Equation 2.6 above to maintain consistency with the other calculated locus frequencies, despite sufficient information to calculate the value based on empirical observations. On average, among the 12 multi-allelic SNPs, $f_{obs,qq}$ and $f_{HWE,qq}$ differ by 25% using global population frequencies, with larger deviations resulting from excess homozygosity, though HWE is a necessary assumption for multi-allelic SNPs in the absence of more comprehensive genotype data for a large population such as that curated in gnomAD.

In sum, for bi-allelic SNPs, calculation of population frequencies at each SNP locus for each genotype utilizes allele frequency and homozygote data reported in gnomAD without assumption of HWE, but for multi-allelic SNPs, population frequencies at each locus are determined from allele frequencies alone by assuming Hardy-Weinberg equilibrium. Hereafter, population frequencies are generalized to genotype frequencies regardless of method of calculation.

Use of frequencies as a representation of genotypes or phenotypes from detected markers differ between SNPs and GVPs due to inherent differences in how typing is derived from marker detection. The use of genotype frequencies f_{pp}, f_{pq}, and f_{qq} assumes complete detection of major and minor variants, or for SNPs specifically, reference and alternate alleles. For SNP identification such as from exome sequencing, 2 alleles are typically detected; detection of 2 reference alleles implies a homozygous-major SNP genotype and detection of 1 each of reference and alternate alleles indicates a heterozygous genotype. But this scheme is not applicable for inferring the homozygous and homozygous-major genotypes in a proteomics experiment, as there may only be 1 variant detected and detection of only 1 variant may then result in some ambiguity in genotype. The 0,0 (homozygous-major) genotype detected during

exome sequencing parallels the detection of the major GVP, represented by the phenotype designated as 0, while detection of the minor GVP is represented by the phenotype 1 as the counterpart for the 1,1 (homozygous) genotype. However, detection of either the 0 or 1 phenotypes is also possible for the 0,1 (heterozygous) genotype when GVP detection is incomplete (i.e., the other variant is not detected), which may occur more often with the data-dependent mass spectrometry analysis utilized in this work owing to incomplete selection of GVP precursor ions in the survey scan for MS/MS spectrum generation. To account for ambiguity in genotype inference at any SNP locus given GVP responses, population frequencies for the various phenotypes, as detected in a proteomics experiment, are determined as listed in Table 2.9. In cases where GVPs are not detected, i.e., '--' responses under "Proteomics – Observed Phenotype", the phenotype frequency is 1 so as to remove any influence of non-detected GVPs on downstream quantification. These summations represent a conservative approach to calculate SNP locus frequencies that will be used to quantify discriminative potential.

Table 2.9. Method for calculation of population frequencies at each SNP locus based on true detection of observed phenotypes from proteomics experiments. 0 and 1 represent the presence of the major and minor allele or GVP, respectively, and '--' denotes the absence of variants.

Exome – Expected Genotype	**Proteomics – Observed Phenotype**	**Genotype Frequency**	**Phenotype Frequency**
0,0	0	f_{pp}	$f_{pp} + f_{pq}$
0,0	--		1
0,1	0	f_{pq}	$f_{pp} + f_{pq}$
0,1	1		$f_{pq} + f_{qq}$
0,1	0,1		f_{pq}
0,1	--		1
1,1	1	f_{qq}	$f_{pq} + f_{qq}$
1,1	--		1

While routinely used as a method to evaluate discriminative power, random match probability, as the product of SNP loci frequencies, assumes SNP loci independence. However, its use may be complicated by SNP co-inheritance within population structure, also known as linkage disequilibrium, the non-random association of two SNPs at different loci that are inherited from a single, ancestral chromosome[33]. As such, their genotypes are not encountered independent of each other in a population and usually occur at higher than expected frequencies[34]. Disequilibria may not be limited to proximal regions around more frequently occurring SNPs and can extend farther than 100 kilobases in distance between linked SNPs[35]. To minimize chances of linkage between SNP pairs, a one-SNP-per-gene rule was adopted, and in instances where multiple SNPs from the same gene were identified, the SNP that yielded the most consistent response among samples in accordance with exome genotypes was selected. Thus, from a selection of 38 identified SNPs, 27 remained for profiling and quantifying discriminative power.

Comparison of GVP profiles pairwise showed large intraindividual variation, predominantly derived from variability in SNP identification among the three sample preparation methods. Figure 2.5a represents a simplified GVP profile from each hair sample, based on the phenotype frequencies associated with the presence of the combination of major and minor GVPs; Appendix Table S-2.3 displays the full GVP responses. On average, 8 ± 2 differences in inferred SNP genotypes between profiles from the same individual prepared by different sample preparation methods were observed, not significantly different from any GVP profile differences when comparing between different individuals, on average, 11 ± 3 differences (Kruskal-Wallis and Dunn post-hoc tests; $p \geq 0.026$, where statistical significance was established at $p \leq 0.025$; Figure 2.5b). Notably, intraindividual responses for SNPs from keratins and KAPs were more

consistent than those from intracellular proteins, especially between samples prepared with the Bulk-HVP and Single-Acetone PPT methods. These two methods also yielded higher protein sequence coverage for keratins and KAPs than the Single-LLE preparation, as discussed in Section 2.3.5; greater protein sequence coverage increases the chances of detecting GVPs from a protein. In contrast, hair samples prepared with the Single-LLE method exhibited many more GVP non-detects, with overall fewer identified proteins and peptides, and poorer hair proteome coverage. These observations indicate that sample preparation method has a large influence on GVP detection; application of single hair analysis to focused studies on hair proteome variation where hair sample preparation methods do not vary will minimize intraindividual variation, permitting greater similarity in GVP profiles among sample replicates.

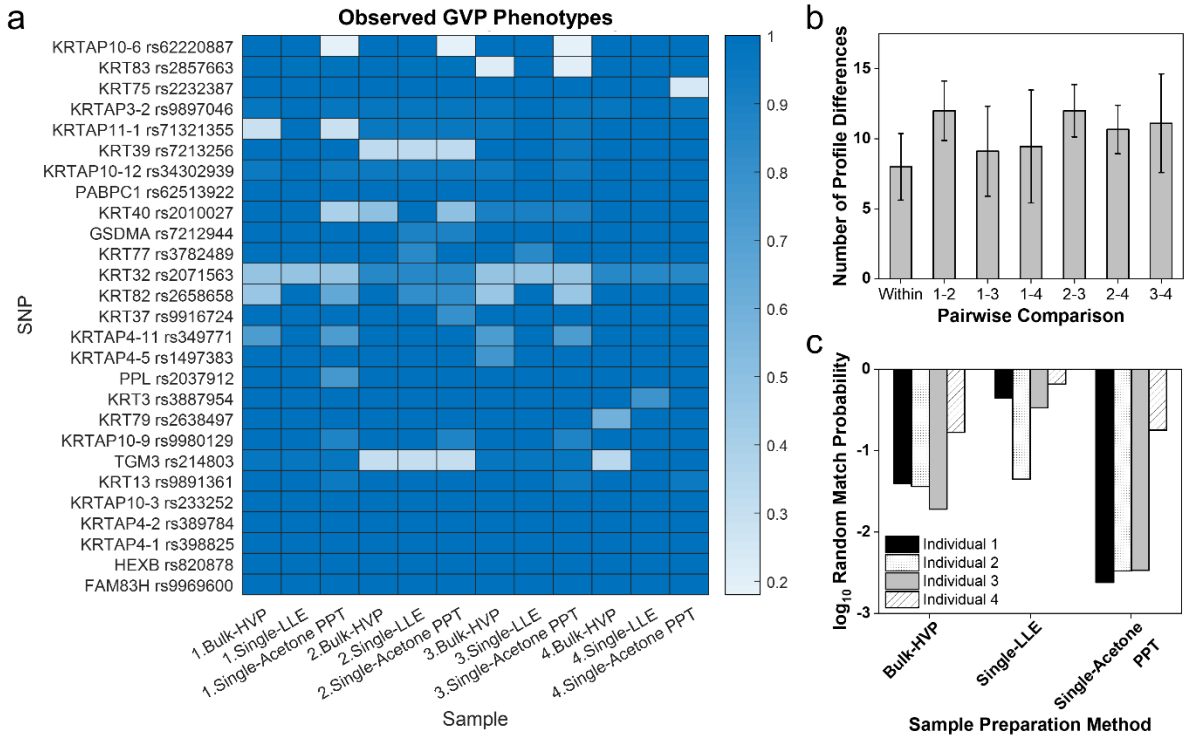


Figure 2.5. (a) GVP profile assembled from SNPs identified in hair samples prepared by three different methods, (b) number of profile differences when comparing GVP profiles pairwise from the same individual (Within) and between individuals (1-2, 1-3, 1-4, 2-3, 2-4, and 3-4), among the three sample preparation methods, and (c) random match probabilities from corresponding GVP profiles. Samples in (a) are denoted x.y, where x is the individual code (of 4 individuals) and y is the hair sample preparation method. Error bars in (b) represent the standard deviation.

As with GVP profiles, discriminative power, evaluated using random match probabilities, varies among hair samples from the same individual depending on sample preparation method (Figure 2.5c). Not surprisingly, lowest discriminative power, ranging from 1 in 2 to 1 in 22, was observed from hair samples prepared using the Single-LLE method, and samples prepared with the Single-Acetone PPT approach attained the highest discriminative power (between 1 in 6 and 1 in 418), owing to more GVP non-detects, from both major and minor GVPs, with the first method. RMPs achieved with the Single-Acetone PPT preparation were slightly higher than those from the Bulk-HVP method, that is, within an order of magnitude (ranging from 1 in 6 to 1 in 52; Figure 2.5c), likely a result of identifying just a slightly greater number of SNPs from

minor GVPs (on average, 13 ± 5 compared to 11 ± 5 SNPs, from Table 2.8), as the presence of minor GVPs typically correlates with lower genotype frequencies at each SNP locus and therefore, lower RMPs and higher discriminative power when aggregated. Among the individuals, Individual 4 exhibits the most common profile while the other three individuals carry SNPs that enable distinction from one another, though presence of the minor GVP for SNP rs2232387 from KRT75, identified in a single one-inch hair sample prepared by acetone precipitation, permits differentiation from the other individuals. Identification of both the major and minor GVPs for SNP rs71321355 from KRTAP11-1 enables distinction of Individual 1 from Individuals 2 and 3. Though this genotype is shared with Individual 4, GVP non-detection of both the major and minor variants, perhaps due to the difficulty with which proteins are extracted from this hair matrix in general, results in omission of this SNP for quantification of discriminative power for this individual. Detection of the minor GVP corresponding to the SNP rs2857663 from KRT83 in hair samples differentiates Individual 3 in this dataset, as the other three individuals exhibit a homozygous-major genotype, and the heterozygous genotype for SNP rs214803 from TGM3 distinguishes Individual 2 from the other subjects, who are homozygotes for either the major or minor variant. While there exists variability in GVP detection which affects discriminative power with the current analytical scheme, it is expected that GVP profiling used in routine operation when acquired with targeted mass spectrometry approaches will minimize this variability for improved differentiative potential.

2.4 Conclusions

This work establishes an informatics-based workflow for GVP identification, from hair sample preparation to GVP discovery to profiling, that integrates mass spectral data and exome sequence information for single hair analysis. The strategies presented herein directly enable

examination of intrinsic variation in and effects of external exposures on the hair proteome and GVP marker identification, all of which encounter restriction of analysis to mass-limited hair samples and are discussed in the next chapters.

Not limited to GVP identification from hair proteins, the optimization process and experimental considerations discussed in this work extend application to other disciplines. Comparison of sample preparation approaches and selection of appropriate search engine parameters for peptide identification represent general concerns in proteomics analyses and should undergo critical evaluation to optimize for analytes of interest, such as membrane proteins, another class of challenging proteins, depending on analytical needs. Exome sequence-driven GVP identification outlines a focused, computationally economical approach for discovery of novel biomarkers, applicable to the biomedical sciences, where the untargeted detection scheme is refined by comparing to an established ground truth for validation purposes. And finally, the profiling process demonstrates a potential method for interpreting findings from GVP analysis when performed routinely in forensic identification.

APPENDIX

Table S-2.1. Numbers of identified proteins and peptides from GPM and PEAKS searches for each bulk volume hair sample across the six different sample preparation methods (n = 4 hair samples per condition). A non-redundant set of peptide sequences was tabulated from raw peptide numbers from PEAKS Set P12 for direct comparison to the number of peptide sequences reported in GPM Set G8.

Sample Preparation Method	GPM Set G8		PEAKS Set P12		
	Proteins	Peptides	Proteins	Peptides	Non-Redundant Peptides
HVP	90	379	125	643	464
	122	600	174	995	765
	132	655	205	1093	830
	111	452	145	758	570
HVP Acetone PPT	82	354	115	532	433
	124	490	128	958	614
	117	402	140	969	520
	81	312	113	570	473
HVP Acetone PPT-Sonication	172	584	262	1165	772
	224	897	315	1599	1189
	225	418	248	1343	960
	192	308	283	1060	722
HVP SDD-Acetone PPT	62	198	77	268	234
	80	234	124	672	452
	68	203	94	351	284
	72	217	99	348	285
HVP SDD-Acetone PPT-Sonication	154	494	186	982	632
	166	633	258	1192	843
	160	612	212	1144	809
	154	512	285	1011	672
HVP SDD-Acetone PPT-Sonication2X	148	473	202	990	633
	142	617	256	1240	840
	118	471	175	910	651
	85	344	105	727	453

Table S-2.2. Average protein sequence coverage for each hair sample preparation method (n = 4 per method). One-way ANOVAs and Tukey HSD post-hoc tests were performed to determine statistical significance. '--' indicates that the post-hoc test was not performed, as results from the one-way ANOVA were not statistically significant.

Gene	Sample Preparation Method Sequence Coverage Average (%)			One-way ANOVA p-Value	Tukey HSD Post-Hoc p-Values		
	Bulk-HVP	Single-LLE	Single-Acetone PPT		Single-Acetone PPT – Bulk-HVP	Single-LLE – Bulk-HVP	Single-LLE – Single-Acetone PPT
KRT31	95.7	84.4	94.7	0.112	--	--	--
KRT39	80.8	51.2	87.6	0.022	0.816	0.061	0.024
KRT34	84.3	74.3	86.4	0.049	0.893	0.111	0.055
KRT35	89.3	75.5	91.4	0.056	--	--	--
KRT33A	99.9	88.2	97.2	0.019	0.726	0.020	0.066
LGALS7	71.5	50.7	72.6	0.152	--	--	--
KRT85	98.8	92.9	98.4	0.006	0.973	0.010	0.014
KRT82	75.5	53.4	65.7	0.333	--	--	--
SELENBP1	69.6	43.3	62.8	0.139	--	--	--
CALML3	57.4	51.7	66.9	0.589	--	--	--
KRT32	74.0	62.3	68.8	0.249	--	--	--
SFN	71.0	43.4	65.2	0.067	--	--	--
KRT33B	95.7	84.6	96.7	0.022	0.965	0.045	0.030
PKP1	60.9	36.4	49.6	0.024	0.302	0.019	0.214
HIST1H4A	49.0	47.3	48.1	0.984	--	--	--
DSP	59.1	25.9	45.7	0.024	0.400	0.020	0.164
KRT36	50.2	47.9	52.4	0.724	--	--	--
DUSP14	40.2	26.8	47.1	0.166	--	--	--
VDAC2	31.7	32.3	40.6	0.829	--	--	--
JUP	43.0	25.0	42.5	0.236	--	--	--
LGALS3	32.2	27.2	31.3	0.736	--	--	--
DSG4	40.7	19.2	26.7	4.57×10^{-4}	0.007	3.78×10^{-4}	0.127
KRT80	21.4	13.7	47.0	0.017	0.062	0.709	0.018
KRT83	91.8	87.6	94.4	0.227	--	--	--
PPIA	41.1	20.5	29.4	0.360	--	--	--
LAP3	28.0	16.9	27.3	0.636	--	--	--
KRT40	30.6	20.0	42.5	0.142	--	--	--
KRT86	99.7	97.2	99.6	0.090	--	--	--
HSPB1	36.0	11.6	23.5	0.053	--	--	--
YWHAE	35.0	21.6	34.2	0.271	--	--	--
TGM3	31.5	13.1	23.4	0.319	--	--	--
PKP3	36.0	11.7	19.0	0.052	--	--	--
KRT84	21.6	20.3	22.4	0.753	--	--	--
ATP5B	22.1	9.1	17.2	0.270	--	--	--
KRT87P	82.9	83.6	85.9	0.550	--	--	--
KRT81	91.2	87.9	92.9	0.284	--	--	--
KRT38	46.8	27.5	42.0	0.129	--	--	--
YWHAZ	26.7	15.8	21.0	0.342	--	--	--
HEPHL1	12.1	8.8	16.7	0.658	--	--	--
LMNA	18.6	5.9	10.2	0.147	--	--	--
CTNNB1	15.5	8.9	12.9	0.320	--	--	--
KRT31	12.8	13.1	9.7	0.773	--	--	--
KRT5	7.5	11.0	13.8	0.123	--	--	--

Table S-2.2 (cont'd)

Gene	Sample Preparation Method Sequence Coverage Average (%)			One-way ANOVA p-Value	Tukey HSD Post-Hoc p-Values		
	Bulk-HVP	Single-LLE	Single-Acetone PPT		Single-Acetone PPT – Bulk-HVP	Single-LLE – Bulk-HVP	Single-LLE – Single-Acetone PPT
KRT75	95.7	84.4	94.7	0.112	--	--	--
PPL	8.3	1.2	1.7	0.012	0.026	0.018	0.971
CUX2	0.6	0.5	0.8	0.247	--	--	--
ATP5A1	37.2	11.4	19.7	0.037	0.150	0.033	0.605
FAM26D	26.0	8.8	22.1	0.143	--	--	--
TRIM29	28.7	4.9	20.3	0.032	0.524	0.027	0.156
NEU2	26.9	10.3	12.6	0.276	--	--	--
KRTAP16-1	9.5	7.4	20.7	0.244	--	--	--
ANXA2	39.2	23.9	39.7	0.503	--	--	--
LAMP1	5.2	3.2	9.7	0.008	0.047	0.450	0.007
HSPA5	13.3	5.1	5.3	0.196	--	--	--
HSPA2	20.7	7.6	15.8	0.090	--	--	--
KRT37	33.2	17.0	32.0	0.104	--	--	--
KRTAP11-1	47.2	13.0	47.4	0.022	1.000	0.036	0.035
VSIG8	36.7	28.2	38.7	0.825	--	--	--
LRRC15	22.6	12.4	22.2	0.616	--	--	--
KRTAP1-5	14.9	21.7	23.7	0.749	--	--	--
KRTAP3-3	62.0	21.9	62.0	0.155	--	--	--
KRT10	19.8	28.4	19.1	0.828	--	--	--
HSD17B10	14.2	11.8	7.2	0.539	--	--	--
HNRNPA1	17.0	11.6	6.9	0.244	--	--	--
ALDH2	13.7	7.0	6.5	0.431	--	--	--
CRYBG1	3.7	1.1	4.7	0.162	--	--	--
KRT19	7.0	7.5	5.8	0.826	--	--	--
TUBB2A	44.9	12.1	42.1	0.019	0.961	0.026	0.040
S100A3	44.3	29.5	59.2	0.591	--	--	--
GAPDH	38.4	19.0	31.0	0.597	--	--	--
MIF	43.0	4.3	27.0	0.114	--	--	--
FABP4	25.9	20.8	31.3	0.847	--	--	--
KRT1	19.7	36.8	15.6	0.452	--	--	--
CALML5	21.1	18.5	23.6	0.925	--	--	--
KRTAP3-2	61.5	19.9	60.2	0.174	--	--	--
PRDX6	18.2	14.1	18.5	0.919	--	--	--
SERPINB5	19.7	11.0	19.3	0.776	--	--	--
KRTAP24-1	21.3	6.2	16.3	0.257	--	--	--
MDH2	25.5	5.0	10.7	0.116	--	--	--
TUBA4A	36.8	10.9	29.1	0.124	--	--	--
PRSS1	5.1	2.7	11.5	0.014	0.061	0.620	0.014
HEXB	8.5	2.2	4.9	0.253	--	--	--
HSD17B4	7.5	2.7	4.1	0.313	--	--	--
PLEC	11.1	1.8	2.9	0.309	--	--	--
GFAP	2.1	5.6	3.3	0.132	--	--	--
PLB1	3.0	1.1	0.6	0.096	--	--	--
POTEF	5.8	2.6	4.0	0.222	--	--	--
AHNAK	0.8	0.3	0.2	0.237	--	--	--
KRTAP13-2	68.4	17.1	53.7	0.089	--	--	--

Table S-2.2 (cont'd)

Gene	Sample Preparation Method Sequence Coverage Average (%)			One-way ANOVA p-Value	Tukey HSD Post-Hoc p-Values		
	Bulk-HVP	Single-LLE	Single-Acetone PPT		Single-Acetone PPT – Bulk-HVP	Single-LLE – Bulk-HVP	Single-LLE – Single-Acetone PPT
KRTAP3-1	59.4	17.6	52.8	0.158	--	--	--
KRTAP19-5	53.1	40.3	13.2	0.074	--	--	--
KRTAP9-3	46.9	12.6	42.5	0.125	--	--	--
KRTAP13-1	40.3	9.7	33.7	0.091	--	--	--
TXNRD1	35.0	4.2	19.9	0.094	--	--	--
LYG2	25.5	7.7	22.8	0.504	--	--	--
KRTAP9-8	53.8	15.7	42.8	0.146	--	--	--
RIDA	20.4	3.6	12.2	0.366	--	--	--
BLMH	15.6	4.9	13.5	0.539	--	--	--
KRTAP4-11	25.1	7.8	21.4	0.199	--	--	--
HIST1H1B	16.6	2.5	12.1	0.059	--	--	--
KRTAP4-8	20.9	8.8	21.1	0.389	--	--	--
GPRC5D	2.9	0.7	17.1	1.72×10^{-5}	7.63×10^{-5}	0.490	2.43×10^{-5}
YWHAB	21.8	7.2	18.6	0.286	--	--	--
EEF2	11.7	1.7	5.3	0.222	--	--	--
LDHA	21.4	2.9	14.0	0.009	0.285	0.007	0.085
PLD3	8.3	2.2	3.8	0.160	--	--	--
GPNMB	3.5	4.0	2.5	0.913	--	--	--
SFPQ	5.8	0.4	1.3	0.007	0.021	0.008	0.779
NPC1	0.9	0.5	2.5	0.011	0.038	0.745	0.012
RTN4	1.1	0.3	1.7	0.292	--	--	--
KRTAP4-4	24.7	5.1	20.3	0.243	--	--	--
GSTP1	15.0	8.1	16.5	0.726	--	--	--
KRTAP9-6	28.9	8.8	33.1	0.291	--	--	--
PLCD1	15.9	2.8	13.5	0.359	--	--	--
KRT2	7.2	27.2	5.9	0.164	--	--	--
MEMO1	11.8	2.3	14.6	0.165	--	--	--
KRTAP4-2	24.4	4.8	15.6	0.174	--	--	--
KRTAP10-12	11.2	6.3	25.0	0.173	--	--	--
KRTAP1-3	19.2	9.2	25.4	0.402	--	--	--
KRTAP9-7	25.9	8.7	24.4	0.395	--	--	--
RPSA	13.3	1.8	3.2	0.004	0.012	0.006	0.860
LMNB1	13.7	1.4	3.1	0.073	--	--	--
KRTAP10-11	5.7	1.7	10.1	0.110	--	--	--
HSPA8	23.1	4.3	8.2	0.100	--	--	--
PROCR	1.9	3.7	8.2	0.182	--	--	--
KRTAP9-1	13.9	4.4	13.2	0.305	--	--	--
HSPD1	9.6	1.2	2.5	0.140	--	--	--
PKM	8.2	2.0	1.0	0.065	--	--	--
KRT79	2.9	5.1	10.5	0.133	--	--	--
EIF4A1	5.8	4.4	7.7	0.799	--	--	--
KRT13	1.4	10.5	6.8	0.125	--	--	--
FAM83H	2.9	1.2	0.3	0.186	--	--	--
TPI1	26.0	11.8	20.5	0.731	--	--	--
KRTAP26-1	10.7	6.7	29.4	0.161	--	--	--
KRTAP9-4	32.8	14.4	43.5	0.485	--	--	--

Table S-2.2 (cont'd)

Gene	Sample Preparation Method Sequence Coverage Average (%)			One-way ANOVA p-Value	Tukey HSD Post-Hoc p-Values		
	Bulk-HVP	Single-LLE	Single-Acetone PPT		Single-Acetone PPT – Bulk-HVP	Single-LLE – Bulk-HVP	Single-LLE – Single-Acetone PPT
KRT9	4.9	33.4	1.5	0.051	--	--	--
HSPE1	6.9	14.2	15.2	0.751	--	--	--
ENO1	16.6	5.4	10.8	0.517	--	--	--
KRTAP9-2	27.2	12.9	38.6	0.488	--	--	--
GDPD3	16.3	0.0	10.1	0.194	--	--	--
KRTAP4-3	10.6	4.0	9.5	0.575	--	--	--
CKMT1A	11.5	4.7	4.7	0.531	--	--	--
APOD	12.7	0.0	7.3	0.162	--	--	--
EEF1G	12.1	2.2	8.7	0.311	--	--	--
KRTAP10-10	15.8	4.0	18.9	0.328	--	--	--
KRTAP4-1	6.7	0.0	8.2	0.059	--	--	--
PSAP	9.4	1.0	3.1	0.255	--	--	--
HADHB	8.8	0.0	3.5	0.049	0.245	0.042	0.496
KRTAP10-6	9.0	3.2	18.4	0.165	--	--	--
CTSD	2.8	3.9	2.7	0.887	--	--	--
PADI3	5.7	0.9	2.0	0.216	--	--	--
HSP90AA1	6.7	1.2	1.4	0.200	--	--	--
TGM1	3.5	0.4	2.4	0.286	--	--	--
KRT3	0.0	6.6	4.9	0.104	--	--	--
KRT76	3.6	5.2	1.5	0.488	--	--	--
FABP5	33.9	5.0	7.0	0.235	--	--	--
S100A14	38.9	2.6	0.0	0.008	0.012	0.017	0.965
CRIP2	19.2	3.8	3.8	0.202	--	--	--
KRTAP12-2	3.8	0.0	30.5	0.012	0.030	0.900	0.015
H2AFY	19.0	4.6	3.4	0.137	--	--	--
KRTAP10-9	4.7	5.0	28.5	0.058	--	--	--
HNRNPA2B1	15.6	0.0	3.8	0.026	0.089	0.026	0.719
H1F0	10.3	6.8	2.1	0.480	--	--	--
CFL1	14.6	11.4	5.1	0.653	--	--	--
RAN	7.2	1.2	9.7	0.447	--	--	--
LDHB	14.0	1.7	4.9	0.266	--	--	--
KRTAP2-2	11.4	20.9	48.6	0.314	--	--	--
CS	11.4	0.0	3.1	0.155	--	--	--
ATG9B	9.1	0.0	5.4	0.285	--	--	--
TAGLN2	15.8	0.0	8.5	0.091	--	--	--
KRTAP10-3	1.5	4.4	19.8	0.049	0.055	0.901	0.109
YWHAQ	12.7	0.0	13.4	0.173	--	--	--
MAP7	8.9	0.0	2.3	0.183	--	--	--
KRT7	2.6	7.5	5.5	0.617	--	--	--
RNH1	4.8	0.0	1.9	0.291	--	--	--
UQCRC1	4.8	0.0	1.1	0.022	0.075	0.022	0.715
HADHA	4.4	0.0	0.6	0.126	--	--	--
PABPC1	10.7	0.6	3.3	0.106	--	--	--
HK1	4.3	0.0	0.4	0.084	--	--	--
XPNPEP3	2.4	0.0	2.1	0.185	--	--	--
DSC3	3.9	0.0	0.7	0.197	--	--	--

Table S-2.2 (cont'd)

Gene	Sample Preparation Method Sequence Coverage Average (%)			One-way ANOVA p-Value	Tukey HSD Post-Hoc p-Values		
	Bulk-HVP	Single-LLE	Single-Acetone PPT		Single-Acetone PPT – Bulk-HVP	Single-LLE – Bulk-HVP	Single-LLE – Single-Acetone PPT
TUBB4B	43.5	0.0	10.6	0.005	0.026	0.005	0.571
PDIA3	2.3	0.0	1.6	0.297	--	--	--
KRTAP7-1	14.4	6.9	6.6	0.706	--	--	--
HIST1H3A	13.2	11.9	28.3	0.668	--	--	--
H3F3A	26.1	11.9	14.9	0.760	--	--	--
HIST2H3A	26.1	11.9	14.9	0.760	--	--	--
CRYAB	14.6	0.0	8.6	0.314	--	--	--
COMT	11.0	0.0	10.3	0.324	--	--	--
S100A6	13.1	0.0	4.4	0.190	--	--	--
GDI2	16.2	0.0	8.5	0.250	--	--	--
KRTAP6-1	10.9	0.0	3.5	0.070	--	--	--
ARL8B	12.1	1.2	8.2	0.495	--	--	--
RPS3	7.9	2.3	3.0	0.491	--	--	--
ANXA1	7.8	0.0	4.6	0.317	--	--	--
VDAC1	7.3	0.0	3.0	0.213	--	--	--
GSDMA	7.8	0.8	0.8	0.223	--	--	--
GGH	2.6	2.1	4.2	0.810	--	--	--
KRTAP4-5	21.1	0.0	6.8	0.082	--	--	--
GJA1	6.3	0.0	1.8	0.052	--	--	--
YWHAG	15.6	4.0	4.0	0.353	--	--	--
RBM14	5.4	0.0	1.3	0.087	--	--	--
MT-CO2	1.9	1.1	3.7	0.574	--	--	--
KRT14	4.0	11.8	0.0	0.168	--	--	--
SERPINB13	3.3	0.0	2.9	0.393	--	--	--
RPL13	3.1	0.0	3.1	0.329	--	--	--
SLC25A3	3.0	1.5	1.5	0.732	--	--	--
RACK1	1.0	2.4	2.5	0.810	--	--	--
RPL8	3.7	0.0	2.1	0.309	--	--	--
KRT72	0.0	2.6	5.5	0.237	--	--	--
KRT23	3.1	2.7	0.0	0.298	--	--	--
ASPRV1	2.8	0.0	2.3	0.338	--	--	--
KRT4	0.0	6.7	3.5	0.262	--	--	--
ALYREF	2.1	0.0	2.1	0.274	--	--	--
KRTAP29-1	1.8	0.0	1.8	0.274	--	--	--
RPS6	1.6	0.8	0.8	0.748	--	--	--
ACTBL2	3.2	0.0	8.6	0.096	--	--	--
KRT28	2.0	2.0	3.3	0.875	--	--	--
VIM	0.0	4.6	0.0	5.97×10^{-5}	1.000	1.34×10^{-4}	1.34×10^{-4}
KRT16	5.7	1.1	0.9	0.280	--	--	--
VCP	1.4	0.5	0.4	0.540	--	--	--
ENGASE	1.9	0.0	0.3	0.028	0.065	0.033	0.905
CHUK	0.0	0.0	2.0	1.47×10^{-143}	5.45×10^{-10}	0.858	5.45×10^{-10}
KRT71	2.6	1.4	0.0	0.225	--	--	--
HIST3H3	12.5	11.9	12.7	0.999	--	--	--
RPLP1	14.0	0.0	3.5	0.286	--	--	--
DCD	10.9	5.7	0.0	0.528	--	--	--

Table S-2.2 (cont'd)

Gene	Sample Preparation Method Sequence Coverage Average (%)			One-way ANOVA p-Value	Tukey HSD Post-Hoc p-Values		
	Bulk-HVP	Single-LLE	Single-Acetone PPT		Single-Acetone PPT – Bulk-HVP	Single-LLE – Bulk-HVP	Single-LLE – Single-Acetone PPT
PGK1	12.7	0.0	3.8	0.229	--	--	--
KRTAP19-7	8.7	0.0	4.4	0.323	--	--	--
PRSS3	0.0	7.6	6.6	0.518	--	--	--
CSTB	5.4	0.0	6.1	0.475	--	--	--
EIF6	9.7	0.0	1.0	0.193	--	--	--
CLIC3	10.7	0.0	0.0	0.067	--	--	--
CSNK1A1	10.9	0.0	3.0	0.229	--	--	--
KRTAP4-9	9.8	0.0	16.9	0.366	--	--	--
HNRNPA3	5.4	2.9	0.0	0.344	--	--	--
PHB2	5.4	0.0	1.7	0.243	--	--	--
SUN2	5.4	0.0	1.3	0.201	--	--	--
RTN3	0.0	0.8	5.7	0.153	--	--	--
KRTAP10-5	0.0	0.0	17.5	0.008	0.014	1.000	0.014
KRT27	2.5	8.5	0.0	0.207	--	--	--
NONO	5.6	0.0	1.3	0.199	--	--	--
PHGDH	4.3	0.0	0.9	0.175	--	--	--
ATP6V1A	4.6	0.0	0.5	0.154	--	--	--
RPS15A	3.1	0.0	1.5	0.323	--	--	--
CD9	2.2	0.0	2.4	0.513	--	--	--
PPME1	1.8	0.0	2.8	0.551	--	--	--
ENO3	3.6	0.9	1.9	0.725	--	--	--
TUFM	3.4	0.0	0.9	0.200	--	--	--
DNASE1L2	0.0	0.0	4.1	0.007	0.013	1.000	0.013
RPL12	2.7	0.0	1.4	0.323	--	--	--
RPS10	0.0	0.0	4.1	0.007	0.013	1.000	0.013
RPL6	3.0	0.0	0.8	0.304	--	--	--
LMNB2	3.7	0.0	1.0	0.199	--	--	--
DDX39B	3.3	0.0	5.6	0.367	--	--	--
CTSB	1.9	0.0	1.3	0.454	--	--	--
ASS1	2.2	0.0	0.8	0.259	--	--	--
TUBB3	0.0	0.0	16.3	0.016	0.028	1.000	0.028
KIF20A	0.8	0.0	1.7	0.323	--	--	--
KRT20	1.1	0.0	3.8	0.207	--	--	--
PCBP1	4.5	0.0	0.0	0.007	0.013	0.013	1.000
CPT1A	2.1	0.0	0.0	0.013	0.023	0.023	1.000
SDR16C5	0.0	0.0	1.9	0.007	0.013	1.000	0.013
KRT17	0.0	6.2	0.0	0.014	1.000	0.023	0.023
KRT77	1.8	3.5	0.0	0.337	--	--	--
DES	0.7	0.4	0.0	0.323	--	--	--
HSP90AB1	1.1	0.6	0.6	0.891	--	--	--
PGM2L1	0.0	1.0	0.0	0.007	1.000	0.013	0.013
TXN	23.3	0.0	0.0	0.100	--	--	--
KRTAP9-9	38.0	0.0	0.0	0.113	--	--	--
LGALS1	9.8	0.0	4.3	0.552	--	--	--
KRTAP1-1	0.0	10.9	10.6	0.622	--	--	--
RPLP2	8.0	0.0	3.5	0.552	--	--	--

Table S-2.2 (cont'd)

Gene	Sample Preparation Method Sequence Coverage Average (%)			One-way ANOVA p-Value	Tukey HSD Post-Hoc p-Values		
	Bulk-HVP	Single-LLE	Single-Acetone PPT		Single-Acetone PPT – Bulk-HVP	Single-LLE – Bulk-HVP	Single-LLE – Single-Acetone PPT
ECHDC1	10.3	0.0	0.0	0.100	--	--	--
G6PD	10.2	0.0	0.0	0.277	--	--	--
RPS11	10.0	0.0	0.0	0.141	--	--	--
ALDOA	9.1	0.0	0.0	0.106	--	--	--
HMGN2	8.3	0.0	0.0	0.100	--	--	--
RPS25	3.8	0.0	3.8	0.622	--	--	--
DYNLL1	13.5	0.0	0.0	0.100	--	--	--
BOLA2	7.3	0.0	0.0	0.129	--	--	--
PRDX2	9.7	0.0	0.0	0.118	--	--	--
RPL18	3.5	3.5	0.0	0.622	--	--	--
PGAM1	4.5	0.0	4.1	0.621	--	--	--
FKBP1A	6.0	0.0	0.0	0.100	--	--	--
S100A8	0.0	3.0	3.0	0.622	--	--	--
SH3BGRL3	5.4	0.0	0.0	0.100	--	--	--
RPS20	5.0	0.0	0.0	0.100	--	--	--
CAT	4.9	0.0	0.0	0.211	--	--	--
RPS18	4.8	0.0	0.0	0.153	--	--	--
RPS2	4.4	0.0	0.0	0.127	--	--	--
RPS14	4.3	0.0	0.0	0.100	--	--	--
KRTAP4-12	13.4	0.0	0.0	0.102	--	--	--
RPL15	4.2	0.0	0.0	0.181	--	--	--
SPINT1	4.1	0.0	0.0	0.105	--	--	--
METAP1	4.0	0.0	0.0	0.138	--	--	--
GPI	3.9	0.0	0.0	0.110	--	--	--
ACAA1	3.8	0.0	0.0	0.105	--	--	--
DBI	3.7	0.0	0.0	0.101	--	--	--
RPL31	1.8	0.0	1.8	0.622	--	--	--
GRN	3.5	0.0	0.0	0.118	--	--	--
CTNNA1	3.7	0.0	0.0	0.116	--	--	--
HSPA9	3.1	0.0	0.0	0.141	--	--	--
PPP1CB	2.9	0.0	0.0	0.182	--	--	--
FAM49A	3.8	0.0	0.0	0.126	--	--	--
MTCH2	2.7	0.0	0.0	0.117	--	--	--
ATP5O	1.4	0.0	1.3	0.621	--	--	--
RPL7A	2.6	0.0	0.0	0.100	--	--	--
P4HB	2.0	0.4	0.0	0.474	--	--	--
EFHD1	2.4	0.0	0.0	0.101	--	--	--
CDH1	2.4	0.0	0.0	0.100	--	--	--
FBP1	2.4	0.0	0.0	0.100	--	--	--
RTCB	2.3	0.0	0.0	0.160	--	--	--
GOT2	2.3	0.0	0.0	0.100	--	--	--
ILF2	2.2	0.0	0.0	0.100	--	--	--
RPS3A	2.0	0.0	0.0	0.108	--	--	--
NDUFV1	1.9	0.0	0.0	0.100	--	--	--
DNAJB6	3.8	0.0	0.0	0.100	--	--	--
SLC40A1	0.0	1.0	0.8	0.614	--	--	--

Table S-2.2 (cont'd)

Gene	Sample Preparation Method Sequence Coverage Average (%)			One-way ANOVA p-Value	Tukey HSD Post-Hoc p-Values		
	Bulk-HVP	Single-LLE	Single-Acetone PPT		Single-Acetone PPT – Bulk-HVP	Single-LLE – Bulk-HVP	Single-LLE – Single-Acetone PPT
SHMT2	1.8	0.0	0.0	0.191	--	--	--
TRIM25	1.8	0.0	0.0	0.153	--	--	--
KRTAP10-7	0.0	0.0	14.5	0.101	--	--	--
TALDO1	1.6	0.0	0.0	0.100	--	--	--
EEF1A2	0.0	3.3	4.8	0.605	--	--	--
KRT25	2.2	0.0	0.4	0.478	--	--	--
TUBB1	3.9	0.0	0.0	0.109	--	--	--
PCMT1	1.6	0.0	0.0	0.100	--	--	--
WNT3A	1.6	0.0	0.0	0.100	--	--	--
KRTAP10-4	0.0	0.0	8.7	0.103	--	--	--
KRT24	0.0	1.4	1.8	0.613	--	--	--
CXADR	1.4	0.0	0.0	0.104	--	--	--
KRT18	0.0	0.0	4.2	0.100	--	--	--
IGHA1	0.6	1.3	0.0	0.563	--	--	--
KRT15	1.8	0.0	2.2	0.616	--	--	--
FASN	1.0	0.0	0.0	0.104	--	--	--
NEFL	0.4	0.6	0.0	0.593	--	--	--
DHCR7	0.8	0.0	0.0	0.100	--	--	--
KRT8	4.0	0.0	0.0	0.101	--	--	--
GARS	0.8	0.0	0.0	0.141	--	--	--
SEC23B	0.8	0.0	0.0	0.103	--	--	--
CLTC	0.6	0.0	0.0	0.165	--	--	--
CAPN12	0.0	0.0	0.6	0.100	--	--	--
KIF20B	0.3	0.0	0.0	0.191	--	--	--
PKD2	0.2	0.0	0.2	0.622	--	--	--
DMXL1	0.3	0.0	0.0	0.224	--	--	--
S100A10	14.4	0.0	0.0	0.405	--	--	--
NUTF2	11.4	0.0	0.0	0.405	--	--	--
PRDX5	11.1	0.0	0.0	0.405	--	--	--
S100A9	10.3	0.0	0.0	0.405	--	--	--
PLP2	0.0	0.0	9.9	0.405	--	--	--
S100A7	9.4	0.0	0.0	0.405	--	--	--
KRTAP12-1	0.0	0.0	9.1	0.405	--	--	--
ACOT7	8.9	0.0	0.0	0.405	--	--	--
LYPLA1	7.0	0.0	0.0	0.405	--	--	--
ATOX1	6.6	0.0	0.0	0.405	--	--	--
ATP6V0C	0.0	0.0	5.8	0.405	--	--	--
PDIA6	5.6	0.0	0.0	0.405	--	--	--
NDUFA13	4.9	0.0	0.0	0.405	--	--	--
HINT1	4.8	0.0	0.0	0.405	--	--	--
HAGH	4.4	0.0	0.0	0.405	--	--	--
RPS28	0.0	4.3	0.0	0.405	--	--	--
PSMA5	3.9	0.0	0.0	0.405	--	--	--
HMGN3	3.8	0.0	0.0	0.405	--	--	--
COX5A	3.5	0.0	0.0	0.405	--	--	--
SRI	3.0	0.0	0.0	0.405	--	--	--

Table S-2.2 (cont'd)

Gene	Sample Preparation Method Sequence Coverage Average (%)			One-way ANOVA p-Value	Tukey HSD Post-Hoc p-Values		
	Bulk-HVP	Single-LLE	Single-Acetone PPT		Single-Acetone PPT – Bulk-HVP	Single-LLE – Bulk-HVP	Single-LLE – Single-Acetone PPT
RPS21	3.0	0.0	0.0	0.405	--	--	--
RAB1A	3.7	0.0	0.0	0.405	--	--	--
RPL10A	2.9	0.0	0.0	0.405	--	--	--
H2BFS	0.0	8.9	0.0	0.405	--	--	--
HIST1H2BA	7.3	0.0	0.0	0.405	--	--	--
PFN1	2.7	0.0	0.0	0.405	--	--	--
HIST2H2AB	15.2	0.0	0.0	0.405	--	--	--
ACADVL	2.7	0.0	0.0	0.405	--	--	--
S100A16	0.0	0.0	2.7	0.405	--	--	--
GGCT	2.7	0.0	0.0	0.405	--	--	--
KRTAP4-6	9.8	0.0	0.0	0.405	--	--	--
RPL22	0.0	2.5	0.0	0.405	--	--	--
YPEL5	0.0	0.0	2.5	0.405	--	--	--
UQCRH	2.5	0.0	0.0	0.405	--	--	--
GIPC1	2.3	0.0	0.0	0.405	--	--	--
C1QBP	2.3	0.0	0.0	0.405	--	--	--
EDF1	2.2	0.0	0.0	0.405	--	--	--
CTNND1	2.2	0.0	0.0	0.405	--	--	--
CHAC1	2.1	0.0	0.0	0.405	--	--	--
TUBA4B	3.5	0.0	0.0	0.405	--	--	--
HNRNPH1	5.1	0.0	0.0	0.405	--	--	--
PARK7	2.0	0.0	0.0	0.405	--	--	--
SNRPD3	2.0	0.0	0.0	0.405	--	--	--
HNRNPD	2.0	0.0	0.0	0.405	--	--	--
RPL27A	1.9	0.0	0.0	0.405	--	--	--
KRTAP4-7	6.2	0.0	0.0	0.405	--	--	--
DNAJC7	1.8	0.0	0.0	0.405	--	--	--
TUBB	13.4	0.0	0.0	0.405	--	--	--
MYL4	1.6	0.0	0.0	0.405	--	--	--
IL1F10	1.6	0.0	0.0	0.405	--	--	--
COX4I1	1.6	0.0	0.0	0.405	--	--	--
RPS19	1.6	0.0	0.0	0.405	--	--	--
TSPAN7	0.0	1.5	0.0	0.405	--	--	--
ATP5F1	1.5	0.0	0.0	0.405	--	--	--
TARS	1.5	0.0	0.0	0.405	--	--	--
IDH2	0.0	1.4	0.0	0.405	--	--	--
PLA2G2E	1.4	0.0	0.0	0.405	--	--	--
HIST3H2BB	0.0	10.3	0.0	0.405	--	--	--
SSBP1	1.4	0.0	0.0	0.405	--	--	--
CCL21	0.0	0.0	1.3	0.405	--	--	--
RAB15	0.0	0.0	1.3	0.405	--	--	--
ECHS1	1.3	0.0	0.0	0.405	--	--	--
RPL23A	1.3	0.0	0.0	0.405	--	--	--
LAMP2	0.0	1.3	0.0	0.405	--	--	--
TPP1	1.2	0.0	0.0	0.405	--	--	--
KHDRBS1	1.2	0.0	0.0	0.405	--	--	--

Table S-2.2 (cont'd)

Gene	Sample Preparation Method Sequence Coverage Average (%)			One-way ANOVA p-Value	Tukey HSD Post-Hoc p-Values		
	Bulk-HVP	Single-LLE	Single-Acetone PPT		Single-Acetone PPT – Bulk-HVP	Single-LLE – Bulk-HVP	Single-LLE – Single-Acetone PPT
PSMD2	1.2	0.0	0.0	0.405	--	--	--
TMED10	1.1	0.0	0.0	0.405	--	--	--
TUBA1C	0.0	0.0	3.9	0.405	--	--	--
APRT	1.1	0.0	0.0	0.405	--	--	--
PHB	1.1	0.0	0.0	0.405	--	--	--
CAPZB	1.1	0.0	0.0	0.405	--	--	--
EEF1D	0.0	1.1	0.0	0.405	--	--	--
PGLS	1.1	0.0	0.0	0.405	--	--	--
HNRNPH2	4.1	0.0	0.0	0.405	--	--	--
RPS7	1.0	0.0	0.0	0.405	--	--	--
EIF3I	1.0	0.0	0.0	0.405	--	--	--
CBR1	1.0	0.0	0.0	0.405	--	--	--
HNRNPK	0.0	0.0	0.9	0.405	--	--	--
RNF39	0.9	0.0	0.0	0.405	--	--	--
DNAJA1	0.8	0.0	0.0	0.405	--	--	--
FSCN1	0.8	0.0	0.0	0.405	--	--	--
WNT3	0.8	0.0	0.0	0.405	--	--	--
TACSTD2	0.8	0.0	0.0	0.405	--	--	--
IDH3A	0.8	0.0	0.0	0.405	--	--	--
PPT2	0.7	0.0	0.0	0.405	--	--	--
USP6NL	0.7	0.0	0.0	0.405	--	--	--
CSTF1	0.7	0.0	0.0	0.405	--	--	--
IMMT	0.7	0.0	0.0	0.405	--	--	--
AP1B1	0.7	0.0	0.0	0.405	--	--	--
RNASET2	0.0	0.0	0.7	0.405	--	--	--
RPL3	0.7	0.0	0.0	0.405	--	--	--
SPECC1	0.7	0.0	0.0	0.405	--	--	--
RPS4X	1.6	0.0	0.0	0.405	--	--	--
GNA13	0.7	0.0	0.0	0.405	--	--	--
RPL4	0.0	0.0	0.6	0.405	--	--	--
PTBP1	0.6	0.0	0.0	0.405	--	--	--
SLC1A5	0.0	0.5	0.0	0.405	--	--	--
CAPG	0.5	0.0	0.0	0.405	--	--	--
EIF4A2	2.1	0.0	0.0	0.405	--	--	--
FLG2	0.4	0.0	0.0	0.405	--	--	--
PABPC3	1.0	0.0	0.0	0.405	--	--	--
SLC27A6	0.4	0.0	0.0	0.405	--	--	--
ATP2B4	0.4	0.0	0.0	0.405	--	--	--
PMEL	0.0	0.3	0.0	0.405	--	--	--
DSG1	0.6	0.0	0.0	0.405	--	--	--
CRAT	0.0	0.0	0.3	0.405	--	--	--
ILF3	0.3	0.0	0.0	0.405	--	--	--
AHCTF1	0.3	0.0	0.0	0.405	--	--	--
PREP	0.3	0.0	0.0	0.405	--	--	--
ANK1	0.0	0.0	0.3	0.405	--	--	--
POTEE	1.1	0.0	0.0	0.405	--	--	--

Table S-2.2 (cont'd)

Gene	Sample Preparation Method Sequence Coverage Average (%)			One-way ANOVA p-Value	Tukey HSD Post-Hoc p-Values		
	Bulk-HVP	Single-LLE	Single-Acetone PPT		Single-Acetone PPT – Bulk-HVP	Single-LLE – Bulk-HVP	Single-LLE – Single-Acetone PPT
CDC42BPA	0.0	0.0	0.2	0.405	--	--	--
SEC24C	0.2	0.0	0.0	0.405	--	--	--
TNIK	0.0	0.0	0.2	0.405	--	--	--
HERC4	0.2	0.0	0.0	0.405	--	--	--
ITSN2	0.2	0.0	0.0	0.405	--	--	--
ATP7A	0.2	0.0	0.0	0.405	--	--	--
USP35	0.2	0.0	0.0	0.405	--	--	--
CEP250	0.0	0.0	0.3	0.405	--	--	--
NPC1L1	0.1	0.0	0.0	0.405	--	--	--
RB1CC1	0.1	0.0	0.0	0.405	--	--	--
TPR	0.1	0.0	0.0	0.405	--	--	--
DYNC1H1	0.1	0.0	0.0	0.405	--	--	--
NEB	0.0	0.0	0.0	0.405	--	--	--
KIAA1109	0.0	0.0	0.0	0.405	--	--	--

Table S-2.3. Comprehensive GVP profile, based on the detection of major and minor GVPs. 0 and 1 represent detection of the major and minor GVP, respectively. '--' indicates non-detection of GVPs. Samples are denoted x.y, where x is the individual code (of 4 individuals) and y is the sample preparation method.

Gene SNP	Individual 1			Individual 2			Individual 3			Individual 4		
	1.Bulk-HVP	1.Single-LLE	1.Single-Acetone PPT	2.Bulk-HVP	2.Single-LLE	2.Single-Acetone PPT	3.Bulk-HVP	3.Single-LLE	3.Single-Acetone PPT	4.Bulk-HVP	4.Single-LLE	4.Single-Acetone PPT
KRTAP10-6 rs62220887	--	--	1	--	--	1	--	--	1	--	--	--
KRT83 rs2857663	--	0	0	--	0	0	1	--	0,1	--	--	0
KRT75 rs2232387	--	--	--	--	--	--	--	--	--	--	--	1
KRTAP3-2 rs9897046	0	--	0	0	0	0	0	--	0	0	--	0
KRTAP11-1 rs71321355	0,1	--	0,1	0	0	0	0	--	0	--	--	--
KRT39 rs7213256	--	--	--	1	1	1	--	--	0	--	--	--
KRTAP10-12 rs34302939	0	--	0	0	0	0	0	--	0	--	--	--
PABPC1 rs62513922	--	--	--	0	--	--	0	--	--	0	--	--
KRT40 rs2010027	--	--	0,1	1	--	1	0	0	0	--	--	--
GSDMA rs7212944	--	--	--	--	0	0	--	--	--	--	--	--
KRT77 rs3782489	--	--	--	--	0	--	--	0	--	--	--	--
KRT32 rs2071563	0,1	0,1	0,1	0	0	0	0,1	0,1	0,1	0	0	0
KRT82 rs2658658	0,1	--	1	--	0	0	0,1	--	0,1	--	--	--
KRT37 rs9916724	--	--	--	--	--	0	--	--	--	--	--	--
KRTAP4-11 rs349771	1	--	1	--	--	--	1	--	1	--	--	--
KRTAP4-5 rs1497383	--	--	--	--	--	--	0	--	--	--	--	--

Table S-2.3 (cont'd)

Gene SNP	Individual 1			Individual 2			Individual 3			Individual 4		
	1.Bulk-HVP	1.Single-LLE	1.Single-Acetone PPT	2.Bulk-HVP	2.Single-LLE	2.Single-Acetone PPT	3.Bulk-HVP	3.Single-LLE	3.Single-Acetone PPT	4.Bulk-HVP	4.Single-LLE	4.Single-Acetone PPT
PPL rs2037912	--	--	1	--	--	--	--	--	--	--	--	--
KRT3 rs3887954	--	--	--	--	--	--	--	--	--	--	1	--
KRT79 rs2638497	--	--	--	--	--	--	--	--	--	0	--	--
KRTAP10-9 rs9980129	--	--	1	--	--	1	--	--	1	--	--	--
TGM3 rs214803	1	1	1	0,1	0,1	0,1	1	1	1	0	--	--
KRT13 rs9891361	--	--	1	--	--	--	--	--	1	--	--	1
KRTAP10-3 rs233252	--	--	1	--	--	--	--	--	--	--	--	--
KRTAP4-2 rs389784	--	--	--	--	--	--	1	--	--	--	--	--
KRTAP4-1 rs398825	1	--	--	--	--	--	--	--	--	--	--	--
HEXB rs820878	--	--	--	1	--	--	--	--	--	--	--	--
FAM83H rs9969600	--	--	--	1	--	--	1	--	--	--	--	--

REFERENCES

REFERENCES

1. Chu, F.; Mason, K. E.; Anex, D. S.; Jones, A. D.; Hart, B. R., Hair Proteome Variation at Different Body Locations on Genetically Variant Peptide Detection for Protein-Based Human Identification. *Scientific Reports* **2019,** *9* (1), 7641.

2. Parker, G. J.; Leppert, T.; Anex, D. S.; Hilmer, J. K.; Matsunami, N.; Baird, L.; Stevens, J.; Parsawar, K.; Durbin-Johnson, B. P.; Rocke, D. M.; Nelson, C.; Fairbanks, D. J.; Wilson, A. S.; Rice, R. H.; Woodward, S. R.; Bothner, B.; Hart, B. R.; Leppert, M., Demonstration of Protein-Based Human Identification Using the Hair Shaft Proteome. *PLoS ONE* **2016,** *11* (9), e0160653.

3. Waridel, P.; Frank, A.; Thomas, H.; Surendranath, V.; Sunyaev, S.; Pevzner, P.; Shevchenko, A., Sequence Similarity-Driven Proteomics in Organisms with Unknown Genomes by LC-MS/MS and Automated De Novo Sequencing. *Proteomics* **2007,** *7* (14), 2318-2329.

4. Ma, B.; Zhang, K.; Hendrie, C.; Liang, C.; Li, M.; Doherty-Kirby, A.; Lajoie, G., PEAKS: Powerful Software for Peptide De Novo Sequencing by Tandem Mass Spectrometry. *Rapid Communications in Mass Spectrometry* **2003,** *17* (20), 2337-2342.

5. Eng, J. K.; McCormack, A. L.; Yates, J. R., An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *Journal of the American Society for Mass Spectrometry* **1994,** *5* (11), 976-989.

6. Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S., Probability-Based Protein Identification by Searching Sequence Databases Using Mass Spectrometry Data. *Electrophoresis* **1999,** *20* (18), 3551-3567.

7. Craig, R.; Beavis, R. C., TANDEM: Matching Proteins with Tandem Mass Spectra. *Bioinformatics* **2004,** *20* (9), 1466-1467.

8. Craig, R.; Cortens, J. C.; Fenyo, D.; Beavis, R. C., Using Annotated Peptide Mass Spectrum Libraries for Protein Identification. *Journal of Proteome Research* **2006,** *5* (8), 1843-1849.

9. Yen, C.-Y.; Houel, S.; Ahn, N. G.; Old, W. M., Spectrum-to-Spectrum Searching Using a Proteome-Wide Spectral Library. *Molecular & Cellular Proteomics* **2011,** *10* (7), M111.007666.

10. Fenyö, D.; Eriksson, J.; Beavis, R., Mass Spectrometric Protein Identification Using the Global Proteome Machine. *Methods in Molecular Biology* **2010,** *673*, 189-202.

11. Zhang, J.; Xin, L.; Shan, B.; Chen, W.; Xie, M.; Yuen, D.; Zhang, W.; Zhang, Z.; Lajoie, G. A.; Ma, B., PEAKS DB: De Novo Sequencing Assisted Database Search for Sensitive

and Accurate Peptide Identification. *Molecular & Cellular Proteomics* **2012,** *11* (4), M111.010587.

12. Cruz, C. F.; Azoia, N. G.; Matamá, T.; Cavaco-Paulo, A., Peptide—Protein Interactions Within Human Hair Keratins. *International Journal of Biological Macromolecules* **2017,** *101*, 805-814.

13. Rice, R. H., Proteomic Analysis of Hair Shaft and Nail Plate. *Journal of Cosmetic Science* **2011,** *62* (2), 229-236.

14. Feist, P.; Hummon, B. A., Proteomic Challenges: Sample Preparation Techniques for Microgram-Quantity Protein Analysis from Biological Samples. *International Journal of Molecular Sciences* **2015,** *16* (2).

15. Mason, K. E.; Paul, P. H.; Chu, F.; Anex, D. S.; Hart, B. R., Development of a Protein-Based Human Identification Capability from a Single Hair. *Journal of Forensic Sciences* **2019,** *0* (0).

16. Crowell, A. M. J.; Wall, M. J.; Doucette, A. A., Maximizing Recovery of Water-Soluble Proteins Through Acetone Precipitation. *Analytica Chimica Acta* **2013,** *796*, 48-54.

17. Keller, A.; Eng, J.; Zhang, N.; Li, X.-j.; Aebersold, R., A Uniform Proteomics MS/MS Analysis Platform Utilizing Open XML File Formats. *Molecular Systems Biology* **2005,** *1* (1), 2005.0017.

18. Pedrioli, P. G. A.; Eng, J. K.; Hubley, R.; Vogelzang, M.; Deutsch, E. W.; Raught, B.; Pratt, B.; Nilsson, E.; Angeletti, R. H.; Apweiler, R.; Cheung, K.; Costello, C. E.; Hermjakob, H.; Huang, S.; Julian, R. K.; Kapp, E.; McComb, M. E.; Oliver, S. G.; Omenn, G.; Paton, N. W.; Simpson, R.; Smith, R.; Taylor, C. F.; Zhu, W.; Aebersold, R., A Common Open Representation of Mass Spectrometry Data and Its Application to Proteomics Research. *Nature Biotechnology* **2004,** *22* (11), 1459-1466.

19. Hubbard, T.; Barker, D.; Birney, E.; Cameron, G.; Chen, Y.; Clark, L.; Cox, T.; Cuff, J.; Curwen, V.; Down, T.; Durbin, R.; Eyras, E.; Gilbert, J.; Hammond, M.; Huminiecki, L.; Kasprzyk, A.; Lehvaslaiho, H.; Lijnzaad, P.; Melsopp, C.; Mongin, E.; Pettett, R.; Pocock, M.; Potter, S.; Rust, A.; Schmidt, E.; Searle, S.; Slater, G.; Smith, J.; Spooner, W.; Stabenau, A.; Stalker, J.; Stupka, E.; Ureta-Vidal, A.; Vastrik, I.; Clamp, M., The Ensembl Genome Database Project. *Nucleic Acids Research* **2002,** *30* (1), 38-41.

20. Boutet, E.; Liberherr, D.; Tognolli, M.; Schneider, M.; Bairoch, A., UniProtKB/Swiss-Prot. *Methods in Molecular Biology* **2007,** *406*, 89-112.

21. Fenyö, D.; Beavis, R. C., A Method for Assessing the Statistical Significance of Mass Spectrometry-Based Protein Identifications Using General Scoring Schemes. *Analytical Chemistry* **2003,** *75* (4), 768-774.

22. Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P., ProteoWizard: Open Source Software for Rapid Proteomics Tools Development. *Bioinformatics* **2008,** *24* (21), 2534-2536.

23. Kempkes, L. J. M.; Martens, J.; Grzetic, J.; Berden, G.; Oomens, J., Deamidation Reactions of Asparagine- and Glutamine-Containing Dipeptides Investigated by Ion Spectroscopy. *Journal of the American Society for Mass Spectrometry* **2016,** *27* (11), 1855-1869.

24. Brosch, M.; Yu, L.; Hubbard, T.; Choudhary, J., Accurate and Sensitive Peptide Identification with Mascot Percolator. *Journal of Proteome Research* **2009,** *8* (6), 3176-3181.

25. Nesvizhskii, A. I.; Aebersold, R., Interpretation of Shotgun Proteomic Data. *Molecular & Cellular Proteomics* **2005,** *4* (10), 1419-1440.

26. Sun, S.; Zhou, J.-Y.; Yang, W.; Zhang, H., Inhibition of Protein Carbamylation in Urea Solution Using Ammonium-Containing Buffers. *Analytical Biochemistry* **2014,** *446*, 76-81.

27. Kitts, A.; Sherry, S., The Single Nucleotide Polymorphism Database (dbSNP) of Nucleotide Sequence Variation. In *The NCBI Handbook*, McEntyre, J.; Ostell, J., Eds. National Center for Biotechnology Information (US): Bethesda, MD, 2002 [Updated 2011].

28. Karczewski, K. J.; Francioli, L. C.; Tiao, G.; Cummings, B. B.; Alföldi, J.; Wang, Q.; Collins, R. L.; Laricchia, K. M.; Ganna, A.; Birnbaum, D. P.; Gauthier, L. D.; Brand, H.; Solomonson, M.; Watts, N. A.; Rhodes, D.; Singer-Berk, M.; England, E. M.; Seaby, E. G.; Kosmicki, J. A.; Walters, R. K.; Tashman, K.; Farjoun, Y.; Banks, E.; Poterba, T.; Wang, A.; Seed, C.; Whiffin, N.; Chong, J. X.; Samocha, K. E.; Pierce-Hoffman, E.; Zappala, Z.; O'Donnell-Luria, A. H.; Minikel, E. V.; Weisburd, B.; Lek, M.; Ware, J. S.; Vittal, C.; Armean, I. M.; Bergelson, L.; Cibulskis, K.; Connolly, K. M.; Covarrubias, M.; Donnelly, S.; Ferriera, S.; Gabriel, S.; Gentry, J.; Gupta, N.; Jeandet, T.; Kaplan, D.; Llanwarne, C.; Munshi, R.; Novod, S.; Petrillo, N.; Roazen, D.; Ruano-Rubio, V.; Saltzman, A.; Schleicher, M.; Soto, J.; Tibbetts, K.; Tolonen, C.; Wade, G.; Talkowski, M. E.; Neale, B. M.; Daly, M. J.; MacArthur, D. G., Variation Across 141,456 Human Exomes and Genomes Reveals the Spectrum of Loss-of-Function Intolerance Across Human Protein-Coding Genes. *bioRxiv* **2019**, 531210.

29. Lek, M.; Karczewski, K. J.; Minikel, E. V.; Samocha, K. E.; Banks, E.; Fennell, T.; O'Donnell-Luria, A. H.; Ware, J. S.; Hill, A. J.; Cummings, B. B.; Tukiainen, T.; Birnbaum, D. P.; Kosmicki, J. A.; Duncan, L. E.; Estrada, K.; Zhao, F.; Zou, J.; Pierce-Hoffman, E.; Berghout, J.; Cooper, D. N.; Deflaux, N.; DePristo, M.; Do, R.; Flannick, J.; Fromer, M.; Gauthier, L.; Goldstein, J.; Gupta, N.; Howrigan, D.; Kiezun, A.; Kurki, M. I.; Moonshine, A. L.; Natarajan, P.; Orozco, L.; Peloso, G. M.; Poplin, R.; Rivas, M. A.; Ruano-Rubio, V.; Rose, S. A.; Ruderfer, D. M.; Shakir, K.; Stenson, P. D.; Stevens, C.; Thomas, B. P.; Tiao, G.; Tusie-Luna, M. T.; Weisburd, B.; Won, H.-H.; Yu, D.; Altshuler, D. M.; Ardissino, D.; Boehnke, M.; Danesh, J.; Donnelly, S.; Elosua, R.; Florez, J. C.; Gabriel, S. B.; Getz, G.; Glatt, S. J.; Hultman, C. M.; Kathiresan, S.; Laakso, M.; McCarroll, S.; McCarthy, M. I.; McGovern, D.; McPherson, R.; Neale, B. M.; Palotie, A.; Purcell, S. M.; Saleheen, D.; Scharf, J. M.; Sklar, P.; Sullivan, P. F.; Tuomilehto, J.; Tsuang, M. T.; Watkins, H. C.;

Wilson, J. G.; Daly, M. J.; MacArthur, D. G.; Exome Aggregation Consortium, Analysis of Protein-Coding Genetic Variation in 60,706 Humans. *Nature* **2016,** *536*, 285.

30. Wigginton, J. E.; Cutler, D. J.; Abecasis, G. R., A Note on Exact Tests of Hardy-Weinberg Equilibrium. *The American Journal of Human Genetics* **2005,** *76* (5), 887-893.

31. Graffelman, J., Exploring Diallelic Genetic Markers: The HardyWeinberg Package. *Universitat Politecnica de Catalunya* **2019**.

32. Pemberton, T. J.; Rosenberg, N. A., Population-Genetic Influences on Genomic Estimates of the Inbreeding Coefficient: A Global Perspective. *Human Heredity* **2014,** *77* (1-4), 37-48.

33. Reich, D. E.; Cargill, M.; Bolk, S.; Ireland, J.; Sabeti, P. C.; Richter, D. J.; Lavery, T.; Kouyoumjian, R.; Farhadian, S. F.; Ward, R.; Lander, E. S., Linkage Disequilibrium in the Human Genome. *Nature* **2001,** *411*, 199.

34. Ardlie, K. G.; Kruglyak, L.; Seielstad, M., Patterns of Linkage Disequilibrium in the Human Genome. *Nature Reviews Genetics* **2002,** *3* (4), 299-309.

35. Abecasis, G. R.; Noguchi, E.; Heinzmann, A.; Traherne, J. A.; Bhattacharyya, S.; Leaves, N. I.; Anderson, G. G.; Zhang, Y.; Lench, N. J.; Carey, A.; Cardon, L. R.; Moffatt, M. F.; Cookson, W. O. C., Extent and Distribution of Linkage Disequilibrium in Three Genomic Regions. *The American Journal of Human Genetics* **2001,** *68* (1), 191-197.

CHAPTER 3: Effects of Hair Proteome Variation at Different Body Locations on Identification of Genetically Variant Peptides

Foreword

This chapter describes work that has been adapted from a published paper[1]. Contributions from others to the conduct of the experiments described in this chapter are as follows, in no particular order: P. H. Paul provided single nucleotide variant lists and individualized mutated protein FASTA files, and K. E. Mason and D. S. Anex acquired the mass spectrometry data.

3.1 Introduction

Chapter 2 established an optimized workflow for single hair analysis, which has demonstrated relevance for protein-based human identification using genetically variant peptides (GVPs) from mass-limited hair evidence. To further develop GVP analysis for forensic applications, there is a need to determine the forensic contexts in which GVP analysis of single hairs can be used, such as effects of intrinsic variation in hair protein chemistry on GVP detection. One key knowledge gap lies in whether the same GVP markers can be identified in hair from different body locations, which may exhibit different chemistries arising from different environments or location-dependent gene expression; hair origin is often not known from hair specimens recovered for forensic investigations.

Application of GVP analysis for forensic investigations demands that detected GVP markers reflect an individual's genetics, i.e., an individual's SNP genotypes. Variability in their detection that results from matrix differences, such from different body locations, might exert potential to compromise identifications. Transcription and translation of SNPs from nuclear DNA into proteins are not known to vary with different body locations, and this study is based on the premise that common hair proteins are expressed in hairs from all body locations.

However, GVP detection may be compromised if associated proteins carrying the amino acid polymorphisms are not expressed in the tissue of interest or expressed at levels too low to be detected. It is well known that proteins are differentially expressed among tissues and cell types; a large-scale survey of data from proteomics experiments of tissues and body fluids found that protein abundances can vary drastically among biological specimens[2]. For example, the ubiquitous housekeeping protein glyceraldehyde-3-phosphate dehydrogenase (GAPDH) is highly expressed in skin and brain tissue, but exhibits low abundance in stem cells[2]. Wilhelm and co-workers demonstrated that approximately 70 highly expressed proteins are found in all tissues and bodily fluids, including hair follicle, but can span up to 5 orders of magnitude difference in protein expression[2]. If proteins carrying amino acid polymorphisms are expressed in lower abundance in hair from different body locations, GVPs from those mutated proteins may not be consistently detected in a particular hair type, and the power to distinguish an individual's genetics would be diminished.

However, examination of effects of differential protein expression in hair with body location on GVP detection requires knowledge of body location-specific hair protein expression, which has not been well-characterized. Reports of differences in hair protein chemistry among the various body locations have been scarce[3, 4]; however, morphological differences among these hair types may be linked to intrinsic variation in hair protein chemistry, as these differences may influence variation in hair physicochemical properties. Clear morphological differences exist among hair from different body locations, such as thickness and length between pubic and head hair, which are dictated by hair follicle differentiation and growth[5]. Hair follicle growth is modulated by androgens and other hormones; androgen signaling in the hair follicle transforms fine vellus hairs into thicker and longer pigmented hairs[5]. Conversely, the regulation of large hair

follicles to those producing vellus hairs, which result in balding, also rely on androgen signaling[6]. Through binding to receptors in the hair bulb, androgens are known to regulate growth of and gene expression in the hair follicle[7], which exhibit different responses among body locations[5], and variation of androgen levels with hair follicle type have the potential to change hair protein composition. As hair shaft grows out of the follicle and comprises mostly protein, it is likely that differences in protein composition and abundance manifested from differential gene expression contribute to the morphological differences. Further, from work performed by Laatsch et al., differential protein expression in hair from a few body locations has been characterized by measuring protein abundances via a bottom-up, or shotgun, proteomics approach[3], demonstrating detectable intrinsic variation in hair protein chemistry at different body sites, which ranges between 2- and 15-fold differences. Therefore, detection of some GVPs associated with body location-specific differential protein expression may be compromised. With the exception of a similar study that was performed concurrent with the present work, though analyzed with a slightly different approach with regards to variant peptide marker identification[4], effects of body location-specific protein abundance variation on robust identification of GVP markers have received limited attention.

In this investigation, proteomics technologies and methodologies developed for single hair analysis were employed to examine GVP markers identified from head, arm, and pubic hair for any differences in the potential for differentiation of individuals. Aims of this research include: 1. determination of body location-specific variation in hair proteome composition, 2. evaluation of the effects of differential protein expression on GVP identification and subsequent SNP inference, and 3. quantification of the extent to which individuals are differentiated with robust, i.e., consistently identified and body location-invariant, hair GVP markers. This work

aims to establish the independence of forensic SNP identification from body location-specific hair proteomic variation and further identify viable GVP markers that yield similar distinction of individuals irrespective of body location origin in single one-inch (25 mm) hairs.

3.2 Experimental

3.2.1 Hair Sample Preparation for Mass Spectrometry

Head, arm, and pubic hair specimens from three subjects (ages 25, 31, and 35) were collected to profile the protein variation in non-chemically treated hair from different body locations, under approval by the Institutional Review Board at Lawrence Livermore National Laboratory (Protocol ID# 15-008) and in accordance with the Declaration of Helsinki. Written informed consent for specimen collection and analysis was obtained prior to collection. Samples were stored in the dark at room temperature (RT). A one-inch (25 mm) single hair was segmented from a hair sample, and each was further segmented into four pieces of equal length (~6 mm) for full immersion into the denaturation solution for protein extraction. To account for biological variation within individuals, different 25-mm single hair specimens from each body location were prepared as n = 4 for each individual. The same protocol was followed to prepare the first three sets of replicates for proteomics-only analysis; a fourth set of replicates was prepared with a slightly modified protocol for protein/DNA co-extraction.

Aliquots of 100 µL of an aqueous denaturation solution (50 mM dithiothreitol (DTT), 50 mM ammonium bicarbonate (ABC), and 20 mg/mL sodium dodecanoate (SDD)) were added to each single-inch hair specimen contained in a microcentrifuge tube. The tubes were sealed, placed in a 70 °C water bath, and ultrasonicated at a frequency of 37 kHz and 100% power (Elma, Singen, Germany) until each hair sample was entirely solubilized (on average, 2 h). 10

µL of 1 M aqueous iodoacetamide solution was added to each sample and the extracts were incubated in the dark at RT for 45 min.

To remove SDD (an ionization-suppressing compound in LC-MS/MS analysis) from the extracts, liquid-liquid extraction was performed by addition of 100 µL of 0.75% (v/v) trifluoroacetic acid in ethyl acetate. The upper organic layer was removed after phase separation, taking care to not remove the precipitated protein layer at the organic-aqueous interface, and each extract was re-adjusted to pH = 8 by addition of 10 µL of 1 M ABC. Protein concentration was performed using spin filter concentrators with a lock volume of 20 µL (PES, 10 kDa MWCO; Thermo Fisher Scientific Inc., Waltham, MA). Samples were centrifuged at 3,000 × *g* for 15 min at RT, and 60 µL of buffer solution (50 mM DTT and 50 mM ABC) were added to wash the retentates, followed by a wash with 30 µL of buffer after centrifugation for 15 min. Finally, spin filter retentates were centrifuged for 30 min and reconstituted to 50 µL with buffer solution (50 mM DTT, 50 mM ABC, and 0.1 mg/mL ProteaseMAX™ Surfactant (Promega, Madison, WI)) prior to overnight trypsin digestion (TPCK-treated, sequencing grade) for at least 18 h accompanied by magnetic stirring with micro stir bars at RT.

Protein digests were filtered for particulates using centrifugal filter tubes (PVDF, 0.1 µm; MilliporeSigma, Burlington, MA) with centrifugation at 9,000 × *g* for 10 min at RT. Filtered digests for the first three sets of replicates were then analyzed by LC-MS/MS. For the fourth set of replicates, a protein/DNA co-extraction procedure was performed by adding 200 µL of ethanol to each filtrate, and the mixture was transferred to a DNA-binding column from the QIAamp DNA Micro Kit (Qiagen, Germantown, MD) and fractionated via centrifugation into a protein fraction (flow-through) and a DNA fraction (retentate). Protein digest eluate collected after centrifugation at 6,000 × *g* for 1 min at RT was evaporated to dryness under vacuum and

reconstituted in 50 µL of 50 mM ABC, 50 mM DTT, and 0.1 mg/mL ProteaseMAX™ Surfactant. The reconstituted protein digest was filtered as above and analyzed via LC-MS/MS.

3.2.2 Liquid Chromatography-Tandem Mass Spectrometry Analysis

Protein digests were analyzed on an EASY-nLC 1200 system coupled to a Q Exactive Plus Orbitrap mass spectrometer (Thermo Fisher Scientific Inc., Waltham, MA). 1 µL injections were loaded onto an Acclaim™ PepMap™ 100 C18 trap (75 µm × 20 mm, 3 µm particle size), washed with 6 µL of mobile phase A, and separated on an Easy-Spray™ C18 analytical column (50 µm × 150 mm, 2 µm particle size). Separations were performed at a flow rate of 300 nL/min using mobile phases A (0.1% formic acid in water) and B (0.1% formic acid in 90% acetonitrile/10% water) over a 107-min gradient: 2 to 3% B in 1 min, 3 to 11% B in 75 min, 11 to 39% B in 15 min, ramped to 100% B in 1 min, and held at 100% B for 15 min. Positive mode nano-electrospray ionization was achieved at a voltage of 1.9 kV. Full MS survey scans were acquired at a resolution of 70,000 at *m/z* 200, with a maximum ion accumulation time of 30 ms, and a scan range between *m/z* 380 and 1800. Data-dependent MS/MS scans were triggered for the 10 most abundant ions at an intensity threshold of 3.3×10^4 and acquired at a resolution of 17,500 with a maximum ion accumulation time of 60 ms, dynamic exclusion of 24 s, and an isolation window of 2 Da. HCD fragmentation was performed at a collision energy setting of 27. Singly-charged species and ions with unassigned charge states were excluded from MS/MS.

3.2.3 Protein and Peptide Identification

Mass spectral data were imported into PEAKS Studio 8.5 (Bioinformatics Solutions Inc., Waterloo, ON, Canada) for peptide identification via *de novo* sequencing and subsequent database searching. Precursor ion mass tolerance was selected as ±20 ppm, while a mass error of 0.05 Da was allowed for fragment ions. A list of 313 potential post-translational modifications,

which includes cysteine carbamidomethylation, methionine oxidation, and asparagine and glutamine deamidation, was allowed as variable modifications for peptide identification, based on results described in Chapter 2, Section 2.2.3 of this dissertation. The maximum number of PTMs allowed per peptide was three, and a combined total of 3 tryptic missed cleavages on either end of the peptide were permitted. All *de novo*-sequenced peptides with a confidence score (-10lgP) greater than 15% were matched to protein sequences in a reference database. To capture non-mutated proteins in the samples, the UniProtKB SwissProt Human protein database (downloaded September 21, 2017) was used for protein inference from identified peptides[8]. A second protein and peptide identification using the same raw mass spectral files was performed in PEAKS, where *de novo*-sequenced peptides were searched against individualized protein databases created from exome sequence data from the donor who provided the hair sample (see Chapter 2, Section 2.3.3 of this dissertation for more details). This PEAKS analysis enabled a focused search for proteins with expected mutations to identify GVPs in each sample. Each individualized protein database contains protein sequences from a list of 691 common gene products found in hair with the appropriate mutations expected in an individual based on their exome sequences. GVPs identified from each hair specimen were matched to mutated protein sequences in individualized protein databases in this peptide identification analysis.

Identified proteins and peptides from both PEAKS analyses were further filtered with a 1% false discovery rate threshold for peptide-spectrum matches and then exported from PEAKS. An in-house Python-based script was applied to the output files to merge results from the two PEAKS analyses and generate a non-redundant protein profile for each sample. Protein profile metrics include the number of proteins, unique peptide sequences, amino acids, and SNPs identified from both major and minor GVPs.

3.2.4 Label-Free Protein Quantification

Examination of differences in hair protein expression levels among head, arm, and pubic hair requires measurement of protein abundances. An automated, label-free protein quantification approach was applied in this work, where proteins were quantified using only unique peptides identified in the PEAKS analysis. Shared peptides between different proteins, that is, peptides that cannot be uniquely attributed to one human protein, and thus identified as belonging to multiple proteins in the output files from PEAKS analyses, were excluded from quantification. Using an in-house Python script, the extracted ion chromatogram peak area corresponding to the precursor *m/z* of each unique peptide in the MS1 survey scan was quantified from the raw mass spectral file using the *m/z* and retention time identified in the PEAKS analysis; isotopic ions were not quantified. Only ions with an abundance greater than 1000 counts in consecutive survey spectra within 0.05 min of one another were included, as inclusion of consecutive spectra beyond this retention window may incorporate signal from isobaric precursors. Optimization and validation of this automated method were performed by comparing results to those obtained via manual integration using Thermo Xcalibur Qual Browser (version 3.1.66.10).

To maximize quantification of the entirety of a peak but minimize interference from overlapping isobaric precursor ions so that peak integration for protein quantification using this automated approach approximates manual peak integration, a mass error tolerance of 5 ppm and a flexible retention time window were selected (Figure 3.1). A mass error tolerance of 5 ppm represents the smallest window that permits automated and manual peak integration to converge on similar peak areas observed among the range of mass error tolerances examined in Figure 3.1a (from 1 ppm to 50 ppm). A lower peak area was observed after applying a 1-ppm mass error

tolerance, indicating that peak integration achieved at 1 ppm did not encompass the entire peak and was thus, too restrictive. However, among the mass error windows from 5 ppm to 30 ppm, selection of a conservative mass error window (at 5 ppm) avoids integration of isobaric precursor ions.

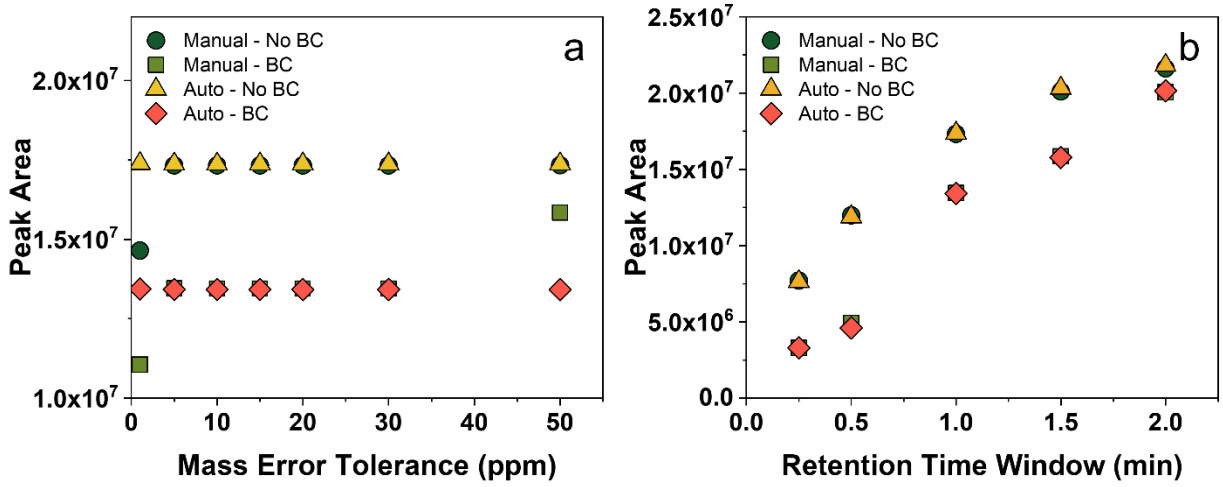


Figure 3.1. Optimization of (a) mass error tolerance and (b) retention time window for automated protein quantification using unique peptide AFDQDGDGHITVDELRR from CALML5 in one sample. Automated (Auto) and manual peak integration were compared before and after applying baseline subtraction as a baseline correction method (BC). A mass error tolerance of 5 ppm and a flexible retention time window were selected as optimal parameters to maximize peptide peak quantification and minimize noise from overlapping isobaric precursor ions.

A flexible retention time window was implemented to allow automated integration of the entire chromatographic peak without prior knowledge of chromatographic peak widths, which may differ among ions in the same run and are influenced by ion abundance and interaction with the column stationary phase. The retention time window (chosen from any of 0.25 min, 1 min, or 2 min) determined for peptide quantification represents the window that yielded the smallest difference in peak area between the uncorrected value and the value after baseline subtraction. This process included baseline subtraction to remove signal contribution from noise and background ions. To perform baseline subtraction, the baseline of each survey scan extracted ion

chromatographic peak was first defined, as a line extending between the first and last ions of the peak, where ion counts of both exceeded the threshold value of 1000. The area below the baseline, trapezoidal in shape, was then determined. Baseline subtraction was performed by subtracting the trapezoidal area below the baseline from the survey scan extracted ion chromatographic peak area. For example, in Figure 3.1b, a 2-min retention time window for the peptide AFDQDGDGHITVDELRR from CALML5 yielded the smallest difference in peak area before and after baseline subtraction. Convergence of peak area between the uncorrected and baseline subtracted values with a 2-min retention time window indicates that the entire peak was integrated, whereas a smaller retention time window only captures a portion of the peak equidistant from the peak apex. The baseline area should be minimal if the entire peak is captured, given that at the base, the ion intensities of the first and last ions that bound the peak will be low. Therefore, to quantify this peptide, a 2-min window was selected. To quantify other peptides, this optimization was performed to select the appropriate retention time windows of 0.25, 1, or 2 min.

This automated method was validated for analytical reliability using a repeated measures ANCOVA (by fitting a linear mixed-effects model via the *lme* function in the *nlme v3.1-141* package in R) for the 8 unique peptides identified as belonging to CALML5 in one sample. While the application of baseline subtraction results in statistically smaller peak areas ($p = 6.92 \times 10^{-8}$; n = 352 data points across all conditions of the four variables: integration method, baseline correction, mass error tolerance, and retention time window), peak areas determined using this automated method are not different from those quantified via manual integration ($p = 0.136$), for both areas calculated with or without baseline subtraction. This result indicates that the described method for peptide quantification is statistically indistinguishable from manual integration and

that automated peak integration can be implemented in lieu of manual integration to facilitate quantification of thousands of peptides identified from each single hair sample.

To account for run-to-run chromatographic variation in total peak area for each sample, which may arise from hair-to-hair differences in the amount of material loaded onto the chromatographic column, each peptide peak area was normalized to the total peak area (from all identified peptides, including shared peptides) and then rescaled so that the total peak area is equal to the average total peak area of the entire dataset. Each protein identified in a sample was then quantified by summing the normalized survey scan extracted ion chromatogram peak areas of their respective unique peptide ions.

Biological reliability was assessed by statistically comparing the coefficients of variation in CALML5 abundance obtained from biological replicate arm hairs (n = 4) from each individual (n = 3). The coefficient of variation from each individual was determined to be equivalent (p = 0.709) using the asymptotic test (*cvequality* package in R); similar protein abundance variation among biological replicates between individuals for the same condition demonstrates the reliability of the described method for protein quantification.

3.2.5 GVP Profile Generation – Observed Phenotype Frequencies

Each GVP profile was established using the presence or non-detection of the major and minor GVPs at each SNP locus. Observed phenotype frequencies, derived from SNP genotype frequencies as described in Chapter 2, Section 2.3.7 of this dissertation, were used to represent the presence or non-detection of major and minor GVPs that imply a phenotype. Conventionally, SNPs are associated with population genotype frequencies, obtained from the Genome Aggregation Database (gnomAD)[9]. However, to account for uncertainty in establishing a genotype with proteomic responses from incomplete GVP detection, observed phenotype

frequencies were used as sums of genotype frequencies (i.e., sum of either major homozygote or minor homozygote genotype frequency with heterozygote genotype frequency) when only either a major or minor GVP was detected. Total population genotype frequencies from gnomAD (v2.1.1) were used; these frequencies were updated from those used in Chu et al.[1] For example, detection of only the minor GVP for a SNP resulted in an observed phenotype frequency as the sum of the heterozygote and minor homozygote frequencies. Observed phenotype frequency was not reported for absent GVPs (i.e., one true negative and one false negative response) as the SNP was not considered in that sample. For multi-allelic SNPs, Hardy-Weinberg equilibrium was assumed (see Chapter 2, Section 2.3.7 of this dissertation for more details).

3.2.6 Statistical Analysis

All statistical comparisons were performed in R (x64 version 3.4.4). Significance was established at $\alpha = 0.05$. Two-way ANOVAs were performed using the *aov* function after fitting a linear model in the *stats v3.5.0* package. Tukey HSD post-hoc tests, two sample t-tests, and tests for association of Pearson product-moment correlations were performed using the same package. Equal variances were not assumed for t-tests. For non-parametric Kruskal-Wallis and Dunn post-hoc tests, the *agricolae v.1.2-8* package and *dunn.test v.1.1.0* package, respectively, were used, and a Bonferroni correction was applied to adjust p-values. Principal components analysis (PCA) was performed in MATLAB (R2017a, MathWorks, Natick, MA). All plots were drawn in OriginPro 2018 (OriginLab Corp., Northampton, MA) except for MATLAB outputs; PCA plots were drawn in Microsoft Excel 2016 (Redmond, WA). All values are reported as mean ± s.d. unless otherwise specified.

3.3 Results and Discussion

3.3.1 Single Inch Hair Sample Preparation Performance

Single one-inch hairs yield rich protein profiles that are comparable to profiles established with greater hair quantities; on average, 142 ± 33 (mean ± s.d.) proteins were identified from each of 9 head hairs (i.e., from three sets of proteomics-only biological replicates from three individuals), and the average number of identified unique peptides was 1,031 ± 219. From unique peptides, the average numbers of identified amino acids were 15,527 ± 3,056. The presence of a subset of unique peptides known as genetically variant peptides (GVPs) enabled inference of 16 ± 5 SNPs from major GVPs, and 17 ± 3 SNPs from minor GVPs (i.e., GVPs corresponding to the major and minor alleles, respectively). Because both major or minor GVPs allow SNP inference, nonsynonymous, or missense, SNPs were reported for both types of GVPs. However, in some cases, detection of both GVPs for the same SNP may not be possible. In previous studies, Parker et al. identified at least 180 proteins in 10 mg of head hair samples from 60 subjects and detected between 156 and 2,011 unique peptides[10], and Adav et al. identified, on average, 195 ± 12 proteins in human hair using various sample preparation methods[11]. Commensurate performance to previous works is achieved even when sample size is substantially reduced to simulate amounts of material available from forensic samples.

In addition, performing co-extraction of protein and mitochondrial DNA (mtDNA) yielded no loss in protein information relative to processing for protein alone. Proteomic results from co-extraction were not statistically different from proteomics-only sample preparation for each of the above metrics (two sample t-test; $p \geq 0.106$; Appendix Figure S-3.1); for example, averaged from head, arm, and pubic hairs, 156 ± 56 proteins were identified from proteomics-only samples and 151 ± 39 proteins were detected in co-extracted samples. These observations

indicate that additional steps taken to co-extract DNA with protein did not adversely affect protein identification or detection of unique peptides and missense SNPs from GVPs. As both sample preparation methods yielded the same proteomic information, the protein/DNA co-extracted sample set was included in this study for all further analyses. Analysis of GVPs and mtDNA can provide corroborating evidence for more confident profiling of individuals, which will be explored in a later chapter.

3.3.2 Proteomic Variation at Different Body Locations

This study sought to empirically define body location-specific proteomic variation and its effects on GVP identification and subsequent SNP inference, and quantify the extent to which individuals are distinguished using single one-inch hairs from different body locations. Hair proteomic variation at three different body locations in 36 hair specimens was first assessed by comparing five metrics: the numbers of detected proteins, unique peptides, amino acids, and missense SNPs from major and minor GVPs (Figure 3.2). Two-way ANOVAs with Tukey HSD post-hoc tests were performed for each metric to account for effects of body location and individual. Statistical testing revealed significant effects of body location on the numbers of detected proteins ($p = 1.07 \times 10^{-4}$), unique peptides ($p = 5.66 \times 10^{-4}$), and amino acids ($p = 2.21 \times 10^{-3}$), while effects of individual and the interaction between body location and individual were not significant. A single inch of pubic hair yields more proteins, unique peptides, and amino acids, than head or arm hair. A significant effect of body location on the number of SNPs inferred from GVPs was observed for major ($p = 7.56 \times 10^{-3}$) and minor GVPs ($p = 1.91 \times 10^{-5}$). These results suggest that compared to head and arm hair, the protein composition of pubic hair is more complex, from which many GVPs and SNPs can be identified for human identification.

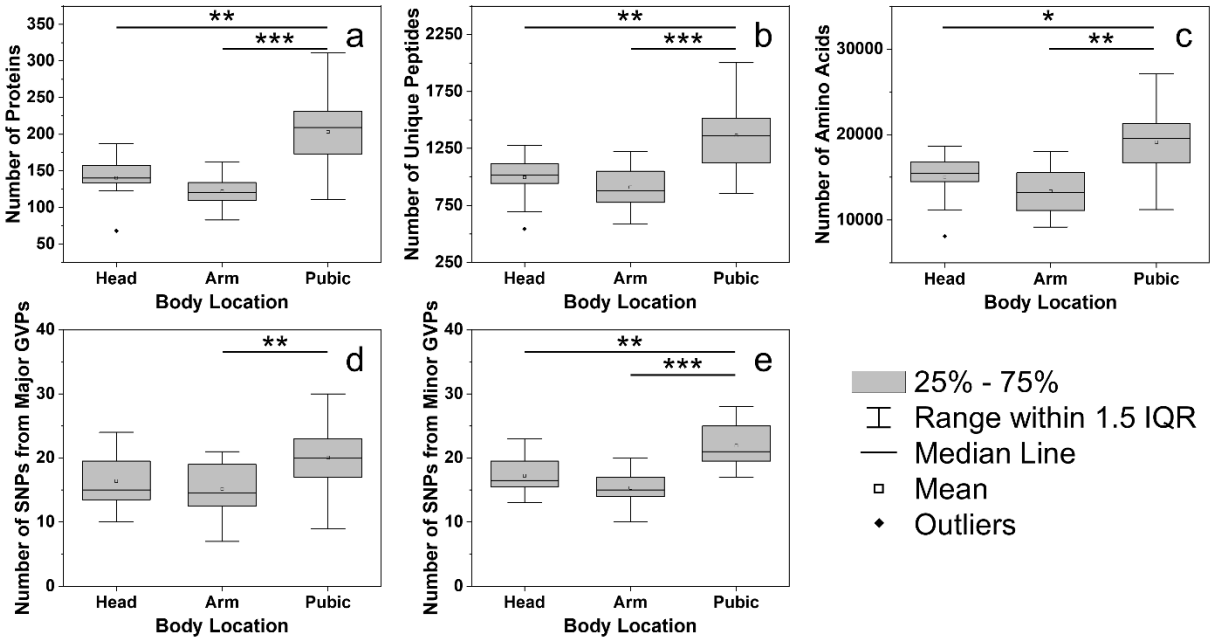


Figure 3.2. Comparison of numbers of identified (a) proteins, (b) unique peptides, (c) amino acids, and missense SNPs inferred from (d) major and (e) minor GVPs at different body locations. Black lines represent statistically significant comparisons and significance levels are represented as $p \leq 0.05$ (*), $p \leq 0.01$ (**), and $p \leq 0.001$ (***). Pubic hair samples yield statistically greater numbers of proteins, peptides, amino acids, and inferred SNPs (two-way ANOVA and Tukey HSD; n = 36).

Significant effects of body location observed for these five metrics may arise from differences in mass per unit length of hair. The mass of a single inch of pubic hair (200.1 ± 39.6 µg) was statistically greater than an inch of head (84.4 ± 27.7 µg; two-way ANOVA and Tukey HSD; $p = 1.76 \times 10^{-9}$) or arm hair (49.4 ± 22.2 µg; $p = 1.74 \times 10^{-11}$). Despite mass differences in hair, the same injection volume was used for each sample, and thus, different quantities of material were loaded onto the column for LC-MS/MS. It is proposed that more proteins, unique peptides, amino acids, and inferred SNPs were identified in pubic hair samples owing to larger on-column mass loadings.

To assess body location-specific proteomic variation and its effects on GVP identification without bias from different on-column mass loadings, protein abundances were quantified from identified peptides and compared. A 2014 study by Laatsch and co-workers reported differential

expression for a subset of hair proteins at different body locations using bottom-up proteomics and data-dependent mass spectrometry approaches, and quantified protein abundance by spectral counts[3]. To confirm these observations in this study for head and pubic hair, and to assess differential protein expression in arm hair, which was not examined previously, protein quantities derived from mass spectral data were compared. Various approaches have been used to quantify proteins using mass spectral data, including spectral counts[3, 12, 13], precursor ion peak areas from survey scans[14, 15], and MS/MS fragment ion abundances[16] to represent peptide abundance. Because dynamic exclusion was used during data acquisition to maximize peptide identification and protein coverage, MS/MS spectral counts may not reflect peptide abundance, especially for lower abundance peptides[17], where only one or two spectral counts may be obtained regardless of the order of magnitude of abundance. Of the latter two methods, protein quantification using precursor ion peak areas from survey scans was chosen as the simpler approach; use of the second method requires additional knowledge and selection of characteristic fragment ions from each identified peptide, which is not easily accessible in outputs from search engine results. As such, MS scan precursor ion peak areas in mass spectral data from a complete list of unique peptides identified from their MS/MS spectra were tabulated. Normalizing each precursor ion peak area to the total peak area of all identified peptides permits comparison of relative amounts of protein among hair samples of similar length but whose masses may vary, as it may not be practical to measure one-inch segments in more routine analyses. Protein abundance in each sample was calculated as the sum of all normalized peak areas assigned to the protein, as detailed in Section 3.2.4.

Protein abundance was examined in this study to determine effects of body location, as low protein abundance derived from body location-specific expression patterns may compromise

GVP identification in hair samples. Statistical comparison of protein abundances identified 37 proteins with body location-specific differential expression, of which a subset is shown in Figure 3.3 (two-way ANOVA and Tukey HSD). Even after accounting for mass differences, pubic hair remains the most protein-rich, both in composition and abundance, compared to head and arm hair; 89% of differentially expressed proteins, measured by normalized protein abundances, exhibit greater abundance in pubic hair and are least abundant in arm hair. Not surprisingly, keratins and KAPs comprised only 27% of body location-specific differentially expressed proteins (i.e., 10 proteins), in agreement with Laatsch et al.[3], while intracellular proteins such as FABP4, MIF, and ATP5B made up the majority (i.e., 27 proteins, or 73% of body location-specific proteins). Laatsch et al. reported differential expression levels for 12 proteins between head and pubic hair using spectral counts[3]. Of these 12 proteins, 6 proteins (K37, K38, KAP11-1, KAP13-1, KAP13-2, and KAP19-5) had statistically different abundances between head and pubic hair samples in the current study, indicating agreement between studies using different protein quantification methods.

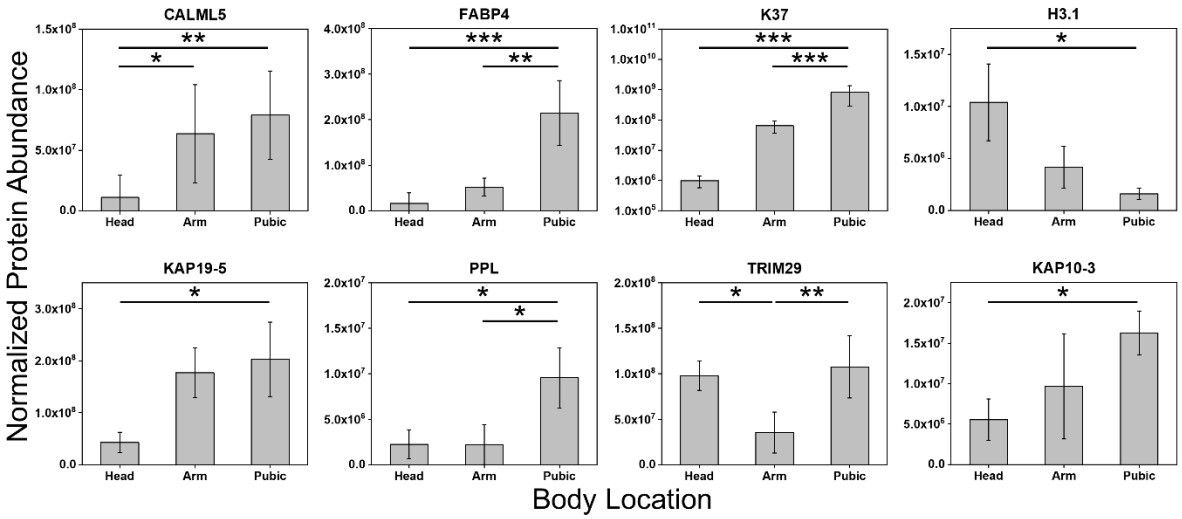


Figure 3.3. Average normalized abundances for a subset of eight differentially expressed hair proteins at different body locations (two-way ANOVA and Tukey HSD; n = 36). Error bars represent standard deviation from 4 replicate measurements of each of three individuals. Black lines represent statistically significant comparisons and significance levels are represented as $p \leq 0.05$ (*), $p \leq 0.01$ (**), and $p \leq 0.001$ (***).

Among the differentially expressed proteins, expression of K37 and K38 was notable as the only keratins in the set of 37 markers. Exhibiting differential expression between head and pubic hair, these keratins are also statistically more abundant in pubic hair compared to arm hair. The presence of these proteins in head hair, albeit at low levels for K37, agrees with findings of previous studies[3, 10, 11]. Unique peptides from K37 were observed in only 25% of head hairs, while the protein was consistently expressed in arm and pubic hair (83% and 100% identification, respectively). Because the hair samples prior to segmentation were of similar length (1 – 2 inches), failure to detect K37 in many head hairs cannot be attributed to degradation with age of hair, i.e., time since biosynthesis. Instead, K37 expression is linked to hair follicle keratinocyte differentiation differences at the various body locations; the protein is known to be expressed in the medulla of sexual hairs, including pubic hair, matured from unmedullated vellus hairs[18]. With the exception of K37 and K38, other hair keratins were not differentially expressed from the various body locations examined.

As keratins and KAPs primarily contribute to the highly conserved structural integrity of hair, it was considered unlikely that many hair structural proteins would exhibit differential expression at the various body locations. However, variation in KAP abundance observed in this dataset was more prevalent compared to keratins or other peripheral structural proteins. Of the 37 hair protein markers that exhibit body location-specific differential expression, 7 are KAPs (19%). For example, KAP19-5 is highly abundant in arm and pubic hair compared to head hair, between 4- and 5-fold greater, on average, and pubic hair is significantly enriched with many other KAPs including KAP11-1 and KAP10-3 (two-way ANOVA and Tukey HSD; Appendix Figure S-3.2). Differential expression of KAPs may be linked to the structural conformation of intermediate filaments and affect physicochemical properties (e.g., rigidity, tensile strength, thickness) of hair fibers, and can serve as useful markers to differentiate hair fibers from different body locations, as evident in the PCA scores plot (Figure 3.4).

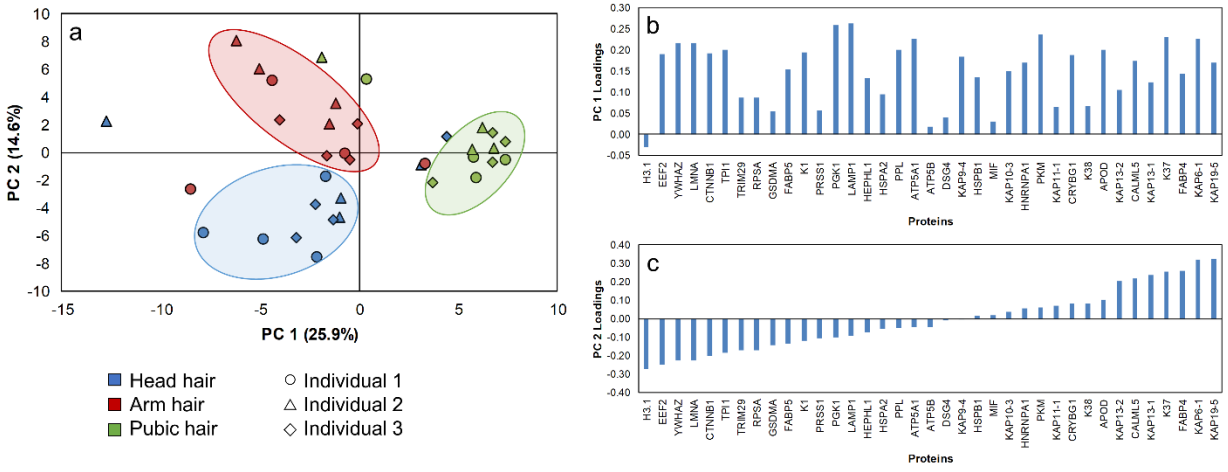


Figure 3.4. Principal components analysis (a) scores and loadings on PCs (b) 1 and (c) 2 using 37 differentially expressed hair proteins. Protein abundances were log-transformed and Pareto-scaled to reduce dominance of highly abundant proteins. This subset of proteins allows distinction of hair from different body locations; 75% of samples are captured in non-overlapping clusters.

Distinction of head, arm, and pubic hair is further enhanced by differential expression of intracellular proteins that dominate the hair proteome[19]. Many intracellular proteins are least

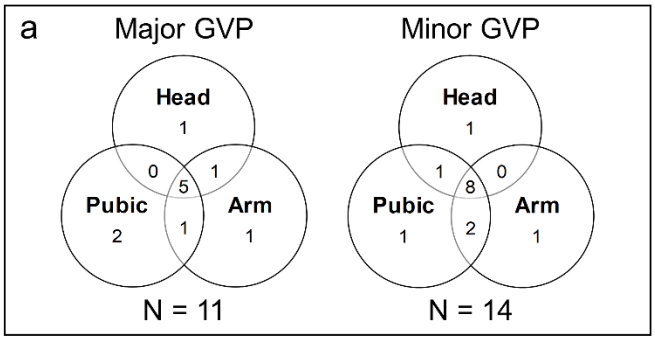
abundant in arm hair, although arm hair samples have notably high abundances of CALML5, GSDMA, and KAP19-5 compared to head hair samples. Protein expression is much more similar between head and arm hair; notably, histone H3.1 exhibits higher expression in head and arm hair compared to pubic hair and is a key protein in differentiating these hair types (Figure 3.4). Found in the nucleosome, this protein is an integral component in chromatin structure and is linked to DNA synthesis and repair[20], but also exhibits antimicrobial activity[11]. Protein abundance variation in 37 markers enabled distinction of hair fibers from different body locations via principal components analysis (Figure 3.4). Differential protein expression captured with protein abundance confirms proteomic variation in hair from different body locations and may be a valuable tool for screening in forensic investigations, but examination for any downstream effects on GVP and SNP detection is critically important to establish the power of this approach for forensic identification purposes.

3.3.3 Effects of Proteomic Variation on GVP Identification

Because protein abundances vary for a subset of hair proteins at different body locations and GVPs result from hair protein digests, it was considered that GVP identification may be affected by body location-specific differential protein expression. Therefore, it was imperative to examine the SNPs identified in each sample and determine whether differential protein expression affects success in GVP identification and subsequent SNP inference. Further comparison of identified SNPs in each sample was performed to determine whether some SNPs are only identified at specific body locations. Only SNP inferences consistent with an individual's genotype determined from exome sequencing were considered. SNPs with false positive responses are not robust candidates for a GVP panel and were removed; 65 SNPs remained for further analysis.

To observe effects of differential protein expression by body location on identification of SNPs, distributions of inferred SNPs from detected major and minor GVPs were compared across body locations. Of 65 SNPs, only exome-proteome consistent SNPs, in which the proteomic response corresponded with the exome response, i.e., true positive and true negative responses, across all 12 samples per body location for either major or minor GVPs, were retained (Figure 3.5). Figure 3.5a-b illustrates the amount of overlap in consistent SNPs across samples from different body locations. From 11 and 14 consistent SNPs identified from major and minor GVPs, respectively, 5 and 8 SNPs were identified at all body locations, which comprise the majority (on average, 69%) of exome-proteome consistent SNPs. For the remainder of SNPs in the minority, either the major or minor GVPs corresponding to these SNPs were not always identified among all four sample replicates for each body location and individual, resulting in SNP detection variability, which is further discussed below. But because the majority of exome-proteome consistent SNPs were identified among samples regardless of body location, it is likely that if a SNP is identified in samples within a body location, the same SNP is also identified in hair samples from a different body location, i.e., SNPs are not body location-specific. Only 11 SNPs in total are not identified at all body locations; there is one unreliably identified SNP (rs214803 from TGM3) that overlaps between major and minor GVPs. Response rates among the hair samples from this subset of SNPs were compared and correlated with respective protein abundances below to assess whether the unreliability in SNP identification derives from low protein abundance.

a



b

Body Location	SNPs from Major GVPs	SNPs from Minor GVPs
Head	7	10
Arm	8	11
Pubic	8	12
Total	11	14
Consistent across all samples	**5**	**8**

c

Protein	Differentially Expressed	SNPs with Unreliable Identifications
APOD	X	
ATP5A1	X	
ATP5B	X	
CALML5	X	
CRYBG1	X	
CTNNB1	X	
DSG4	X	
DSP		X
EEF2	X	
FABP4	X	
FABP5	X	
GSDMA	X	
HEPHL1	X	
HIST1H2AB		X
HIST1H3A	X	
HNRNPA1	X	
HSPA2	X	
HSPB1	X	
JUP		X
K1	X	
K32		X
K33B		X
K37	X	
K38	X	
K83		X
KAP10-12		X
KAP10-3	X	
KAP11-1	X	
KAP13-1	X	
KAP13-2	X	
KAP19-5	X	
KAP4-11		X
KAP6-1	X	
KAP9-4	X	
LAMP1	X	
LMNA	X	
MIF	X	
PGK1	X	
PKM	X	
PPL	X	
PRSS1	X	
RPSA	X	
TGM3		X
TPI1	X	
TRIM29	X	
VSIG8		X
YWHAZ	X	

Figure 3.5. Comparison and distribution of exome-proteome consistent SNPs across different body locations. (a) Distribution of inferred consistent SNPs across the three body locations for major and minor GVPs, respectively. (b) Summary of the number of consistent SNPs inferred from GVPs at each body location. (c) Comparison of differentially expressed proteins to proteins containing 11 SNPs with unreliable identifications at one or two body locations (i.e., not identified at all body locations). The majority of exome-proteome consistent SNPs identified at each body location were identified in all samples. Unreliably identified SNPs at either one or two body locations originate from a set of proteins that are not differentially expressed; there is no overlap between these sets of proteins. Therefore, the SNPs identified from this dataset are not body location-specific.

The possibility that body location-specific SNP localization results from proteomic variation was further examined by comparing subsets of proteins. The subset of 37 proteins with body location-specific differential expression was compared with the proteins of 11 inconsistently identified SNPs (Figure 3.5c). Any overlap in composition would suggest that differential expression of the protein affects downstream GVP identification and SNP inference within that protein. However, no overlap existed between differentially expressed proteins and

proteins containing unreliably identified SNPs. With the exception of five proteins (APOD, CALML5, GSDMA, K37, KAP10-3), SNPs were not identified in body location-specific differentially expressed proteins in this dataset. However, SNPs have been documented in these proteins, and as such, in a larger population, SNPs may be identified from these differentially expressed proteins, which may warrant their exclusion from consideration as sources for viable GVP markers. Nevertheless, despite significant positive correlations between the frequency of identifying SNPs from 3 of these proteins and protein abundance (Pearson product-moment correlation; $p \leq 0.043$; Appendix Figure S-3.3), identification of these SNPs remained variable among sample replicates, regardless of body location and protein abundance. This replicate variability indicates that their non-detection is more likely attributed to precursor ion competition for data-dependent MS/MS. Given the complexity of hair and that only the 10 most abundant ions per MS scan are fragmented, the GVP signal may not be sufficiently abundant in survey scans to be selected for MS/MS in some samples. Although the key advantage of using data-dependent mass spectrometry is its breadth of proteome coverage with minimal false positives, crucial in GVP discovery, slight run-to-run differences in chromatographic separation may result in irreproducibility in peptide selection for fragmentation[21, 22], contributing to unreliable SNP identification. Future development and operational use of targeted mass spectrometry approaches for GVP panels is expected to improve reliability of SNP identification.

In sum, all exome-proteome consistent SNPs derive from hair proteins that did not exhibit body location-specific differential protein expression. Comparison of differentially expressed proteins by body location with proteins carrying unreliably identified SNPs showed no overlap between the two sets of proteins, indicating that variability in SNP detection is unrelated to differential protein expression. Furthermore, no positive correlation between SNP

identification frequency and protein abundance was found for unreliably identified SNPs (Appendix Figure S-3.3), suggesting that body location-specific differential expression is not linked to SNP identification for all exome-proteome consistent SNPs. Therefore, while levels of APOD, GSDMA, and K37 display positive correlation with SNP identification (Pearson product-moment correlation coefficient $r \geq 0.71$), the vast majority (on average, 97%) of GVP identifications were not associated with differential protein expression, especially when the peptides were consistently identified among sample replicates. This investigation concludes that SNP identification in hair specimens is not dependent on body location, similar to the findings reported in a concurrent study[4]. GVP identification from protein digests of hair specimens is equally viable regardless of body location origin and all detected GVPs are candidates for a GVP panel.

3.3.4 GVP Candidates for Human Identification Panel

A series of criteria were established to evaluate GVP candidates from this dataset for a robust panel for human identification; these criteria can form the basis for criteria used to select GVP markers for inclusion in a human identification panel in future studies and routine operation. In particular, GVPs included in a human identification panel should be easily detected from mass spectrometry analysis but only reported in accordance with the appropriate SNP genotype, that is, GVP candidates should be consistently detected when the corresponding SNP genotype indicates its presence and should not exhibit false positives. As such, first, only GVPs that indicate exome-proteome consistent SNPs were considered. This criterion is necessary in GVP discovery and for marker validation to evaluate performance, with the intent of selecting markers for application to routine forensic analysis, but it is expected that in practice, exome sequence information will not be needed or available. In addition, only consistent SNPs

identified in all samples from this dataset were selected, as these SNPs, when identified in future studies and routine operation in forensic analyses, would have the lowest false negative rates and their GVP counterparts would have the highest chance of being consistently detected. After de-duplicating the list of SNPs inferred from detection of major and minor GVPs, 12 SNPs remained for consideration. SNP identifiers, the two most abundant forms of the GVP, and their MS scan precursor ion abundances are reported in Table 3.1. See Appendix Table S-3.1 for a complete list of GVPs.

Table 3.1. SNP and GVP candidates for GVP panel. *No SNP identifier associated with SNP (HGVS notation used); Larger, bold red text denotes location of amino acid variant in genetically variant peptide; †Preceding amino acid in peptide sequence denoted by "X."

Gene	SNP Identifier	Amino Acid Polymorphism	GVP Type	Peptide†	PTM	Average Normalized Abundance	Observation Frequency
FAM83H	rs9969600	Q/H	Minor	R.VNL**H**HVDFLR		1.10×10^6	2
KRT32	rs2071563	T/M	Major	R.ARLEGEIN**T**YR	A1:Formylation	5.72×10^7	27
KRT32	rs2071563	T/M	Major	R.LEGEIN**T**YR		1.71×10^8	28
KRT32	rs2071563	T/M	Minor	R.ARLEGEIN**M**YR	M9:Oxidation (M)	4.79×10^7	10
KRT32	rs2071563	T/M	Minor	R.LEGEIN**M**YR	M7:Oxidation (M)	1.69×10^7	7
KRT33A	rs148752041	D/H	Minor	R.**H**NAELENLIR	N2:Deamidation (NQ)	4.38×10^7	8
KRT33A	rs148752041	D/H	Minor	R.**H**NAELENLIRER		1.11×10^8	7
KRT33B	17:g.41366553 G>T *	L/M	Major	R.ILDE**L**TLCRSDLEAQM ESLKEELLSLKQNHEQEV NTLR	C8:Carbamidomethylation; N29:Deamidation (NQ)	9.11×10^7	12
KRT33B	17:g.41366553 G>T *	L/M	Major	R.ILDE**L**TLCRSDLEAQM ESLKEELLSLK	C8:Carbamidomethylation	1.49×10^7	11
KRT33B	17:g.41366553 G>T *	L/M	Minor	R.ILDE**M**TLCR	C8:Carbamidomethylation	7.10×10^7	10
KRT33B	17:g.41366553 G>T *	L/M	Minor	R.RILDE**M**TLCR	C9:Carbamidomethylation	2.20×10^7	7
KRT81	rs2071588	G/R	Minor	R.GLTGGFGSHSVC**R**	C12:Carbamidomethylation	3.35×10^8	19
KRT81	rs2071588	G/R	Minor	L.TGGFGSHSVC**R**	C10:Carbamidomethylation	1.18×10^7	13
KRT83	rs2852464	I/M	Major	R.DLNMDC**I**VAEIK	C6:Carbamidomethylation	1.08×10^8	32
KRT83	rs2852464	I/M	Major	R.DLNMDC**I**VAEIKAQYD DIATR	K12:Carbamylation	1.83×10^8	23
KRT83	rs2852464	I/M	Minor	R.DLNMDC**M**VAEIK	C6:Carbamidomethylation	2.97×10^7	19
KRT83	rs2852464	I/M	Minor	R.DLNMDC**M**VAEIKAQY DDIATR	C6:Carbamidomethylation; M7:Oxidation (M)	4.29×10^6	16
KRTAP10-3	rs233252	C/Y	Minor	R.ST**Y**CVPIPSC	C4:Carbamidomethylation; C10:Carbamidomethylation	2.97×10^6	4
KRTAP10-3	rs233252	C/Y	Minor	R.ST**Y**CVPIPS	C4:Carbamidomethylation	1.65×10^6	2
KRTAP10-9	rs9980129	R/C	Minor	C.CAPTSS**C**QPSYCR	C1:Carbamidomethylation; C7:Carbamidomethylation; C12:Carbamidomethylation	6.99×10^6	10

Table 3.1 (cont'd)

Gene	SNP Identifier	Amino Acid Polymorphism	GVP Type	Peptide†	PTM	Average Normalized Abundance	Observation Frequency
KRTAP4-11	rs760092771	S/C	Major	R.TTYCRPSCCVS**S**	C4:Carbamidomethylation; C8:Carbamidomethylation; C9:Carbamidomethylation	1.75×10^8	5
KRTAP4-11	rs760092771	S/C	Minor	R.TTYCRPSYSVS**C**C	C4:Carbamidomethylation; C12:Carbamidomethylation; C13:Carbamidomethylation	1.32×10^8	12
KRTAP4-11	rs760092771	S/C	Minor	R.TTYCRPSYSVS**C**	C4:Carbamidomethylation; C12:Carbamidomethylation	8.10×10^7	11
KRTAP4-11	rs763737606	C/S	Major	R.TTYCRPSC**C**VSS	C4:Carbamidomethylation; C8:Carbamidomethylation; C9:Carbamidomethylation	1.75×10^8	5
KRTAP4-11	rs763737606	C/S	Minor	R.TTYCRPSY**S**VSCC	C4:Carbamidomethylation; C12:Carbamidomethylation; C13:Carbamidomethylation	1.32×10^8	12
KRTAP4-11	rs763737606	C/S	Minor	R.TTYCRPSY**S**VSC	C4:Carbamidomethylation; C12:Carbamidomethylation	8.10×10^7	11
KRTAP4-11	rs774046661	C/Y	Major	R.TTYCRPS**C**CVSS	C4:Carbamidomethylation; C8:Carbamidomethylation; C9:Carbamidomethylation	1.75×10^8	5
KRTAP4-11	rs774046661	C/Y	Minor	R.TTYCRPS**Y**SVSCC	C4:Carbamidomethylation; C12:Carbamidomethylation; C13:Carbamidomethylation	1.32×10^8	12
KRTAP4-11	rs774046661	C/Y	Minor	R.TTYCRPS**Y**SVSC	C4:Carbamidomethylation; C12:Carbamidomethylation	8.10×10^7	11
VSIG8	rs62624468	V/I	Major	R.LGCPY**V**LDPEDYGPNG LDIEWMQVNSDPAHHR	C3:Carbamidomethylation; N15:Deamidation (NQ)	1.36×10^7	18
VSIG8	rs62624468	V/I	Major	R.LGCPY**V**LDPEDYGPNG LDIEWMQVNSDPAHHRE NVFLSYQDKR	C3:Carbamidomethylation; N15:Deamidation (NQ); M22:Oxidation (M)	6.88×10^6	14
VSIG8	rs62624468	V/I	Minor	R.LGCPY**I**LDPEDYGPNGL DIEWMQVNSDPAHHR	C3:Carbamidomethylation; N15:Deamidation (NQ); M22:Oxidation (M)	4.92×10^6	5
VSIG8	rs62624468	V/I	Minor	R.LGCPY**I**LDPEDYGPNGL DIEWMQVNSDPAHHREN VFLSYQDKR	C3:Carbamidomethylation; N15:Deamidation (NQ); M22:Oxidation (M)	2.90×10^6	3

GVP candidates shown in Table 3.1 represent the peptides observed with the highest frequency in this set of single hairs, detected in at least 2 of 36 specimens. Notably, most GVPs are tryptic; however, many peptides from KAPs are non-tryptic. Instead, peptides from KAP4-11 and KAP10-3 show evidence of proteolysis at the C-terminal sides of cysteines or serines. It is perhaps not surprising that non-tryptic GVPs are identified in KAPs, as there are few sites for trypsin to cleave in these cysteine- and serine-rich proteins, and the frequency of identifying GVPs from KAPs has been low compared to keratins[23]. Use of a combination of proteases that cleave at different sites may improve GVP yields from KAPs. Additionally, GVPs identified in structural proteins, such as the peptide ILDE**L**TLCRSDLEAQMESLKEELLSLK from K33B (mutation site denoted in larger, bold red text), also contained missed cleavages, where the peptide contains one or more trypsin cleavage sites within its sequence. The presence of these peptides points to incomplete protein digestion by trypsin. Lower digestion efficiency in keratins is likely linked to their natural existence as coiled-coil structures through hydrophobic interactions[24]. Hair keratins are stabilized as coiled-coil heterodimers with surfactant[25]; removal of surfactant sodium dodecanoate during sample preparation likely resulted in the unfolding (or refolding) of hydrophobic proteins in an environment thermodynamically unfavorable for monomeric forms, leading to protein aggregation[25] that reduces digestion efficiency. However, a less efficient digestion can be beneficial when identifying unique peptides among these highly paralogous proteins[26, 27], as longer peptides that contain missed cleavage sites may include regions proximal to the shorter, shared tryptic peptides that differentiate paralogous proteins, thus permitting identification of unique peptides.

The second criterion used to evaluate GVP candidates in Table 3.1 is marker independence for random match probability (RMP) determination at the SNP level. To assess the

performance of a robust panel for forensic identifications in a population, random match probabilities are calculated as the product of genotype frequencies for each SNP locus. However, genotype frequencies for correlated SNPs, i.e., SNPs in linkage disequilibrium[28, 29], may be biased in the population, which violates the assumption of marker independence for RMP calculations. To reduce the effect of possible disequilibria, a conservative one-SNP-per-gene rule was adopted; though not implemented here, more sophisticated treatment of linkage disequilibrium will allow for inclusion of more GVPs, and thus, lower RMPs. For multiple SNPs from a gene where mutation frequency in a population is known, the SNP with the lowest minor allele frequency was selected, as this contributes to lower RMP values and greater discriminative power. Finally, SNPs without Reference SNP IDs were also not considered further, as genotype frequencies are not known for these candidates. After applying these criteria, 8 SNPs remained for inclusion in a panel from 245 GVPs, from which GVP profiles can be assembled and compared among the three individuals for differentiative potential.

3.3.5 GVP Profiles and Identification Performance

Highly consistent GVP profiles for each sample were established using 8 SNPs. Each GVP profile was established using the presence or non-detection of the major and minor GVPs at each SNP locus. Figure 3.6 displays a simplified version of each profile by using observed phenotype frequencies associated with SNP genotypes to represent the presence of major and minor GVPs, as described in Chapter 2, Section 2.3.7 of this dissertation. The full set of profiles that denotes the presence or non-detection of GVPs is found in Appendix Figure S-3.4.

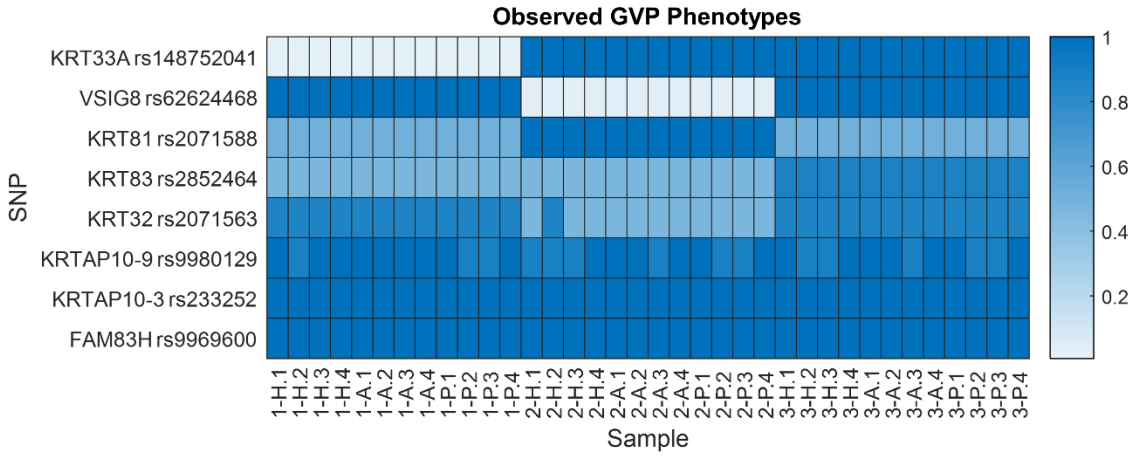


Figure 3.6. GVP profile of 36 samples using observed phenotype frequency to represent the presence or non-detection of major and minor GVPs at 8 SNP loci. Samples are denoted x-y.z, where x is the individual code (of 3 individuals), y represents the body locations from which the samples derived, head (H), arm (A), or pubic (P) regions, and z is the sample replicate. Profiles within an individual are similar, indicating consistent identification of SNPs with robust GVPs.

Comparison of GVP profiles pairwise and quantification of the number of profile differences from these pairwise comparisons showed that, irrespective of body location, GVP profiles are more similar within an individual than those between individuals. Pairwise comparisons of GVP profiles between individuals and replicates allowed quantification of profile similarity, using the number of observed phenotype differences across 8 SNP loci, termed GVP profile differences. For example, when comparing GVP profiles between 1-H.1 and 2-H.3 (head hair samples from Individuals 1 and 2, respectively), a response of 0 for sample 1-H.1, which indicates detection of the major GVP, differs from a response of 0,1 from 2-H.3, which indicates detection of the major and minor GVPs, for the SNP rs2071563; the difference observed at this SNP locus is counted and aggregated with observed differences for the other 7 loci. The number of differences was summed across 8 SNP loci for each pairwise comparison, totaling 630 comparisons (i.e., $(n-1) \times \frac{n}{2}$ for n = 36 hair samples). Replicate comparisons, performed between hair specimens from the same individual and body location, yielded 1.2 ± 1.0 (mean ± s.d.) GVP profile differences (n = 54 profile comparisons), and within-individual comparisons,

between hair samples from the same individual but different body locations, showed 1.1 ± 0.9 differences (n = 144 profile comparisons). Both types of intraindividual variation can be attributed to variability in peptide ion selection with the data-dependent mass spectrometry approach used in this work, discussed in Section 3.3.3, which may result in GVP non-detection among samples. As expected, between-individual comparisons exhibited the greatest number of GVP profile differences, with 4.9 ± 0.8, 5.1 ± 0.9, and 2.8 ± 0.7 differences, respectively, between Individuals 1-2, 2-3, and 1-3 (Figure 3.7a). All observed profile differences approximate expected GVP profile differences, that is, observed average differences are within 1 profile difference of expected values (Figure 3.7b). Greatest profile variation lies between individuals (Kruskal-Wallis test; $p = 2.96 \times 10^{-108}$), demonstrating that despite some sample replicate and within-individual variation (e.g., body location), distinct GVP profiles are observed in samples from different individuals.

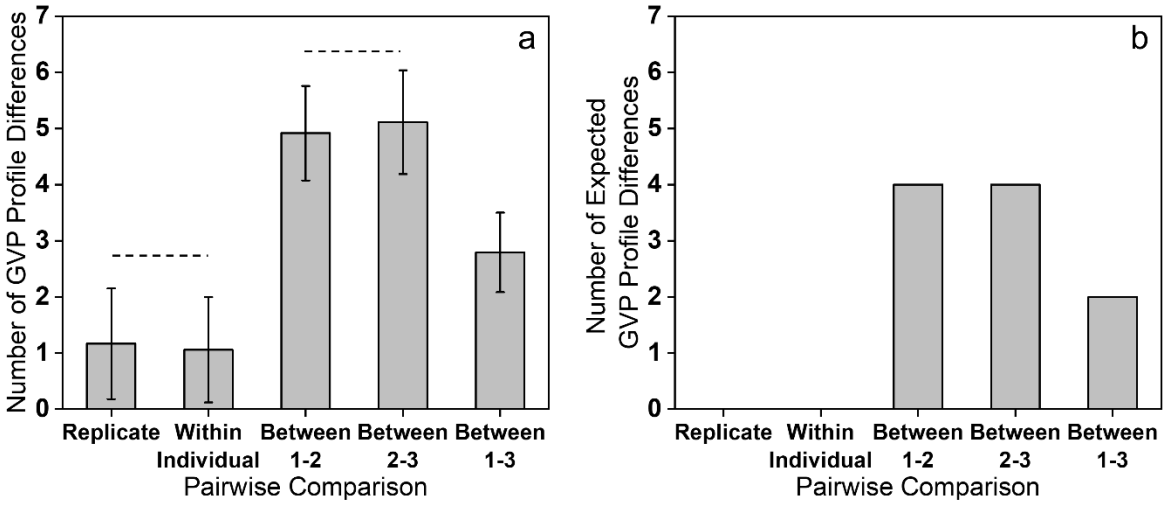


Figure 3.7. (a) Average number of GVP profile differences from different pairwise comparison categories compared to (b) expected number of GVP profile differences. Error bars represent the standard deviation. All but two (---) comparisons are statistically significant (Kruskal-Wallis and Dunn tests; n = 630; $p \leq 3.80 \times 10^{-6}$). The numbers of observed profile differences approximate expected GVP profile differences. Between Individual profile differences are statistically greater than Replicate and Within Individual profile differences.

Not surprisingly, the majority of GVPs derive from keratins and KAPs, though markers from keratins provide more discriminative power with respect to phenotype frequencies for RMP calculations. Greater discriminative power of GVPs from keratins is attributed to consistent identification of minor GVPs, as opposed to the more variable identification of peptide markers in KAPs for reasons discussed above. Sporadic GVP identification also contributes to sample replicate and within-individual variation, although, as expected, interindividual variation in GVP profile differences is the predominant type of variation observed among the entire suite of GVP profile differences, ranging between 3 and 5 times that of intraindividual variation.

Furthermore, RMPs, derived as products of observed phenotype frequencies from GVP profiles of each sample, align with the individual (Figure 3.8). Experimental RMPs range between 1 in 3 and 1 in 870, within an order of magnitude of expected RMPs for each individual. In contrast to the RMPs reported previously from GVP analysis[10, 30], the discriminative power attained in this study represents a lower range, owing to the strict criteria applied for selection of the exome-proteome consistent GVP markers among sample replicates to compare marker detection among hair samples from different body locations. Implementation of targeted mass spectrometry methods to obtain better reproducibility in GVP detection for routine forensic analysis of single hairs is expected to improve discriminative power. Most importantly, this work finds that GVP profiles of samples belonging to the same individual enable distinction of the individual to the same extent regardless of body origin, demonstrating not only body location invariance with a robust panel of inferred SNPs from GVPs, but also that the probative value of one-inch head, arm, and pubic hair samples is equivalent for an individual.

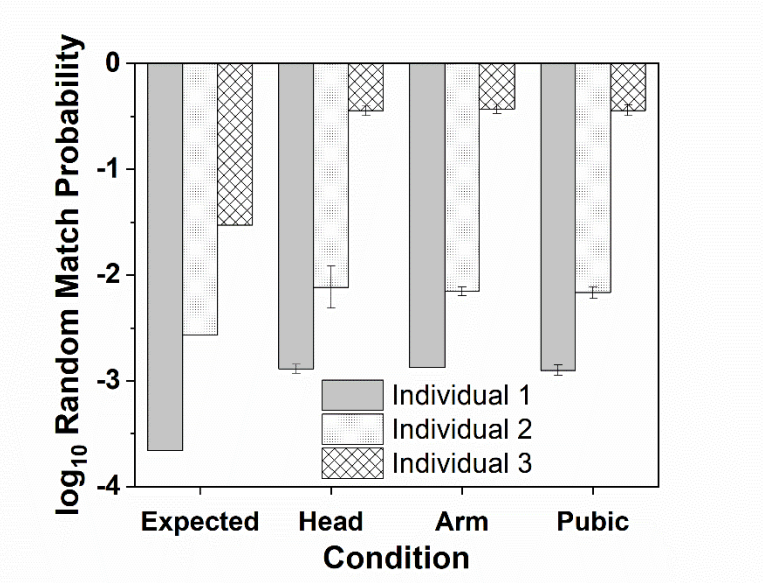


Figure 3.8. Experimentally observed random match probabilities (mean ± 95% CI) compared to expected RMP values for each individual. Expected RMPs are theoretically-derived values based on the detection of all GVPs consistent with an individual's genotype for the same 8 SNPs. RMP values of different body location samples from the same individual are not different; the extent to which individuals are distinguished from one another is not affected by hair origin. Observed RMP values from a robust set of SNPs approximate expected values within an order of magnitude.

3.4 Conclusions

This work demonstrates equivalent evidentiary value of head, arm, and pubic hair for protein-based human identification using genetically variant peptide markers, with GVP identification and SNP inference invariant across hair from different body locations. Furthermore, a set of robust SNPs inferred from exome-proteome consistent GVPs yielded similar potential to differentiate individuals from hair specimens irrespective of body location. The SNPs inferred from exome-proteome consistent GVPs represent a conservative pool of markers. It is expected that deeper interrogation of the hair proteome using data-dependent and data-independent mass spectrometry will achieve future measurements of substantially more markers. The criteria for GVP marker selection discussed herein represent a minimum set of criteria for choosing GVP markers for a human identification panel. Development of targeted

mass spectrometry methods for GVP identification, which is currently underway for use in routine forensic analyses, will reduce false-negative intraindividual variation that occurs using data-dependent protocols, improving reliability in GVP detection and discriminative power.

Characterization of body location-specific proteomic variation has not only improved understanding of intrinsic variation in hair protein chemistry, especially for arm hair whose protein abundance levels had not previously been elucidated, but also showed that these differences in protein levels enable differentiation of hair types. Pubic hair, richer in protein composition and abundance than head and arm hair, may be particularly valuable as a source of GVPs when applied to forensic investigations of sexual assault cases. These findings may also be applicable to further elucidate the underlying biochemical mechanisms responsible for the various pathologies associated with hair growth.

APPENDIX

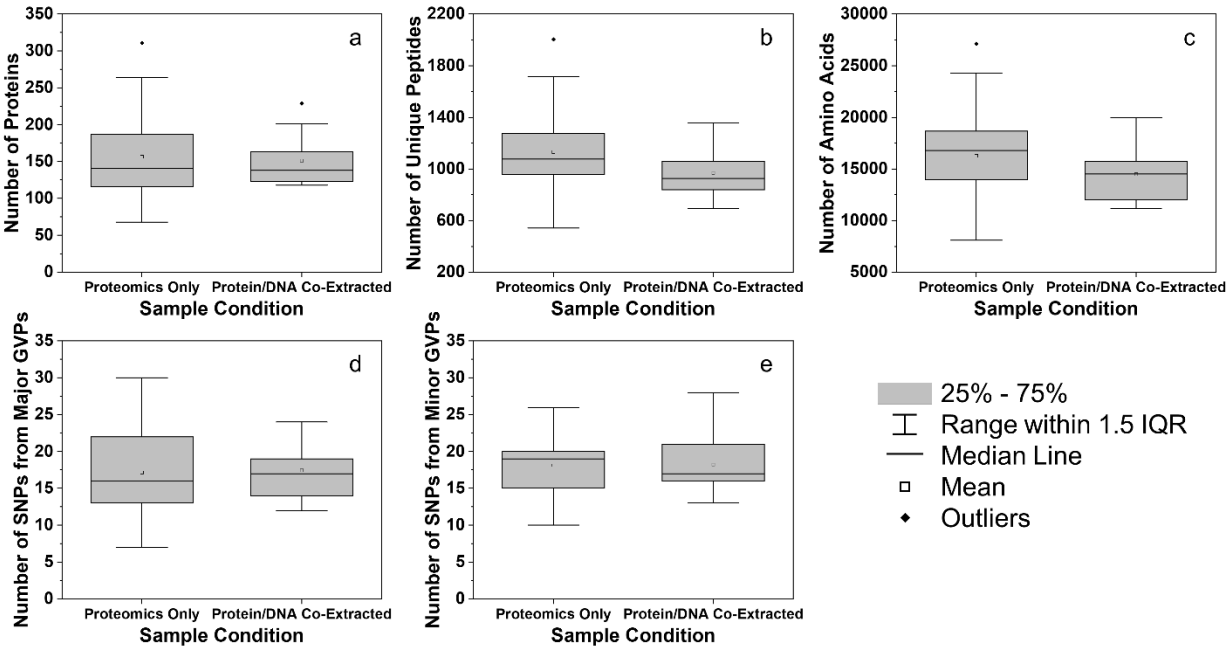


Figure S-3.1. Comparison of average numbers of identified (a) proteins, (b) unique peptides, (c) amino acids, and missense SNPs inferred from (d) major and (e) minor GVPs between Proteomics Only (n = 27) and Co-Extracted (n = 9) samples. Numbers of proteins detected in Protein/DNA Co-Extracted samples are not statistically different from Proteomics Only samples (two sample t-test; $p \geq 0.106$).

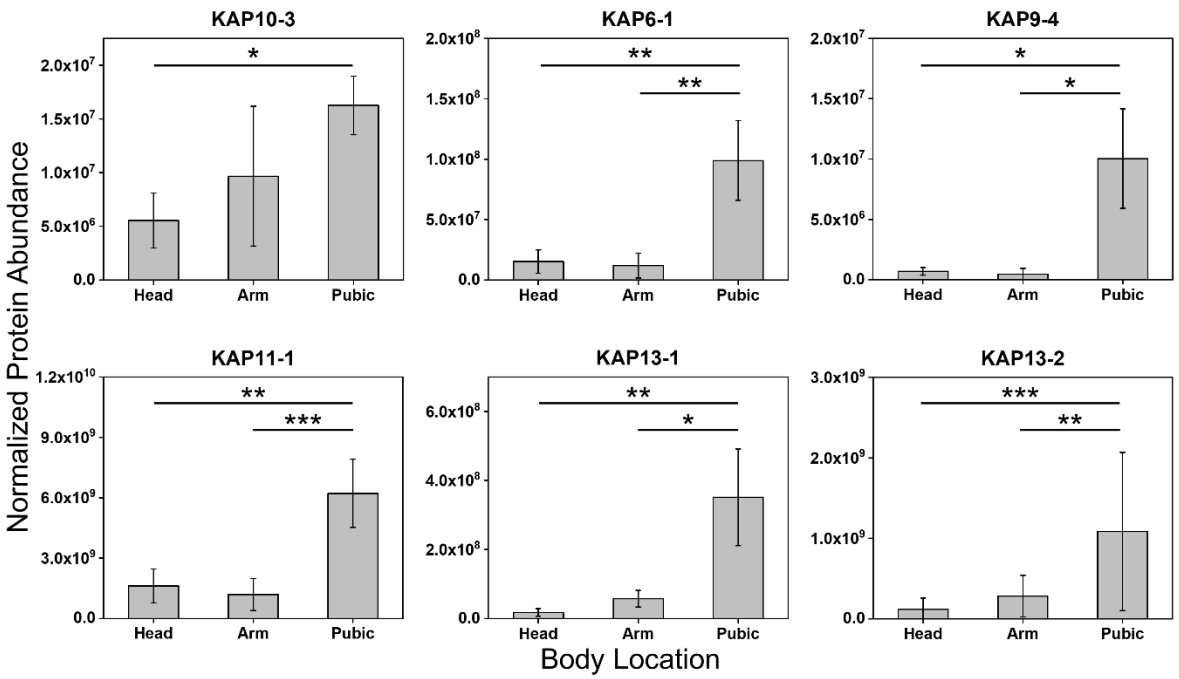


Figure S-3.2. Average abundances for a subset of differentially expressed hair proteins at different body locations (two-way ANOVA and Tukey HSD; n = 36). Error bars represent standard deviation from 4 replicate measurements of each of three individuals. Black lines represent statistically significant comparisons and significance levels are represented as $p \leq 0.05$ (*), $p \leq 0.01$ (**), and $p \leq 0.001$ (***).

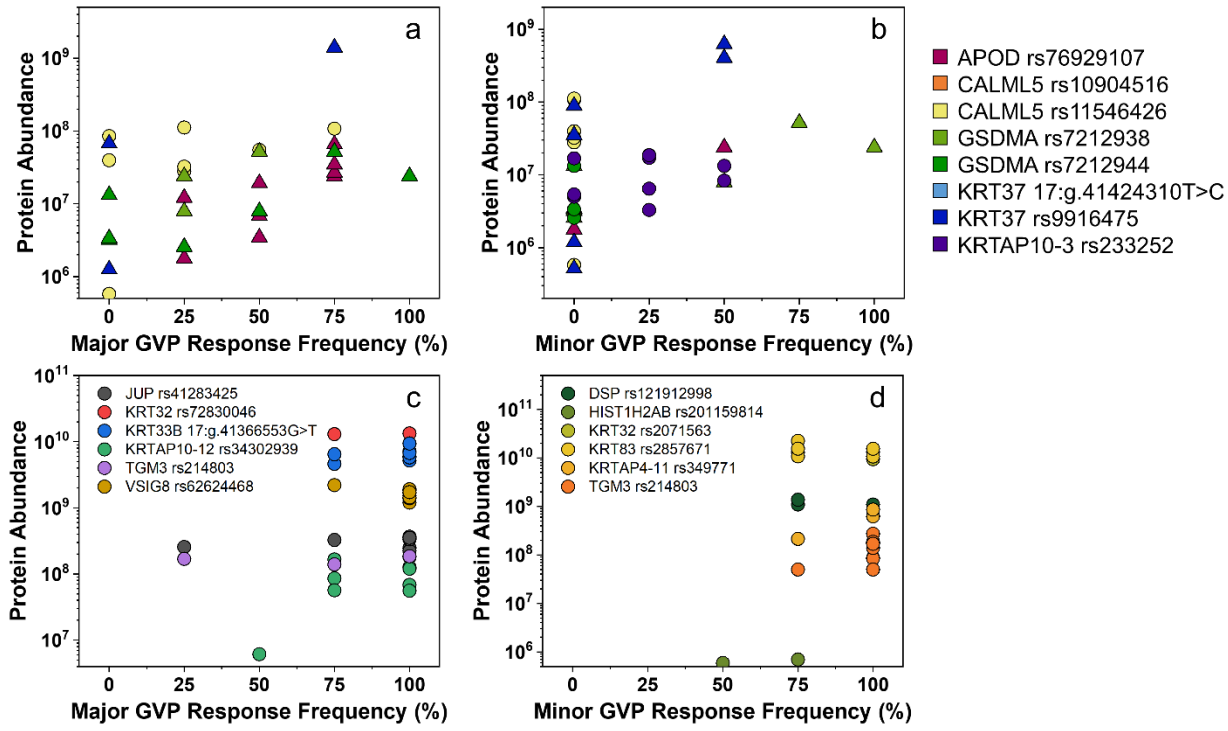


Figure S-3.3. Correlations between GVP response frequency and abundances of differentially expressed proteins for SNPs identified from (a) major GVPs and (b) minor GVPs. Identified SNPs in (a) and (b) are not exome-proteome consistent and display variation in sample replicates. (c) and (d) illustrate the relationship between GVP response frequency of unreliably identified exome-proteome consistent SNPs and protein abundance. Triangles denote significant positive correlations between GVP response frequency for a SNP and corresponding protein abundance (Pearson product-moment correlation; $n = 9$; $p \leq 0.043$). GVP responses show positive correlation with protein abundance for SNPs in APOD, GSDMA, and KRT37, but the majority of GVP identification is not affected by differential protein expression.

Table S-3.1. Complete list of SNP and GVP candidates for GVP panel.

Gene	dbSNP	Variant Type	Peptide	PTM	Avg Adj RT	RT Min	RT Max	Avg *m/z*	*m/z* Min	*m/z* Max	Avg Peak Area	Obs Freq
VSIG8	rs62624468	Minor	LGCPYILDPEDYGPNGLDIEWMQVNSDPAHHRENVFLSYQDKRINHGSLPHLQQR	C3:Carbamidomethylation; M22:Oxidation (M);Q23:Deamidation (NQ)	86.87	86.87	86.87	811.6431	811.6431	811.6431	2.88E+06	1
KRT83	rs2852464	Major	SRDLNMDCIVAEIK	C8:Propionamide	86.94	86.64	87.24	559.9474	559.9473	559.9474	1.19E+07	2
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATR	C6:Carbamidomethylation;Q14:Deamidation (NQ)	92.12	91.77	92.29	819.0556	819.0544	819.0569	1.05E+08	3
VSIG8	rs62624468	Minor	LGCPYILDPEDYGPNGLDIEWMQVNSDPAHHRENVFLSYQDKR	C3:Carbamidomethylation; N15:Deamidation (NQ);M22:Oxidation (M)	88.98	88.39	89.38	851.7304	851.7271	851.7332	2.90E+06	3
KRT81	rs2071588	Minor	GLTGGFGSHSVCR	C12:Carbamidomethylation	36.86	30.71	41.83	667.8165	667.8154	667.8176	3.35E+08	19
KRT81	rs2071588	Minor	GLTGGFGSHSVCR	T3:Dehydration;C12:Carbamidomethylation	36.44	30.63	40.33	439.5445	439.5425	439.5473	1.26E+07	10
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATR	M4:Oxidation (M);C6:Ethanolation (C)	91.78	91.70	91.87	819.7278	819.7238	819.7317	2.08E+07	2
KRT32	rs2071563	Major	ARLEGEINTYR	Y10:Dehydration	42.78	42.78	42.78	435.2304	435.2304	435.2304	2.36E+05	1
FAM83H	rs9969600	Minor	VNLHHVDFLR		56.84	53.53	60.14	417.2314	417.2309	417.2319	1.10E+06	2
KRT32	rs2071563	Minor	ARLEGEINMYR	M9:Oxidation (M)	43.08	36.54	46.48	456.5622	456.5611	456.5632	4.79E+07	10
KRT81	rs2071588	Minor	GLTGGFGSHSVCRGFR	R13:Arginine oxidation to glutamic semialdehyde	71.24	71.24	71.24	797.8712	797.8712	797.8712	1.14E+05	1
VSIG8	rs62624468	Major	LGCPYVLDPEDYGPNGLDIEWMQVNSDPAHHRENVFLSYQDKR	C3:Carbamidomethylation;Q23:Deamidation (NQ)	88.63	88.51	88.85	846.7279	846.7243	846.7329	1.37E+07	3
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	C8:Carbamidomethylation;K27:Acetylation (K)	90.53	90.53	90.53	914.2674	914.2674	914.2674	1.35E+08	1
VSIG8	rs62624468	Major	LGCPYVLDPEDYGPNGLDIEWMQVNSDPAHHRENVFLSYQDK	C3:Carbamidomethylation; N15:Deamidation (NQ);M22:Oxidation (M)	88.97	88.97	88.97	987.8554	987.8554	987.8554	1.89E+06	1
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATRSR	K12:Carbamylation	91.60	91.32	91.89	671.5756	671.5755	671.5756	6.99E+05	2
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATRSRAEAESWYR	N3:Deamidation (NQ);C6:Carbamidomethylation	92.09	92.09	92.09	738.9525	738.9525	738.9525	1.73E+06	1
KRT81	rs2071588	Minor	LTGGFGSHSVCR	C11:Carbamidomethylation	24.55	21.57	27.24	426.5396	426.5393	426.5398	6.22E+06	3
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	C8:Carbamidomethylation; Q28:Deamidation (NQ);N29:Deamidation (NQ)	89.49	89.49	89.49	1132.5847	1132.5847	1132.5847	2.53E+05	1
KRT32	rs2071563	Major	ARLEGEINTYR	N8:Beta-methylthiolation (ND)	43.05	36.54	46.48	456.5623	456.5611	456.5632	5.72E+07	10
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATRSRAEAESWYR	C6:Carbamidomethylation	91.65	91.51	91.79	738.7542	738.7527	738.7556	4.67E+05	2
KRT83	rs2852464	Minor	DLNMDCMVAEIK	M4:Oxidation (M); C6:Carbamidomethylation;M7:Oxidation (M)	59.10	59.10	59.10	735.8144	735.8144	735.8144	1.73E+07	1
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	C8:Carbamidomethylation;N29:Deamidation (NQ)	90.00	89.49	90.62	755.2196	755.2177	755.2234	9.11E+07	12
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATR	D1:Carbamylation;M4:Oxidation (M)	91.84	91.80	91.89	1228.5840	1228.5804	1228.5876	1.15E+06	2
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	C8:Carbamidomethylation; Q15:Deamidation (NQ);N35:Deamidation (NQ)	89.71	89.71	89.71	1132.5741	1132.5741	1132.5741	1.34E+07	1
KRT83	rs2852464	Major	DLNMDCIVAEIKA	M4:Oxidation (M)	86.80	86.41	87.70	725.8491	725.8484	725.8495	7.53E+07	4
VSIG8	rs62624468	Major	LGCPYVLDPEDYGPNGLDIEWMQVNSDPAHHRENVFLSYQDKR	C3:Carbamidomethylation;M22:Oxidation (M)	88.34	87.79	89.13	849.2297	849.2265	849.2327	2.04E+06	5
VSIG8	rs62624468	Major	LGCPYVLDPEDYGPNGLDIEWMQVNSDPAHHR	C3:Carbamidomethylation; N15:Deamidation (NQ);Q23:Deamidation (NQ)	89.34	89.21	89.56	924.9190	924.9138	924.9230	5.03E+07	4
KRT32	rs2071563	Major	LEGEINTYR	L1:Acetylation (N-term)	67.76	67.76	67.76	568.7823	568.7823	568.7823	1.81E+06	1
VSIG8	rs62624468	Major	LGCPYVLDPEDYGPNGLDIEWMQVNSDPAHHRENVFLSYQDK	C3:Carbamidomethylation	89.71	88.87	90.55	984.4526	984.4505	984.4546	2.54E+06	2
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	C8:Carbamidomethylation;N29:Deamidation (NQ)	89.56	89.47	89.72	906.0603	906.0582	906.0617	1.99E+08	4
KRT83	rs2852464	Minor	SRDLNMDCMVAEIKAQYDDIATR	R2:Methylation(KR);N5:Deamidation (NQ)	91.65	91.62	91.69	891.7481	891.7481	891.7481	2.67E+06	2
KRT83	rs2852464	Major	DLNMDCIVAEIK	M4:Oxidation (M);C6:Propionamide	86.47	86.47	86.47	725.8500	725.8500	725.8500	3.18E+07	1
KRT81	rs2071588	Minor	GLTGGFGSHSVCR	C12:Carbamidomethylation	34.40	34.40	34.40	445.5472	445.5472	445.5472	2.25E+08	1
KRT33A	rs148752041	Minor	HNAELENLIR	N7:Deamidation (NQ)	55.09	54.82	55.35	403.8785	403.8784	403.8785	8.26E+06	2
KRT33A	rs148752041	Minor	HNAELENLIRER	N2:Deamidation (NQ)	66.49	63.44	69.69	498.9269	498.9263	498.9279	9.52E+06	4
KRT83	rs2852464	Major	MDCIVAEIKAQYDDIATR	C3:Carbamidomethylation	88.36	88.35	88.36	704.6745	704.6742	704.6748	1.05E+06	2
VSIG8	rs62624468	Major	LGCPYVLDPEDYGPNGLDIEWMQVNSDPAHHRENVFLSYQDK	C3:Carbamidomethylation; M22:Oxidation (M);Q23:Deamidation (NQ)	90.26	90.26	90.26	987.8550	987.8550	987.8550	3.54E+05	1
VSIG8	rs62624468	Major	LGCPYVLDPEDYGPNGLDIEWMQVNSDPAHHRENVFLSYQDKR	C3:Carbamidomethylation; N15:Deamidation (NQ);M22:Oxidation (M)	88.42	88.13	89.10	1019.0730	1019.0701	1019.0743	2.40E+06	4
KRTAP4-11	rs763737606	Minor	TTYCRPSYSVSCC	C4:Carbamidomethylation;C12:Carbamidomethylation ;C13:Carbamidomethylation	54.35	50.18	58.83	820.8288	820.8278	820.8297	1.32E+08	12

Table S-3.1 (cont'd)

Gene	dbSNP	Variant Type	Peptide	PTM	Avg Adj RT	RT Min	RT Max	Avg *m/z*	*m/z* Min	*m/z* Max	Avg Peak Area	Obs Freq
KRT32	rs2071563	Major	LEGEINTYR		38.69	30.91	42.37	547.7778	547.7762	547.7791	1.71E+08	28
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATR	K12:Methylation(KR); Q14:Deamidation (NQ);Y15:Phosphorylation (STY)	92.29	92.29	92.29	623.7826	623.7812	623.7840	9.82E+05	2
KRTAP4-11	rs774046661	Major	TTYCRPSCCVSS	C4:Carbamidomethylation;C8:Carbamidomethylation; C9:Carbamidomethylation	26.63	23.02	30.58	739.2960	739.2935	739.2977	1.75E+08	5
KRT81	rs2071588	Minor	GLTGGFGSHSVCRGFR	C12:Propionamide	56.81	52.83	60.79	427.9643	427.9635	427.9650	1.15E+07	2
KRT33A	rs148752041	Minor	HNAELENLIR	H1:Acetylation (N-term)	80.08	80.08	80.08	625.8270	625.8270	625.8270	5.37E+05	1
VSIG8	rs62624468	Minor	LGCPYILDPEDYGPNGLDIEWMQVNSDPAHHR	C3:Carbamidomethylation; N15:Deamidation (NQ);Q23:Deamidation (NQ)	90.04	89.65	90.44	928.4216	928.4215	928.4216	6.03E+06	2
KRT33A	rs148752041	Minor	VRQLERHNAELENLIRER		74.11	64.54	79.82	455.8527	455.8527	455.8528	3.10E+06	3
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLK	C8:Carboxymethyl	92.69	92.69	92.69	795.1663	795.1663	795.1663	4.35E+06	1
KRT83	rs2852464	Minor	LDNSRDLNMDCMVAEIKAQYDDIATR	N8:Deamidation (NQ);C11:Carbamidomethylation	90.05	90.05	90.05	765.3552	765.3552	765.3552	4.28E+06	1
VSIG8	rs62624468	Major	LGCPYVLDPEDYGPNGLDIEWMQVNSDPAHHRENVFLSYQDKR	C3:Carbamidomethylation; N15:Deamidation (NQ);M22:Oxidation (M)	88.35	87.77	89.01	849.3941	849.3884	849.3976	6.88E+06	14
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	Q15:Deamidation (NQ); Q28:Deamidation (NQ);N29:Deamidation (NQ)	89.82	89.82	89.82	746.0431	746.0431	746.0431	5.19E+06	1
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATR	K12:Acetylation (K);Q14:Deamidation (NQ)	92.19	92.01	92.38	814.0508	814.0505	814.0511	1.64E+08	2
VSIG8	rs62624468	Major	LGCPYVLDPEDYGPNGLDIEWMQVNSDPAHHR	C3:Carbamidomethylation; N15:Deamidation (NQ);M22:Oxidation (M)	88.66	88.28	89.36	928.6667	928.6594	928.6712	5.91E+06	11
KRT33B	17:g.41366553G>T	Minor	ILDEMTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	M5:Oxidation (M);M16:Oxidation (M)	88.75	88.75	88.75	904.4546	904.4546	904.4546	2.72E+07	1
KRTAP4-11	rs763737606	Major	TTYCRPSCCVSS	C4:Carbamidomethylation;C8:Carbamidomethylation; C9:Carbamidomethylation	26.63	23.02	30.58	739.2960	739.2935	739.2977	1.75E+08	5
VSIG8	rs62624468	Major	LGCPYVLDPEDYGPNGLDIEWMQVNSDPAHHRENVFLSYQDK	C3:Carbamidomethylation; N15:Deamidation (NQ);Q23:Deamidation (NQ)	89.35	88.80	89.58	984.8551	984.8522	984.8575	9.88E+06	5
KRT32	rs2071563	Minor	ARLEGEINMYR		56.14	50.57	59.94	451.2309	451.2304	451.2314	1.52E+07	6
KRT33A	rs148752041	Minor	QLERHNAELENLIRER		65.92	65.92	65.92	404.8192	404.8192	404.8192	6.00E+07	1
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATRSRA	D16:Ethylation	93.19	93.19	93.19	685.5915	685.5915	685.5915	3.09E+06	1
KRT83	rs2852464	Major	DLNMDCIVAEIK	C6:Beta-methylthiolation	92.14	92.14	92.14	705.3351	705.3351	705.3351	6.47E+06	1
KRT83	rs2852464	Minor	LDNSRDLNMDCMVAEIKAQYDDIATR	C11:Carboxyl modification with ethanolamine	89.96	89.96	89.96	1015.1418	1015.1418	1015.1418	3.48E+07	1
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	C8:Carbamidomethylation;K20:Dihydroxy	89.98	89.81	90.20	912.2610	912.2548	912.2681	2.25E+07	5
KRT83	rs2852464	Minor	LDNSRDLNMDCMVAEIK	N3:Deamidation (NQ); C11:Carbamidomethylation;M12:Oxidation (M)	88.47	88.47	88.47	680.9747	680.9747	680.9747	3.72E+07	1
KRT81	rs2071588	Minor	GISCYRGLTGGFGSHSVCRGFR	C4:Carbamidomethylation;C18:Carbamidomethylation	73.81	73.81	73.81	487.0358	487.0358	487.0358	5.62E+04	1
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	C8:Carbamidomethylation;Q15:Deamidation (NQ)	89.72	89.29	90.57	755.2209	755.2194	755.2228	4.59E+07	9
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATR	N3:Deamidation (NQ);M4:Sulphone	93.56	93.56	93.56	810.7171	810.7171	810.7171	8.17E+05	1
KRT81	rs2071588	Minor	GLTGGFGSHSVCR	C12:Dihydroxy	44.80	31.53	47.97	655.3009	655.2999	655.3025	4.59E+06	8
KRT83	rs2852464	Major	NSRDLNMDCIVAEIKAQYDDIATR	S2:Phosphorylation (STY);M7:Oxidation (M)	91.74	91.72	91.77	950.7714	950.7704	950.7723	1.06E+06	2
KRT81	rs2071588	Minor	GISCYRGLTGGFGSHSVCR	C4:Carbamidomethylation;C18:Carbamidomethylation	61.16	60.71	61.62	690.9911	690.9905	690.9917	4.09E+07	2
KRT81	rs2071588	Minor	GLTGGFGSHSVCRG	C12:Dipyrrolylmethanemethyl	81.88	81.88	81.88	584.9248	584.9248	584.9248	4.22E+06	1
KRT32	rs2071563	Major	ARLEGEINTYR	E4:Monoglutamyl	59.83	45.85	65.31	484.2436	484.2415	484.2483	1.92E+07	5
KRTAP4-11	rs774046661	Major	TTYCRPSC	C4:Carbamidomethylation;C8:Carbamidomethylation	13.67	11.88	15.96	522.7159	522.7155	522.7161	1.86E+08	4
VSIG8	rs62624468	Major	LGCPYVLDPEDYGPNGLDIEWMQVNSDPAHHRENVFLSYQDKR	C3:Carbamidomethylation; N15:Deamidation (NQ);Q23:Deamidation (NQ)	88.84	88.37	89.26	846.8948	846.8918	846.8975	4.13E+07	8
KRT81	rs2071588	Minor	GLTGGFGSHSVCR	C12:Ubiquitin	30.96	28.67	33.25	464.5541	464.5540	464.5542	4.53E+07	2
KRT33A	rs148752041	Minor	HNAELENLIR		52.70	37.81	58.54	604.8221	604.8216	604.8226	4.83E+08	6
KRT81	rs2071588	Minor	GLTGGFGSHSVCR	H9:Oxidation (HW);C12:Carbamidomethylation	52.70	37.65	60.57	675.8135	675.8130	675.8138	4.49E+05	3
KRT32	rs2071563	Major	ARLEGEINTYR	E4:Carboxylation (E)	77.05	77.05	77.05	455.8969	455.8969	455.8969	2.20E+03	1
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	C8:Carbamidomethylation;N29:Deamidation (NQ)	89.47	89.47	89.47	1132.3224	1132.3224	1132.3224	7.20E+07	1
KRT32	rs2071563	Minor	LEGEINMYR	M7:Oxidation (M)	39.68	37.46	42.73	570.7713	570.7701	570.7725	1.69E+07	7

Table S-3.1 (cont'd)

Gene	dbSNP	Variant Type	Peptide	PTM	Avg Adj RT	RT Min	RT Max	Avg *m/z*	*m/z* Min	*m/z* Max	Avg Peak Area	Obs Freq
KRT83	rs2852464	Minor	DLNMDCMVAEIKAQYDDIATR	N3:Deamidation (NQ);C6:Carbamidomethylation	90.65	90.65	90.65	1237.0690	1237.0690	1237.0690	5.68E+05	1
KRT81	rs2071588	Minor	GLTGGFGSHSVCRGFR	C12:Carbamidomethylation	56.92	52.42	61.42	424.4597	424.4593	424.4600	2.42E+06	2
KRT81	rs2071588	Minor	GLTGGFGSHSVCRGFR	C12:Propionamide	56.62	51.63	60.73	570.2814	570.2805	570.2819	2.89E+06	3
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	R9:Dimethylation(KR)	89.72	89.72	89.72	750.2233	750.2233	750.2233	4.59E+05	1
KRT83	rs2852464	Minor	SRDLNMDCMVAEIK	C8:Carbamidomethylation	83.85	83.85	83.85	561.2629	561.2629	561.2629	1.96E+06	1
KRT83	rs2852464	Major	LDNSRDLNMDCIVAEIKAQYDDIATR	C11:Dihydroxy	91.64	91.64	91.64	754.3608	754.3608	754.3608	1.45E+05	1
KRT83	rs2852464	Minor	DLNMDCMVAEIKAQYDDIATR	N3:Deamidation (NQ); C6:Carbamidomethylation;M7:Oxidation (M)	87.88	87.86	87.90	830.3741	830.3724	830.3758	2.11E+06	2
KRT83	rs2852464	Major	DLNMDCIVAEIK	C6:Dihydroxy	91.12	90.85	91.39	698.3283	698.3273	698.3292	1.83E+06	2
KRT83	rs2852464	Minor	DLNMDCMVAEIKAQYDDIATR	M4:Oxidation (M);M7:Oxidation (M)	92.40	92.40	92.40	816.3672	816.3672	816.3672	1.30E+05	1
KRT83	rs2852464	Major	LDNSRDLNMDCIVAEIKA	N3:Deamidation (NQ);N8:Deamidation (NQ)	87.44	87.44	87.44	674.6608	674.6608	674.6608	7.42E+06	1
KRT83	rs2852464	Major	DLNMDCIVAEIKA	A13:Amidation	91.11	90.78	91.44	717.3590	717.3583	717.3596	4.49E+07	2
KRT83	rs2852464	Major	LDNSRDLNMDCIVAEIKAQYDDIATR	N8:Deamidation (NQ);C11:Carbamidomethylation	89.39	88.67	90.11	760.8625	760.8550	760.8699	4.82E+06	2
KRT83	rs2852464	Major	DLNMDCIVAEIKAQ	Q14:Ethylation	87.99	87.72	88.27	530.9377	530.9376	530.9377	7.81E+06	2
VSIG8	rs62624468	Minor	LGCPYILDPEDYGPNGLDIEWMQVNSDPAHHRENVFLSYQDKR	C3:Carbamidomethylation;N15:Deamidation (NQ)	89.78	89.78	89.78	849.0630	849.0630	849.0630	3.24E+06	1
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATR	M4:Oxidation (M); C6:Carbamidomethylation;K12:Acetylation (K)	91.16	91.16	91.16	838.0533	838.0533	838.0533	1.91E+06	1
KRT33A	rs148752041	Minor	HNAELENLIR	N2:Ammonia-loss (N)	82.26	82.26	82.26	596.3096	596.3096	596.3096	1.33E+08	1
KRT83	rs2852464	Minor	LDNSRDLNMDCMVAEIKAQYDDIATR	N3:Deamidation (NQ); N8:Deamidation (NQ);C11:Carbamidomethylation	89.38	88.99	89.76	765.6023	765.6013	765.6033	2.47E+06	2
KRT81	rs2071588	Minor	GLTGGFGSHSVCR	S8:Phosphorylation (STY);C12:Carbamidomethylation	31.58	31.58	31.58	472.2028	472.2028	472.2028	3.98E+07	1
VSIG8	rs62624468	Major	VLDPEDYGPNGLDIEWMQVNSDPAHHRENVFLSYQDKR	N10:Deamidation (NQ)	86.73	86.73	86.73	748.3503	748.3503	748.3503	1.08E+06	1
VSIG8	rs62624468	Major	LGCPYVLDPEDYGPNGLDIEWMQVNSDPAHHRENVFLSYQDKR	C3:Carbamidomethylation; Q23:Deamidation (NQ);N25:Deamidation (NQ)	88.92	88.92	88.93	846.8951	846.8947	846.8954	2.15E+07	2
KRT32	rs2071563	Major	ARLEGEINTYR	A1:Formylation;N8:Deamidation (NQ)	35.16	35.16	35.16	450.8955	450.8955	450.8955	5.97E+05	1
KRT32	rs2071563	Major	ARLEGEINTYR	A1:Iminobiotinylation	47.90	47.90	47.90	516.2703	516.2703	516.2703	2.25E+05	1
KRT33A	rs148752041	Minor	LERHNAELENLIRER	R3:Hydroxyphenylglyoxal arginine	86.73	86.73	86.73	675.3470	675.3470	675.3470	4.24E+06	1
KRT33B	17:g.41366553G>T	Minor	ILDEMTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	C8:Carbamidomethylation;N29:Deamidation (NQ)	89.59	89.25	90.12	758.2092	758.2083	758.2106	1.08E+08	3
KRT32	rs2071563	Major	ARLEGEINTYR	N8:Ammonia-loss (N)	38.45	37.47	39.43	435.5620	435.5618	435.5621	8.96E+05	2
KRT33B	17:g.41366553G>T	Minor	ILDEMTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	C8:Carbamidomethylation; Q15:Deamidation (NQ);Q28:Deamidation (NQ)	90.18	90.18	90.18	1137.0701	1137.0701	1137.0701	7.21E+05	1
KRT83	rs2852464	Major	DNSRDLNMDCIVAEIK	C10:Propionamide	87.43	87.43	87.43	636.3057	636.3057	636.3057	2.67E+06	1
KRT81	rs2071588	Minor	GLTGGFGSHSVCR	C12:Dihydroxy	44.22	44.22	44.22	437.2033	437.2033	437.2033	5.29E+07	1
KRT33B	17:g.41366553G>T	Minor	ILDEMTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	C8:Carbamidomethylation;Q28:Deamidation (NQ)	90.14	90.14	90.14	909.6469	909.6469	909.6469	1.11E+07	1
KRT81	rs2071588	Minor	GLTGGFGSHSVCRG	R13:Arginine oxidation to glutamic semialdehyde	31.41	31.41	31.41	646.2949	646.2949	646.2949	1.81E+06	1
KRTAP4-11	rs774046661	Minor	TTYCRPSYSVSC	C4:Carbamidomethylation;C12:Carbamidomethylation	50.97	46.53	54.83	740.8135	740.8124	740.8143	8.10E+07	11
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATR	C6:Phosphorylation (HCDR)	92.71	92.71	92.71	620.0433	620.0433	620.0433	6.03E+05	1
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATR	M4:Oxidation (M);K12:Carbamylation	88.87	88.54	89.70	819.3858	819.3837	819.3876	6.98E+07	14
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	C8:Carbamidomethylation; Q15:Deamidation (NQ);Q28:Deamidation (NQ)	91.28	91.28	91.28	906.2604	906.2604	906.2604	2.02E+06	1
KRT83	rs2852464	Major	ILQSHISDTSVVVKLDNSRDLNMDCIVAEIK	N17:Deamidation (NQ);C25:Carbamidomethylation	88.51	88.51	88.51	703.5652	703.5652	703.5652	1.99E+06	1
KRT83	rs2852464	Minor	LDNSRDLNMDCMVAEIKAQYDDIATR	N3:Deamidation (NQ);C11:Carbamidomethylation	87.79	87.79	87.79	765.3504	765.3504	765.3504	6.60E+07	1
KRT32	rs2071563	Major	ARLEGEINTYR	A1:Formylation	38.40	30.78	44.37	450.5648	450.5635	450.5655	5.72E+07	27
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATR	K12:Acetylation (K);Q14:Deamidation (NQ)	92.98	92.98	92.98	1220.5709	1220.5709	1220.5709	6.74E+06	1
KRT83	rs2852464	Minor	DLNMDCMVAEIKAQYDDIATR	C6:Beta-methylthiolation	91.50	91.50	91.50	821.0342	821.0342	821.0342	5.82E-11	1
KRT83	rs2852464	Major	LDNSRDLNMDCIVAEIK	C11:Cysteine mercaptoethanol	85.84	85.84	85.84	675.6466	675.6466	675.6466	9.59E+05	1
KRT83	rs2852464	Minor	DLNMDCMVAEIKAQYDDIATR	D1:Carbamylation;M4:Oxidation (M)	90.75	90.75	90.75	825.3837	825.3837	825.3837	2.72E+06	1
KRT83	rs2852464	Minor	LDNSRDLNMDCMVAEIKAQYDDIATR	L1:Carbamylation	90.59	90.59	90.59	1015.1332	1015.1332	1015.1332	8.98E+07	1

Table S-3.1 (cont'd)

Gene	dbSNP	Variant Type	Peptide	PTM	Avg Adj RT	RT Min	RT Max	Avg *m/z*	*m/z* Min	*m/z* Max	Avg Peak Area	Obs Freq
KRT33B	17:g.41366553G>T	Major	LDELTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	K19:Lipoyl	89.28	89.21	89.39	758.0466	758.0438	758.0482	5.15E+07	3
KRT81	rs2071588	Minor	CCITAAPYRGISCYRGLTGGFGSHSVCRG	C1:Carbamidomethylation;C2:Carbamidomethylation; C13:Carbamidomethylation	83.75	83.53	84.21	633.4956	633.4940	633.5009	1.50E+07	7
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATR	M4:Oxidation (M);C6:Ethanolation (C)	91.87	91.87	91.87	1229.0892	1229.0892	1229.0892	1.10E+06	1
KRT83	rs2852464	Major	LDNSRDLNMDCIVAEIK	N8:Deamidation (NQ);C11:Carbamidomethylation	87.25	86.57	87.92	669.6584	669.6577	669.6590	1.18E+07	2
KRTAP10-3	rs233252	Minor	STYCVPIPSC	C4:Carbamidomethylation;C10:Carbamidomethylation	85.43	85.36	85.58	592.2600	592.2596	592.2604	2.97E+06	4
KRTAP4-11	rs760092771	Minor	TTYCRPSYSVSC	C4:Carbamidomethylation;C12:Carbamidomethylation	50.97	46.53	54.83	740.8135	740.8124	740.8143	8.10E+07	11
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLK	C8:Carbamidomethylation	91.50	91.50	91.50	1059.5532	1059.5532	1059.5532	1.97E+07	1
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATR	M4:Oxidation (M); K12:Acetylation (K);Q14:Deamidation (NQ)	88.74	88.74	88.74	819.3806	819.3806	819.3806	3.60E+06	1
KRT32	rs2071563	Major	LEGEINTYR	N6:Beta-methylthiolation (ND)	39.53	37.46	42.73	570.7715	570.7701	570.7723	2.36E+07	5
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLK	S10:Aminoethylcysteine	91.82	91.51	92.14	795.4188	795.4172	795.4203	3.12E+07	2
KRT81	rs2071588	Minor	CCITAAPYRGISCYRGLTGGFGSHSVCRG	C1:Carbamidomethylation;C2:Carbamidomethylation; C13:Carbamidomethylation	83.67	83.55	83.89	791.6165	791.6162	791.6168	2.43E+07	3
KRTAP4-11	rs774046661	Minor	TTYCRPSYSVSCC	C4:Carbamidomethylation;C12:Carbamidomethylation ;C13:Carbamidomethylation	54.35	50.18	58.83	820.8288	820.8278	820.8297	1.32E+08	12
VSIG8	rs62624468	Major	LGCPYVLDPEDYGPNGLDIEWMQVNSDPAHHRENVFLSYQDK	C3:Carbamidomethylation; N15:Deamidation (NQ);N25:Deamidation (NQ)	89.88	89.88	89.88	984.8471	984.8471	984.8471	6.16E+06	1
KRT83	rs2852464	Major	LDNSRDLNMDCIVAEIKAQYDDIATR	C11:Carbamidomethylation;K17:Guanidination	91.03	89.80	92.19	771.1298	771.1240	771.1332	3.02E+07	10
KRT83	rs2852464	Major	LDNSRDLNMDCIVAEIK	C11:Propionamide	87.54	86.82	88.26	1010.4938	1010.4920	1010.4955	4.46E+06	2
VSIG8	rs62624468	Major	LGCPYVLDPEDYGPNGLDIEWMQVNSDPAHHRENVFLSY	C3:Carbamidomethylation;M22:Oxidation (M)	90.15	90.15	90.15	1141.5269	1141.5269	1141.5269	1.65E+07	1
KRT33A	rs148752041	Minor	QLERHNAELENLIR		61.15	58.06	64.25	434.4869	434.4868	434.4869	1.03E+07	2
KRT83	rs2852464	Major	DLNMDCIVAEIK	N3:Deamidation (NQ);C6:Carbamidomethylation	88.22	87.05	88.92	711.3398	711.3365	711.3450	5.22E+07	3
KRT33A	rs148752041	Minor	HNAELENLIRER	N2:Ammonia-loss (N)	83.53	83.53	83.53	492.9218	492.9218	492.9218	6.49E+06	1
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	C8:Carbamidomethylation; M16:Oxidation (M);Q32:Deamidation (NQ)	91.13	91.13	91.13	909.2644	909.2644	909.2644	2.83E+07	1
KRT33A	rs148752041	Minor	QLERHNAELENLIR		58.41	32.24	66.41	578.9792	578.9781	578.9802	2.52E+07	7
KRT81	rs2071588	Minor	GISCYRGLTGGFGSHSVCR	C4:Propionamide;C18:Carbamidomethylation	68.19	68.19	68.19	521.9982	521.9982	521.9982	3.69E+06	1
KRT83	rs2852464	Major	LNMDCIVAEIKAQYDDIATR	C5:Carbamidomethylated Cys that undergoes beta-elimination and Michael addition of ethylamine	88.49	88.49	88.49	765.0682	765.0682	765.0682	5.56E+08	1
KRT33A	rs148752041	Minor	HNAELENLIR	N7:Ammonia-loss (N)	53.39	53.39	53.39	397.8780	397.8780	397.8780	8.61E+04	1
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATRSR	C6:Carbamidomethylation	91.25	91.25	91.25	675.0778	675.0778	675.0778	9.81E+05	1
KRT83	rs2852464	Major	DLNMDCIVAEIK	C6:Crotonaldehyde	88.74	88.74	88.74	717.3529	717.3529	717.3529	1.38E+07	1
KRT81	rs2071588	Minor	GLTGGFGSHSVCRGF	S10:O-Isopropylphosphorylation	85.12	85.01	85.30	802.3611	802.3590	802.3621	5.14E+06	3
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQN	C8:Carbamidomethylation;Q15:Deamidation (NQ)	92.18	92.18	92.18	855.6912	855.6912	855.6912	2.04E+06	1
KRT83	rs2852464	Major	DLNMDCIVAEIK	C6:Carbamidomethylation	88.53	88.41	88.66	474.2315	474.2314	474.2315	3.35E+07	2
KRT33B	17:g.41366553G>T	Minor	RILDEMTLCR	C9:Carbamidomethylation	71.12	66.12	75.46	436.2246	436.2241	436.2257	2.20E+07	7
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATR	M4:Oxidation (M); Q14:Deamidation (NQ);Y15:Phosphorylation (STY)	92.72	92.72	92.72	499.6269	499.6269	499.6269	4.87E+05	1
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	C8:Carbamidomethylation;M16:Oxidation (M)	88.86	88.09	90.13	757.7197	757.7175	757.7222	4.40E+06	6
KRT83	rs2852464	Minor	DLNMDCMVAEIK	C6:Carbamidomethylation;M7:Oxidation (M)	83.38	82.39	83.93	727.8183	727.8170	727.8203	2.43E+07	7
KRT81	rs2071588	Minor	GLTGGFGSHSVCRGFRA		56.21	50.27	60.80	427.9638	427.9631	427.9647	1.80E+07	12
KRT83	rs2852464	Major	DLNMDCIVAEIK	C6:Propionamide	89.46	88.07	90.53	717.8517	717.8509	717.8527	4.54E+08	6
KRT81	rs2071588	Minor	GLTGGFGSHSVCRG	C12:Carbamidomethylation	33.85	28.67	39.47	464.5542	464.5537	464.5548	1.61E+07	8
VSIG8	rs62624468	Major	LGCPYVLDPEDYGPNGLDIEWMQVNSDPAHHRENVFLSYQDKR	C3:Carbamidomethylation	88.45	88.45	88.45	1015.6737	1015.6737	1015.6737	1.35E+04	1

Table S-3.1 (cont'd)

Gene	dbSNP	Variant Type	Peptide	PTM	Avg Adj RT	RT Min	RT Max	Avg *m/z*	*m/z* Min	*m/z* Max	Avg Peak Area	Obs Freq
VSIG8	rs62624468	Major	LGCPYVLDPEDYGPNGLDIEWMQVNSDP AHHRENVFLSYQDK	C3:Carbamidomethylation	90.30	90.30	90.30	984.4457	984.4457	984.4457	1.16E+07	1
KRT32	rs2071563	Major	ARLEGEINTYR	A1:Sulfonation of N-terminus	85.14	85.14	85.14	729.3369	729.3369	729.3369	1.95E+05	1
VSIG8	rs62624468	Major	LGCPYVLDPEDYGPNGLDIEWMQVNSDP AHHRENVFLSYQDKR	C3:Carbamidomethylation; N15:Deamidation (NQ);N25:Deamidation (NQ)	88.79	88.58	89.21	1016.0700	1016.0619	1016.0756	1.83E+07	6
KRT81	rs2071588	Minor	GLTGGFGSHSVCR	S10:Phosphorylation (STY); C12:Carbamidomethylation	47.57	46.62	48.51	472.2038	472.2035	472.2041	1.80E+07	2
VSIG8	rs62624468	Minor	LGCPYILDPEDYGPNGLDIEWMQVNSDPA HHRENVFLSYQDKR	C3:Carbamidomethylation; M22:Oxidation (M);N25:Deamidation (NQ)	88.07	88.00	88.14	851.7306	851.7302	851.7310	2.75E+07	2
VSIG8	rs62624468	Minor	LGCPYILDPEDYGPNGLDIEWMQVNSDPA HHRENVFLSYQDKR	C3:Carbamidomethylation;M22:Oxidation (M)	88.05	88.00	88.10	851.5665	851.5662	851.5668	4.66E+06	2
KRT33B	17:g.41366 553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQN HEQEVNTLR	M16:Oxidation (M);K20:Methylation(KR)	90.39	90.39	90.39	750.5522	750.5522	750.5522	1.13E+08	1
KRT33B	17:g.41366 553G>T	Minor	ILDEMTLCRSDLEAQMESLKEELLSLKQN HEQEVNTLR	C8:Carbamidomethylation; Q15:Deamidation (NQ);N29:Deamidation (NQ)	89.68	89.21	90.14	909.8566	909.8529	909.8602	5.48E+07	2
KRT33B	17:g.41366 553G>T	Minor	ILDEMTLCRSDLEAQMESLKEELLSLKQN HEQEVNTLR	C8:Carbamidomethylation	89.61	89.08	90.13	909.4535	909.4521	909.4548	4.55E+06	2
VSIG8	rs62624468	Minor	LGCPYILDPEDYGPNGLDIEWMQVNSDPA HHRENVFLSYQDKR	C3:Carbamidomethylation; N15:Deamidation (NQ);M22:Oxidation (M)	88.27	87.98	88.56	1021.8776	1021.8740	1021.8812	2.81E+06	2
KRT33B	17:g.41366 553G>T	Major	LDELTLCRSDLEAQMESLKEELLSLKQNH EQEVNTLR	C7:Carbamidomethylation	90.46	90.46	90.46	883.2513	883.2513	883.2513	9.72E+05	1
VSIG8	rs62624468	Major	LGCPYVLDPEDYGPNGLDIEWMQVNSDP AHHR	C3:Carbamidomethylation;N25:Deamidation (NQ)	89.20	88.90	89.50	924.6630	924.6623	924.6636	6.51E+06	2
KRT33A	rs148752041	Minor	VRQLERHNAELENLIR		70.99	70.99	70.99	498.2786	498.2786	498.2786	1.64E+07	1
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATR	M4:Oxidation (M);Y15:Sulfation	92.64	92.34	92.91	624.0311	624.0269	624.0338	1.19E+06	3
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATR	N3:Deamidation (NQ); C6:Carbamidomethylation;Q14:Deamidation (NQ)	91.74	91.73	91.76	819.3889	819.3887	819.3890	5.59E+07	2
KRT83	rs2852464	Minor	DLNMDCMVAEIK	C6:Carbamidomethylation	87.25	86.99	87.52	480.2170	480.2169	480.2170	1.11E+07	2
KRT33A	rs148752041	Minor	QLERHNAELENLIRER	Q1:Pyro-glu from Q	70.66	70.66	70.66	501.5157	501.5157	501.5157	4.66E+05	1
KRT33A	rs148752041	Minor	QLERHNAELENLIRER	N6:Deamidation (NQ)	73.87	73.87	73.87	506.0182	506.0182	506.0182	1.33E+06	1
KRT83	rs2852464	Major	LDNSRDLNMDCIVAEIK	N8:Deamidation (NQ); M9:Oxidation (M);C11:Carbamidomethylation	83.63	83.63	83.63	674.9910	674.9910	674.9910	6.03E+05	1
KRT83	rs2852464	Major	ILQSHISDTSVVVKLDNSRDLNMDCIVAEI K	C25:Carbamidomethylation	87.86	87.86	87.86	703.3714	703.3714	703.3714	2.62E+06	1
KRT33B	17:g.41366 553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQN HEQEVNTLR	C8:Carbamidomethylation; Q15:Deamidation (NQ);M16:Oxidation (M)	89.97	89.97	89.97	1136.3358	1136.3358	1136.3358	3.04E+06	1
KRT83	rs2852464	Minor	NSRDLNMDCMVAEIKAQYDDIATR	N1:Deamidation (NQ);C9:Carbamidomethylation	92.28	92.28	92.28	944.1147	944.1147	944.1147	1.65E+05	1
VSIG8	rs62624468	Major	LGCPYVLDPEDYGPNGLDIEWMQVNSDP AHHRENVFLSYQDKR	C3:Carbamidomethylation;N25:Deamidation (NQ)	89.08	88.37	89.57	846.7260	846.7217	846.7304	1.90E+07	5
VSIG8	rs62624468	Major	VLDPEDYGPNGLDIEWMQVNSDPAHHR	N10:Deamidation (NQ)	87.86	87.70	88.02	777.1069	777.1061	777.1076	2.00E+06	2
KRT33A	rs148752041	Minor	QLERHNAELENLIR	Q1:Deamidation (NQ)	67.28	57.08	73.48	434.7322	434.7317	434.7326	9.92E+06	3
VSIG8	rs62624468	Major	LGCPYVLDPEDYGPNGLDIEWMQVNSDP AHHR	C3:Carbamidomethylation;N15:Deamidation (NQ)	89.56	89.56	89.56	1232.5564	1232.5564	1232.5564	1.54E+06	1
KRT81	rs2071588	Minor	GLTGGFGSHSVCR	G1:Acetylation (N-term);C12:Carbamidomethylation	67.73	67.73	67.73	688.8200	688.8200	688.8200	2.13E+07	1
KRT33B	17:g.41366 553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQN HEQEVNTLR	C8:Carbamidomethylation;Q32:Deamidation (NQ)	90.09	89.67	90.51	755.2218	755.2200	755.2236	1.63E+08	2
KRT33B	17:g.41366 553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQN HEQEVNTLR	R9:Methylation(KR);M16:Oxidation (M)	89.52	89.42	89.76	750.5519	750.5510	750.5523	2.83E+08	4
VSIG8	rs62624468	Major	LGCPYVLDPEDYGPNGLDIEWMQVNSDP AHHRENVFLSYQDKR	C3:Carbamidomethylation;N15:Deamidation (NQ)	88.83	88.67	89.21	1015.8739	1015.8690	1015.8765	5.90E+06	5
KRT83	rs2852464	Major	DNSRDLNMDCIVAEIKAQYDDIATR	C10:Benzyl isothiocyanate	92.27	92.27	92.27	755.3494	755.3494	755.3494	9.41E+05	1
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATR	C6:Carbamidomethylation;K12:Guanidination	93.83	93.51	94.15	1248.6061	1248.6041	1248.6080	1.70E+07	2
KRT33B	17:g.41366 553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQN HEQEVNTLR	C8:Propionamide;N29:Deamidation (NQ)	91.19	91.19	91.19	908.8614	908.8614	908.8614	3.55E+06	1

Table S-3.1 (cont'd)

Gene	dbSNP	Variant Type	Peptide	PTM	Avg Adj RT	RT Min	RT Max	Avg *m/z*	*m/z* Min	*m/z* Max	Avg Peak Area	Obs Freq
VSIG8	rs62624468	Major	LGCPYVLDPEDYGPNGLDIEWMQVNSDP AHHRENVFLSYQDKR	C3:Carbamidomethylation;N15:Deamidation (NQ)	89.02	88.37	89.81	846.7251	846.7179	846.7323	6.36E+06	9
KRT33B	17:g.41366 553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQN HEQEVNTLR	C8:Carbamidomethylation; Q15:Deamidation (NQ);M16:Oxidation (M)	90.08	89.94	90.28	909.2641	909.2607	909.2665	2.22E+07	4
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATR	C6:Dihydroxy	93.78	93.51	94.31	810.3833	810.3820	810.3864	2.58E+06	5
VSIG8	rs62624468	Major	LGCPYVLDPEDYGPNGLDIEWMQVNSDP AHHRENVFLSYQDKR	C3:Carbamidomethylation;M22:Oxidation (M)	88.58	87.96	88.80	1018.8771	1018.8738	1018.8798	4.05E+06	6
KRT81	rs2071588	Minor	YRGLTGGFGSHSVCR	C14:Carbamidomethylation	45.01	45.01	45.01	414.2027	414.2027	414.2027	2.82E+06	1
KRT33B	17:g.41366 553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLK	C8:Carbamidomethylation; Q15:Deamidation (NQ);M16:Oxidation (M)	91.32	91.19	91.50	1065.2201	1065.2181	1065.2223	1.47E+08	3
KRTAP4-11	rs760092771	Major	TTYCRPSCCVSS	C4:Carbamidomethylation;C8:Carbamidomethylation; C9:Carbamidomethylation	26.63	23.02	30.58	739.2960	739.2935	739.2977	1.75E+08	5
KRT83	rs2852464	Minor	LDNSRDLNMDCMVAEIK	R5:Methylation(KR);C11:Carbamidomethylation	88.29	88.29	88.29	1019.4721	1019.4721	1019.4721	9.23E+05	1
KRT32	rs2071563	Major	LEGEINTYR	E2:Carboxylation (E)	39.59	39.59	39.59	569.7719	569.7719	569.7719	3.55E+06	1
KRT83	rs2852464	Minor	LDNSRDLNMDCMVAEIKAQYDDIATR	C11:Carbamidomethylation	89.65	89.06	90.11	765.1066	765.1028	765.1086	9.99E+06	7
KRT81	rs2071588	Minor	GLTGGFGSHSVCR	C12:Carboxymethyl	44.62	44.62	44.62	445.8747	445.8747	445.8747	9.02E+06	1
KRT33B	17:g.41366 553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQN HEQEVNTLR	C8:Carbamidomethylation;K20:Acetylation (K)	90.42	90.42	90.42	914.2693	914.2693	914.2693	1.09E+07	1
KRT32	rs2071563	Major	ARLEGEINTYR	N8:Deamidation (NQ)	38.63	35.85	42.83	441.5613	441.5605	441.5630	7.07E+07	17
KRT83	rs2852464	Major	DLNMDCIVAEIK	C6:Propionamide	89.63	89.51	89.93	478.9037	478.9029	478.9044	1.30E+07	4
KRT33A	rs148752041	Minor	QLERHNAELENLIRER	Q1:Pyro-glu from Q;N6:Deamidation (NQ)	84.12	84.11	84.13	501.7615	501.7613	501.7617	1.27E+07	2
KRT33B	17:g.41366 553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQN HEQEVNTLR	Q15:Deamidation (NQ); Q28:Deamidation (NQ);N29:Deamidation (NQ)	90.68	90.68	90.68	1118.5682	1118.5682	1118.5682	6.87E+07	1
KRT83	rs2852464	Major	LDNSRDLNMDCIVAEIKA	N8:Deamidation (NQ);M9:Oxidation (M)	85.36	85.36	85.36	679.6591	679.6591	679.6591	8.30E+06	1
KRT83	rs2852464	Major	DLNMDCIVAEIK	N3:Deamidation (NQ);C6:Propionamide	89.58	89.58	89.58	718.3435	718.3429	718.3441	4.97E+08	2
KRT33B	17:g.41366 553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQN HEQEVNTLR	C8:S-nitrosylation	89.74	89.42	90.41	900.2622	900.2601	900.2635	4.85E+07	6
KRT83	rs2852464	Major	DNSRDLNMDCIVAEIKAQYDDIATR	D9:Dehydration	91.81	91.81	91.81	951.1085	951.1085	951.1085	8.30E+05	1
KRT33A	rs148752041	Minor	QLERHNAELENLIR	N6:Deamidation (NQ)	69.12	57.08	73.48	434.7325	434.7318	434.7330	6.48E+06	6
KRT81	rs2071588	Minor	GLTGGFGSHSVCRGFR	C12:S-nitrosylation	64.60	64.60	64.60	556.2657	556.2657	556.2657	6.48E+06	1
KRT81	rs2071588	Minor	CITAAPYRGISCYRGLTGGFGSHSVCR	C1:Carbamidomethylation;C12:Carbamidomethylation ;C26:Carbamidomethylation	83.30	83.30	83.30	601.4884	601.4884	601.4884	1.97E+06	1
KRT83	rs2852464	Major	RDLNMDCIVAEIKAQYDDIATR	C7:Cysteinylation	91.67	91.67	91.67	891.4130	891.4130	891.4130	1.09E+06	1
VSIG8	rs62624468	Major	LGCPYVLDPEDYGPNGLDIEWMQVNSDP AHHRENVFLSYQDKR	C3:Carbamidomethylation	88.60	88.31	89.26	846.5635	846.5611	846.5659	2.30E+06	8
KRT81	rs2071588	Minor	GFGSHSVCR	C8:Carbamidomethylation	36.68	36.68	36.68	503.7301	503.7301	503.7301	1.89E+05	1
KRT33B	17:g.41366 553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQN HEQEVNTLR	C8:Carbamidomethylation	89.76	89.42	90.53	905.8633	905.8601	905.8657	2.71E+07	6
KRT33B	17:g.41366 553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQN HEQEVNTLR	R9:Methylation(KR); K20:Acetylation (K);K27:Acetylation (K)	90.53	90.53	90.53	761.8906	761.8906	761.8906	5.81E+07	1
KRT33B	17:g.41366 553G>T	Minor	ILDEMTLCRSDLEAQMESLKEELLSLKQN HEQEVNTLR	C8:Carbamidomethylation;Q15:Deamidation (NQ)	89.06	89.06	89.06	758.2137	758.2137	758.2137	1.34E+07	1
KRT33B	17:g.41366 553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQN HEQEVNTLR	K27:Methylation(KR);H30:Oxidation (HW)	90.41	90.41	90.41	900.4594	900.4594	900.4594	1.60E+08	1
KRT81	rs2071588	Minor	GLTGGFGSHSVCRGFRA	C12:Selenyl	68.97	68.97	68.97	596.9228	596.9228	596.9228	1.59E+06	1
VSIG8	rs62624468	Major	LGCPYVLDPEDYGPNGLDIEWMQVNSDP AHHR	C3:Carbamidomethylation;M22:Oxidation (M)	89.64	89.33	89.94	928.4137	928.4049	928.4215	2.06E+06	3
KRT33B	17:g.41366 553G>T	Major	LDELTLCRSDLEAQMESLKEELLSLKQNH EQEVNTLR	L1:Levuglandinyl-lysine anhyropyrrole adduct	89.19	89.10	89.32	776.4075	776.4065	776.4081	3.26E+07	3
KRT81	rs2071588	Minor	TGGFGSHSVCRGFR	R11:Ornithine from Arginine	69.47	69.47	69.47	475.9035	475.9035	475.9035	1.53E+05	1
KRT33A	rs148752041	Minor	VRQLERHNAELENLIRER	N8:Deamidation (NQ)	70.79	63.96	77.61	456.0538	456.0533	456.0543	1.54E+06	2
KRT33B	17:g.41366 553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQN HEQEVNTLR	R9:Methylation(KR); K20:Acetylation (K);K27:Acetylation (K)	89.59	89.46	89.72	914.0671	914.0654	914.0687	1.70E+07	4
KRT33B	17:g.41366 553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQN HEQEVNTLR	C8:Carbamidomethylation; M16:Oxidation (M);K27:Acetylation (K)	89.12	89.12	89.12	917.4754	917.4754	917.4754	1.77E+06	1

Table S-3.1 (cont'd)

Gene	dbSNP	Variant Type	Peptide	PTM	Avg Adj RT	RT Min	RT Max	Avg *m/z*	*m/z* Min	*m/z* Max	Avg Peak Area	Obs Freq
KRT33B	17:g.41366553G>T	Minor	ILDEMTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	C8:Carbamidomethylation;Q15:Deamidation (NQ)	89.77	89.77	89.77	909.6522	909.6522	909.6522	2.84E+07	1
KRT33B	17:g.41366553G>T	Minor	ILDEMTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	C8:Carbamidomethylation;N29:Deamidation (NQ)	89.77	89.77	89.77	909.6516	909.6516	909.6516	2.82E+07	1
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATR	D17:Carboxylation (DKW)	92.03	92.03	92.03	814.3848	814.3848	814.3848	8.75E+08	1
KRT83	rs2852464	Major	DLNMDCIVAEIKA		89.16	87.77	90.53	717.8516	717.8511	717.8523	3.17E+08	8
KRT33B	17:g.41366553G>T	Major	RILDELTLCRSDLEAQMESLK	S11:Sulfation	90.05	90.05	90.05	848.4076	848.4076	848.4076	6.61E+06	1
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATR	N3:Deamidation (NQ);C6:Carbamidomethylation	91.85	91.85	91.85	614.5465	614.5465	614.5465	7.42E+05	1
KRT33A	rs148752041	Minor	HNAELENLIR	N2:Deamidation (NQ);N7:Deamidation (NQ)	73.08	73.08	73.08	404.2066	404.2066	404.2066	2.24E+05	1
KRT83	rs2852464	Major	LNMDCIVAEIKAQYDDIATR	C5:Carboxymethylated DTT modification of cysteine	92.06	91.84	92.29	623.7837	623.7834	623.7840	6.08E+05	2
KRT32	rs2071563	Minor	ARLEGEINMYR	M9:Sulphone	42.99	42.99	42.99	461.8943	461.8943	461.8943	1.83E+06	1
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATR	C6:Cysteine mercaptoethanol	91.14	91.14	91.14	825.0453	825.0453	825.0453	4.14E+07	1
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLK	C8:Carbamidomethylation;Q15:Deamidation (NQ)	91.49	91.49	91.49	1059.8884	1059.8884	1059.8884	1.92E+07	1
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	T6:EDT	89.59	89.06	90.12	758.2110	758.2083	758.2137	2.75E+07	2
VSIG8	rs62624468	Major	LGCPYVLDPEDYGPNGLDIEWMQVNSDPAHHR	C3:Carbamidomethylation; N15:Deamidation (NQ);N25:Deamidation (NQ)	89.70	89.34	90.30	924.9148	924.9114	924.9175	6.80E+07	7
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	C8:Carbamidomethylation; Q15:Deamidation (NQ);N29:Deamidation (NQ)	89.78	89.69	89.86	755.3906	755.3895	755.3917	2.67E+08	2
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQN	C8:Carbamidomethylation	91.71	91.71	91.71	855.4385	855.4385	855.4385	8.85E+05	1
KRT81	rs2071588	Minor	TGGFGSHSVCR	C10:Carbamidomethylation	9.33	9.33	9.33	388.8395	388.8395	388.8395	6.01E+06	1
KRT81	rs2071588	Minor	GLTGGFGSHSVCRGFR	C12:Carbamidomethylation	58.48	58.48	58.48	565.6094	565.6094	565.6094	1.13E+06	1
KRT32	rs2071563	Minor	ARLEGEINMYR	N8:Deamidation (NQ);M9:Oxidation (M)	44.58	44.58	44.58	456.8965	456.8965	456.8965	3.97E+07	1
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATR	C6:Carbamidomethylation;K12:Guanidination	93.22	92.99	93.84	832.7410	832.7403	832.7424	3.26E+08	5
KRT33B	17:g.41366553G>T	Minor	ILDEMTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	Q28:Deamidation (NQ)	90.45	90.45	90.45	898.2482	898.2482	898.2482	2.40E+07	1
KRT33A	rs148752041	Minor	QLERHNAELENLIRER	Q1:Deamidation (NQ)	69.73	69.73	69.73	674.3593	674.3593	674.3593	6.45E+06	1
KRT83	rs2852464	Major	LDNSRDLNMDCIVAEIKAQYDDIATR	N3:Deamidation (NQ); C11:Carbamidomethylation;K17:Acetylation (K)	91.54	91.24	91.85	771.3749	771.3741	771.3766	8.83E+07	3
KRT83	rs2852464	Major	LDNSRDLNMDCIVAEIKA		87.35	86.14	88.23	673.9992	673.9985	673.9997	2.75E+07	10
VSIG8	rs62624468	Major	LGCPYVLDPEDYGPNGLDIEWMQVNSDPAHHRENVFLSYQDK	C3:Carbamidomethylation;N15:Deamidation (NQ)	89.71	89.31	90.40	984.6536	984.6502	984.6599	2.21E+06	5
KRT83	rs2852464	Minor	DLNMDCMVAEIKAQYDDIATR	C6:Carbamidomethylation;M7:Oxidation (M)	87.93	87.67	88.66	830.0428	830.0399	830.0453	4.29E+06	16
KRT83	rs2852464	Minor	LDNSRDLNMDCMVAEIK	N8:Deamidation (NQ);C11:Carbamidomethylation	85.67	85.49	85.84	675.6451	675.6435	675.6466	6.70E+05	2
KRT83	rs2852464	Major	MDCIVAEIK	C3:Carbamidomethylation	69.99	68.45	72.03	539.7672	539.7666	539.7676	3.48E+06	3
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATR	N3:Deamidation (NQ);C6:Carbamidomethylation	91.98	90.81	92.88	819.0585	819.0544	819.0621	1.84E+08	5
KRT83	rs2852464	Minor	LDNSRDLNMDCMVAEIKAQYDDIATR	C11:Acetylation (TSCYH)	90.29	90.29	90.29	761.3518	761.3518	761.3518	1.11E+07	1
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATR	C6:Cysteine mercaptoethanol	90.91	90.65	91.16	1237.0683	1237.0676	1237.0690	3.14E+05	2
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATR	M4:Oxidation (M);C6:Dihydroxy	92.65	92.35	92.96	815.7172	815.7168	815.7175	1.97E+07	2
KRT81	rs2071588	Minor	GLTGGFGSHSVCRGFRA		44.85	31.53	58.16	570.2828	570.2825	570.2831	1.44E+07	2
KRT32	rs2071563	Minor	LEGEINMYR		59.17	59.00	59.33	562.7737	562.7729	562.7745	8.32E+06	2
KRT83	rs2852464	Major	LDNSRDLNMDCIVAEIKA	M9:Oxidation (M)	85.25	85.09	85.40	679.3301	679.3296	679.3306	7.87E+06	2
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATR	C6:Carbamidomethylation	91.89	91.89	91.89	614.2957	614.2957	614.2957	1.55E+06	1
KRT33B	17:g.41366553G>T	Minor	ILDEMTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	M5:Oxidation (M);M16:Oxidation (M)	89.81	89.81	89.81	904.4490	904.4490	904.4490	1.17E+05	1
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLK	C8:Carbamidomethylation;M16:Oxidation (M)	91.24	91.24	91.24	1064.8813	1064.8813	1064.8813	3.73E+07	1
VSIG8	rs62624468	Minor	ILDPEDYGPNGLDIEWMQVNSDPAHHRENVFLSYQDKR	M17:Oxidation (M)	86.95	86.95	86.95	753.1918	753.1918	753.1918	6.50E+04	1
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATRSRAEAESWYR	C6:Carbamidomethylation;K12:Guanidination	93.23	93.23	93.23	747.1592	747.1592	747.1592	3.28E+06	1
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	C8:Carbamidomethylation	89.51	89.49	89.53	1132.0786	1132.0763	1132.0808	1.35E+07	2

Table S-3.1 (cont'd)

Gene	dbSNP	Variant Type	Peptide	PTM	Avg Adj RT	RT Min	RT Max	Avg *m/z*	*m/z* Min	*m/z* Max	Avg Peak Area	Obs Freq
VSIG8	rs62624468	Major	VLDPEDYGPNGLDIEWMQVNSDPAHHRENVFLSYQDKR	N10:Deamidation (NQ);Q18:Deamidation (NQ)	88.11	88.11	88.11	748.5203	748.5203	748.5203	6.34E+05	1
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATR	D5:Sodium adduct;C6:Carbamidomethylation	92.25	92.25	92.25	826.0534	826.0534	826.0534	4.30E+04	1
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLK	C8:Monobromobimane derivative	92.18	92.18	92.18	828.1797	828.1797	828.1797	1.32E+05	1
KRT83	rs2852464	Major	LDNSRDLNMDCIVAEIKA	N8:Deamidation (NQ)	88.08	87.92	88.23	674.3289	674.3248	674.3329	6.13E+07	2
KRT81	rs2071588	Minor	TGGFGSHSVCR	C10:Carbamidomethylation	37.05	31.03	42.01	582.7636	582.7621	582.7648	1.18E+07	13
KRT32	rs2071563	Major	ARLEGEINTYR	N8:Deamidation (NQ)	30.79	30.79	30.79	661.8369	661.8369	661.8369	5.71E+07	1
KRT83	rs2852464	Major	SRDLNMDCIVAEIKAQYDDIATR	C8:Monobromobimane derivative	92.28	92.28	92.28	944.1147	944.1147	944.1147	1.65E+05	1
KRT83	rs2852464	Major	LDNSRDLNMDCIVAEIKAQYDDIATR	C11:Carbamidomethylation	90.32	89.90	91.00	760.6186	760.6171	760.6194	3.05E+07	13
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATR	N3:Deamidation (NQ);K12:Carbamylation	92.45	91.81	93.09	814.3836	814.3820	814.3852	3.74E+08	2
KRT83	rs2852464	Major	LDNSRDLNMDCIVAEIK	L1:Acetylation (N-term);C11:Carbamidomethylation	88.94	88.94	88.94	683.3307	683.3307	683.3307	4.09E+06	1
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	C8:Carbamidomethylation	89.80	89.35	90.62	755.0532	755.0504	755.0565	3.12E+07	12
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	Q15:Deamidation (NQ);M16:Oxidation (M);N29:Deamidation (NQ)	90.78	90.78	90.78	1122.3082	1122.3082	1122.3082	2.41E+07	1
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	R9:Methylation(KR);M16:Oxidation (M)	89.58	89.58	89.58	900.4594	900.4594	900.4594	3.36E+07	1
KRT33A	rs148752041	Minor	VRQLERHNAELENLIR		70.86	70.86	70.86	398.8237	398.8237	398.8237	1.77E+06	1
KRT33A	rs148752041	Minor	HNAELENLIRER		59.05	53.66	62.44	498.5978	498.5970	498.5995	1.11E+08	7
VSIG8	rs62624468	Major	LGCPYVLDPEDYGPNGLDIEWMQVNSDPAHHRENVFLSYQDKRINHGSLPHLQQR	C3:Carbamidomethylation;N15:Deamidation (NQ);M22:Oxidation (M)	86.82	86.82	86.82	809.8912	809.8912	809.8912	1.72E+05	1
KRT32	rs2071563	Major	ARLEGEINTYR	E6:Carboxylation (E)	73.15	63.03	79.10	455.8964	455.8955	455.8974	1.43E+07	20
KRT81	rs2071588	Minor	GLTGGFGSHSVCR	C12:Cysteine oxidation to cysteic acid	46.59	43.98	48.21	663.2982	663.2966	663.2994	1.25E+07	8
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATR	K12:Carbamylation	91.76	90.27	92.98	814.0542	814.0524	814.0557	1.83E+08	23
KRT81	rs2071588	Minor	GLTGGFGSHSVCR	C12:Ubiquitin	36.88	36.88	36.88	696.3270	696.3270	696.3270	6.11E+05	1
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	C8:Carbamidomethylation;Q15:Deamidation (NQ);M16:Oxidation (M)	89.38	87.96	90.59	757.8853	757.8773	757.8901	1.10E+07	8
KRT32	rs2071563	Major	GEINTYR		10.73	10.73	10.73	426.7142	426.7142	426.7142	8.49E+05	1
KRT83	rs2852464	Minor	DLNMDCMVAEIKAQYDDIATR	R21:Dimethylation(KR)	92.49	92.49	92.49	815.0471	815.0471	815.0471	5.30E+07	1
KRT32	rs2071563	Major	ARLEGEINTYR		38.99	35.98	43.04	441.2337	441.2329	441.2348	1.27E+08	17
VSIG8	rs62624468	Major	LGCPYVLDPEDYGPNGLDIEWMQVNSDPAHHR	C3:Carbamidomethylation;N15:Deamidation (NQ)	89.54	89.03	90.30	924.6693	924.6627	924.6744	1.36E+07	18
KRT81	rs2071588	Minor	GLTGGFGSHSVCRGFR		52.33	52.33	52.33	410.2085	410.2085	410.2085	4.60E+06	1
KRT83	rs2852464	Major	LDNSRDLNMDCIVAEIKAQYDDIATR	L1:Acetylation (N-term);C11:Carbamidomethylation;K17:Acetylation (K)	93.21	93.21	93.21	781.6294	781.6294	781.6294	2.13E+05	1
KRTAP4-11	rs763737606	Minor	TTYCRPSYSVS	C4:Carbamidomethylation	44.12	42.69	46.58	660.7981	660.7974	660.7986	9.05E+07	3
KRT33A	rs148752041	Minor	LERHNAELENLIR	R3:Hydroxyphenylglyoxal arginine	85.81	85.75	85.88	580.3077	580.3074	580.3079	6.66E+06	2
KRT33B	17:g.41366553G>T	Minor	ILDEMTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	C8:Carbamidomethylation	89.39	89.21	89.77	758.0473	758.0460	758.0482	3.38E+07	4
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	M16:Oxidation (M);K27:Methylation(KR)	90.00	89.58	90.41	750.5510	750.5507	750.5513	1.06E+08	3
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	R9:Dimethylation(KR);Q15:Deamidation (NQ)	89.44	89.44	89.44	750.3876	750.3876	750.3876	1.89E+08	1
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLK	C8:Carbamidomethylation	91.63	91.13	92.27	794.9151	794.9127	794.9166	1.49E+07	11
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATR	N3:Deamidation (NQ);C6:Carbamidomethylation	91.72	91.72	91.72	1228.0917	1228.0917	1228.0917	4.79E+06	1
KRT83	rs2852464	Major	DLNMDCIVAEIK	C6:Carboxymethyl	87.05	87.05	87.05	711.3450	711.3450	711.3450	2.89E+07	1
KRT33B	17:g.41366553G>T	Minor	ILDEMTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	C8:Carbamidomethylation	90.14	90.14	90.14	909.4405	909.4405	909.4405	3.70E+05	1
KRT83	rs2852464	Major	DLNMDCIVAEIKA	N3:Deamidation (NQ);M4:Oxidation (M)	86.51	86.46	86.56	726.3397	726.3396	726.3398	6.01E+07	2
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATR	M4:Oxidation (M);Y15:Phosphorylation (STY)	91.87	91.87	91.87	624.0325	624.0325	624.0325	6.12E+04	1
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATR	N3:Deamidation (NQ);C6:Carbamidomethylation;Q14:Deamidation (NQ)	92.03	91.80	92.26	1228.5837	1228.5804	1228.5869	3.62E+06	2
KRT32	rs2071563	Major	ARLEGEINTYR		38.13	31.21	42.83	661.3460	661.3452	661.3473	6.14E+07	20

Table S-3.1 (cont'd)

Gene	dbSNP	Variant Type	Peptide	PTM	Avg Adj RT	RT Min	RT Max	Avg *m/z*	*m/z* Min	*m/z* Max	Avg Peak Area	Obs Freq
KRT81	rs2071588	Minor	GISCYRGLTGGFGSHSVCR	C4:Carbamidomethylation;C18:Carbamidomethylation	64.67	64.00	65.28	518.4962	518.4957	518.4965	1.37E+07	4
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	C8:Carbamidomethylation;N35:Deamidation (NQ)	89.65	89.65	89.65	906.0541	906.0541	906.0541	1.07E+08	1
KRT83	rs2852464	Minor	LDNSRDLNMDCMVAEIK	N3:Deamidation (NQ);C11:Propionamide	87.61	87.61	87.61	680.3162	680.3162	680.3162	2.13E+07	1
KRT83	rs2852464	Major	DLNMDCIVAEIK	C6:Carbamidomethylation	88.63	86.86	89.42	710.8437	710.8414	710.8454	1.08E+08	32
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	Q15:Deamidation (NQ);Q28:Deamidation (NQ);N29:Deamidation (NQ)	89.69	89.23	90.14	895.0551	895.0536	895.0565	4.80E+06	2
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATR	C6:Carbamidomethylation	91.75	91.55	92.24	818.7256	818.7230	818.7285	2.22E+07	10
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLK	C8:Carbamidomethylation;M16:Oxidation (M)	89.43	89.38	89.53	798.9114	798.9089	798.9131	1.45E+06	3
KRT83	rs2852464	Major	LDNSRDLNMDCIVAEIK	N3:Deamidation (NQ);C11:Carbamidomethylation	86.77	86.72	86.82	669.6569	669.6567	669.6571	2.07E+07	2
VSIG8	rs62624468	Major	LGCPYVLDPEDYGPNGLDIEWMQVNSDPAHHRENVFLSYQDKR	C3:Carbamidomethylation;N15:Deamidation (NQ);N25:Deamidation (NQ)	88.97	88.70	89.26	846.8933	846.8923	846.8939	2.17E+07	3
KRT83	rs2852464	Major	LDNSRDLNMDCIVAEIKAQYDDIATR	M9:Oxidation (M);C11:Carbamidomethylation;K17:Acetylation (K)	89.86	89.86	89.86	775.1303	775.1303	775.1303	5.33E+06	1
KRT33A	rs148752041	Minor	QLERHNAELENLIR	Q1:Pyro-glu from Q;N6:Deamidation (NQ)	84.79	84.54	85.26	573.6322	573.6311	573.6330	1.36E+07	3
KRT83	rs2852464	Major	SRDLNMDCIVAEIKAQYDDIATR	C8:Carbamidomethylation	90.05	89.89	90.35	675.0794	675.0790	675.0800	1.57E+06	3
KRT33A	rs148752041	Minor	HNAELENLIR	E6:Replacement of 2 protons by calcium	33.01	33.01	33.01	416.1989	416.1989	416.1989	4.32E+05	1
KRTAP10-9	rs9980129	Minor	CAPTSSCQPSYCR	C1:Carbamidomethylation;C7:Carbamidomethylation;C12:Carbamidomethylation	24.58	23.18	27.20	787.3139	787.3128	787.3152	6.99E+06	10
VSIG8	rs62624468	Minor	LGCPYILDPEDYGPNGLDIEWMQVNSDPAHHRENVFLSYQDKR	C3:Carbamidomethylation;N15:Deamidation (NQ);Q23:Deamidation (NQ)	88.91	88.70	89.13	849.2322	849.2291	849.2352	4.69E+06	2
KRT33A	rs148752041	Minor	QLERHNAELENLIR	Q1:Acetylation (N-term);E3:Sodium adduct	84.69	84.44	84.94	600.3128	600.3124	600.3132	6.69E+06	2
KRT83	rs2852464	Minor	DLNMDCMVAEIK	N3:Deamidation (NQ);C6:Carbamidomethylation;M7:Oxidation (M)	83.75	83.75	83.75	728.3204	728.3204	728.3204	7.94E+05	1
KRT81	rs2071588	Minor	GLTGGFGSHSVCRGFR	S10:Methylphosphonylation	72.44	63.33	77.60	572.5922	572.5916	572.5926	4.19E+07	7
VSIG8	rs62624468	Major	LGCPYVLDPEDYGPNGLDIEWMQVNSDPAHHRENVFLSYQDKR	C3:Carbamidomethylation;M22:Oxidation (M);N25:Deamidation (NQ)	89.75	89.75	89.75	849.3928	849.3928	849.3928	5.38E+06	1
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATR	C6:Carbamidomethylation	92.07	91.69	92.87	1227.5860	1227.5770	1227.5988	3.08E+06	17
KRT32	rs2071563	Major	ARLEGEINTYR	E4:Carboxylation (E)	69.00	63.18	74.82	683.3399	683.3396	683.3401	1.29E+07	2
KRT83	rs2852464	Major	LDNSRDLNMDCIVAEIKAQYDDIATR	N3:Deamidation (NQ);C11:Carbamidomethylation	90.91	90.91	90.91	1014.1579	1014.1579	1014.1579	3.22E+06	1
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLK	C8:Carbamidomethylation;K20:Carbamylation	91.44	91.36	91.53	1073.8945	1073.8940	1073.8949	7.33E+06	2
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	C8:S-nitrosylation	89.47	89.32	89.62	750.3860	750.3848	750.3872	1.37E+07	2
KRT83	rs2852464	Major	LDNSRDLNMDCIVAEIK	C11:Carbamidomethylation	84.11	84.11	84.11	669.3250	669.3250	669.3250	2.06E+08	1
KRT83	rs2852464	Minor	LDNSRDLNMDCMVAEIK	C11:Carbamidomethylation	85.58	85.34	86.12	675.3124	675.3112	675.3140	3.83E+06	8
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATR	D1:O-Ethylphosphorylation	86.14	86.14	86.14	835.7075	835.7075	835.7075	5.12E+05	1
KRTAP4-11	rs763737606	Minor	TTYCRPSYSVSC	C4:Carbamidomethylation;C12:Carbamidomethylation	50.97	46.53	54.83	740.8135	740.8124	740.8143	8.10E+07	11
KRTAP4-11	rs774046661	Minor	TTYCRPSYSVS	C4:Carbamidomethylation	44.12	42.69	46.58	660.7981	660.7974	660.7986	9.05E+07	3
KRT81	rs2071588	Minor	GLTGGFGSHSVCR	S10:Phosphorylation (STY);C12:Carbamidomethylation	43.48	31.57	47.66	707.7996	707.7986	707.8005	2.96E+07	5
KRT83	rs2852464	Major	SRDLNMDCIVAEIK	C8:Carbamidomethylation;K14:Methylation(KR)	86.49	86.46	86.52	559.9475	559.9473	559.9477	7.42E+06	3
KRT83	rs2852464	Minor	DLNMDCMVAEIK	C6:Carbamidomethylation	87.49	87.32	88.05	719.8223	719.8206	719.8235	2.97E+07	19
KRT33A	rs148752041	Minor	QLERHNAELENLIRER		68.00	63.68	70.68	505.7722	505.7718	505.7726	4.41E+07	4
KRTAP4-11	rs760092771	Minor	TTYCRPSYSVSCC	C4:Carbamidomethylation;C12:Carbamidomethylation;C13:Carbamidomethylation	54.35	50.18	58.83	820.8288	820.8278	820.8297	1.32E+08	12
KRT33A	rs148752041	Minor	QLERHNAELENLIRER	N6:Deamidation (NQ)	75.90	75.90	75.90	405.0158	405.0158	405.0158	5.60E+06	1
KRT83	rs2852464	Minor	DLNMDCMVAEIKAQYDDIATR	N3:Deamidation (NQ);C6:Carbamidomethylation	91.15	91.14	91.16	825.0461	825.0453	825.0469	2.91E+07	2
KRT81	rs2071588	Minor	GLTGGFGSHSVCR	S10:Ubiquitin	38.82	38.82	38.82	464.5533	464.5533	464.5533	9.95E+06	1
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	Q15:Deamidation (NQ);M16:Oxidation (M);Q28:Deamidation (NQ)	90.85	90.85	90.85	1122.3102	1122.3102	1122.3102	2.15E+06	1

Table S-3.1 (cont'd)

Gene	dbSNP	Variant Type	Peptide	PTM	Avg Adj RT	RT Min	RT Max	Avg *m/z*	*m/z* Min	*m/z* Max	Avg Peak Area	Obs Freq
KRT33A	rs148752041	Minor	HNAELENLIR		56.80	52.86	59.35	403.5508	403.5503	403.5511	4.53E+07	6
KRT32	rs2071563	Major	ARLEGEINTYR	R2:Methylation(KR);T9:Sulfation	72.93	72.93	72.93	708.3322	708.3322	708.3322	2.90E+05	1
KRT32	rs2071563	Major	LEGEINTYR	N6:Deamidation (NQ)	40.91	40.74	41.08	548.2684	548.2681	548.2687	3.38E+07	2
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	C8:Carbamidomethylation;Q15:Deamidation (NQ)	89.71	89.71	89.71	1132.3273	1132.3273	1132.3273	2.24E+07	1
KRT83	rs2852464	Minor	DLNMDCMVAEIKAQYDDIATR	C6:Carbamidomethylation	90.88	90.54	91.78	824.7102	824.6974	824.7150	2.42E+07	15
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	C8:Carbamidomethylation;M16:Sulphone	89.97	89.94	89.99	912.2602	912.2601	912.2603	6.55E+06	2
KRT33B	17:g.41366553G>T	Minor	ILDEMTLCR	C8:Carbamidomethylation	79.09	75.33	83.15	575.7834	575.7826	575.7842	7.10E+07	10
KRT83	rs2852464	Minor	LDNSRDLNMDCMVAEIKAQYDDIATR	K17:Carbamylation	90.24	89.89	90.72	761.6013	761.6003	761.6023	5.34E+07	7
KRT83	rs2852464	Major	DLNMDCIVAEIK	C6:Carboxymethyl	88.58	88.58	88.58	474.5648	474.5648	474.5648	7.53E+05	1
KRT83	rs2852464	Minor	LDNSRDLNMDCMVAEIK	L1:Propionamide (K_X@N-term)	88.44	88.30	88.58	1019.4705	1019.4701	1019.4708	3.20E+07	2
KRT83	rs2852464	Major	SRDLNMDCIVAEIK	C8:Carbamidomethylation	86.18	86.10	86.26	555.2766	555.2760	555.2772	3.01E+06	2
KRT33A	rs148752041	Minor	QLERHNAELENLIRER	Q1:Deamidation (NQ);N6:Deamidation (NQ)	74.09	74.09	74.09	506.2686	506.2686	506.2686	3.18E+05	1
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	C8:Carbamidomethylation;M16:Oxidation (M)	90.12	87.98	91.03	909.0628	909.0578	909.0659	2.12E+06	9
KRT33A	rs148752041	Minor	HNAELENLIR	N2:Deamidation (NQ)	53.75	53.75	53.75	605.3137	605.3137	605.3137	9.15E+07	1
KRT83	rs2852464	Minor	LDNSRDLNMDCMVAEIK	C11:Propionamide	88.41	88.16	89.01	1019.4736	1019.4694	1019.4766	9.82E+06	4
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATR	M4:Oxidation (M);C6:Carbamidomethylation	89.73	89.46	90.47	824.0568	824.0552	824.0585	8.65E+06	9
KRT83	rs2852464	Major	DLNMDCIVAEIKAQYDDIATR	C6:Carbamidomethylation;E10:Sodium adduct	92.59	92.29	92.88	826.0546	826.0533	826.0558	9.39E+05	2
KRT81	rs2071588	Minor	GLTGGFGSHSVCRGFR	S8:Michael addition with methylamine	51.23	51.23	51.23	550.9444	550.9444	550.9444	2.43E+05	1
KRT83	rs2852464	Major	DLNMDCIVAEIK	M4:Oxidation (M);C6:Propionamide	86.89	86.50	87.21	725.8489	725.8483	725.8498	9.43E+07	5
VSIG8	rs62624468	Minor	LGCPYILDPEDYGPNGLDIEWMQVNSDPAHHR	C3:Carbamidomethylation;N15:Deamidation (NQ);M22:Oxidation (M)	89.20	88.74	89.76	932.1679	932.1654	932.1708	4.92E+06	5
KRT83	rs2852464	Major	DLNMDCIVAEIK	C6:Ubiquitin	88.11	87.91	88.32	739.3547	739.3539	739.3555	1.28E+07	2
KRT83	rs2852464	Major	LDNSRDLNMDCIVAEIKA	N3:Deamidation (NQ)	88.45	87.57	89.10	674.3270	674.3262	674.3281	1.13E+07	4
KRT83	rs2852464	Major	DLNMDCIVAEIK	M4:Oxidation (M);C6:Carbamidomethylation	85.46	85.18	85.95	718.8406	718.8389	718.8428	1.13E+07	15
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	C8:Carbamidomethylation;Q15:Deamidation (NQ);N29:Deamidation (NQ)	89.69	89.68	89.71	906.2596	906.2576	906.2615	5.75E+07	2
KRT83	rs2852464	Major	ILQSHISDTSVVVKLDNSRDLNMDCIVAEIK	N17:Deamidation (NQ);N22:Deamidation (NQ);C25:Carbamidomethylation	88.02	88.02	88.02	879.4531	879.4531	879.4531	1.54E+06	1
KRTAP10-3	rs233252	Minor	STYCVPIPS	C4:Carbamidomethylation	85.02	84.89	85.14	512.2450	512.2449	512.2450	1.65E+06	2
KRT83	rs2852464	Major	LDNSRDLNMDCIVAEIKAQYDDIATR	S4:EDT	89.97	89.97	89.97	765.3486	765.3486	765.3486	2.73E+06	1
VSIG8	rs62624468	Major	LGCPYVLDPEDYGPNGLDIEWMQVNSDPAHHR	C3:Carbamidomethylation	89.51	88.56	91.16	924.4190	924.4116	924.4226	1.11E+06	4
KRT33A	rs148752041	Minor	HNAELENLIR	N2:Deamidation (NQ)	69.82	66.72	72.84	403.8789	403.8781	403.8795	4.38E+07	8
KRT83	rs2852464	Major	DLNMDCIVAEIK	N3:Deamidation (NQ);C6:Carbamidomethylation	88.58	88.58	88.58	474.5648	474.5648	474.5648	7.53E+05	1
KRT33B	17:g.41366553G>T	Minor	ILDEMTLCR	M5:Oxidation (M);C8:Carbamidomethylation	53.35	51.65	55.06	583.7809	583.7805	583.7813	5.72E+06	2
KRT33B	17:g.41366553G>T	Major	ILDELTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	C8:Carbamidomethylation;Q15:Deamidation (NQ)	90.77	90.62	90.93	906.0625	906.0623	906.0627	1.42E+07	2
KRT33B	17:g.41366553G>T	Minor	ILDEMTLCRSDLEAQMESLKEELLSLKQNHEQEVNTLR	C8:Carbamidomethylation;Q15:Deamidation (NQ);N29:Deamidation (NQ)	89.39	89.39	89.39	758.3733	758.3733	758.3733	5.32E+07	1

Gene	SNP Identifier	1-H.1	1-H.2	1-H.3	1-H.4	1-A.1	1-A.2	1-A.3	1-A.4	1-P.1	1-P.2	1-P.3	1-P.4	2-H.1	2-H.2	2-H.3	2-H.4	2-A.1	2-A.2	2-A.3	2-A.4	2-P.1	2-P.2	2-P.3	2-P.4	3-H.1	3-H.2	3-H.3	3-H.4	3-A.1	3-A.2	3-A.3	3-A.4	3-P.1	3-P.2	3-P.3	3-P.4
KRT33A	rs148752041	1	1	1	1	1	1	1	1	1	1	1	1	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	0	--	--	--	--	--	--	--
VSIG8	rs62624468	0	0	0	0	0	0	0	0	0	0	0	0	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	1	0,1	0,1	0,1	0	0	0	0	0	0	0	0	0	0	0	0
KRT81	rs2071588	1	1	1	1	1	1	1	1	1	1	1	1	--	--	--	--	--	--	--	--	--	--	--	--	1	1	1	1	1	1	1	1	1	1	1	1
KRT83	rs2852464	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0	0	0	0	0	0	0	0	0	0	0	0
KRT32	rs2071563	0	0	0	0	0	0	0	0	0	0	0	0	0,1	0	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0	0	0	0	0	0	0	0	0	0	0	0
KRTAP10-9	rs9980129	--	1	--	--	--	--	--	--	--	1	1	--	1	1	1	--	--	--	1	--	--	1	1	--	--	1	1	--	--	--	1	--	--	1	1	--
KRTAP10-3	rs233252	--	--	--	--	--	--	--	--	--	--	--	--	--	--	1	--	--	--	--	1	--	1	1	--	1	--	1	--	1	--	--	--	--	--	1	--
FAM83H	rs9969600	--	--	--	--	--	--	--	--	--	--	--	--	1	--	--	--	--	--	--	--	1	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Figure S-3.4. GVP profiles established for each sample using the presence or non-detection of major and minor GVPs. “0” and “1” represent the presence of the major and minor GVP, respectively, while ‘--’ represents GVPs that were not detected.

REFERENCES

REFERENCES

1. Chu, F.; Mason, K. E.; Anex, D. S.; Jones, A. D.; Hart, B. R., Hair Proteome Variation at Different Body Locations on Genetically Variant Peptide Detection for Protein-Based Human Identification. *Scientific Reports* **2019,** *9* (1), 7641.

2. Wilhelm, M.; Schlegl, J.; Hahne, H.; Gholami, A. M.; Lieberenz, M.; Savitski, M. M.; Ziegler, E.; Butzmann, L.; Gessulat, S.; Marx, H.; Mathieson, T.; Lemeer, S.; Schnatbaum, K.; Reimer, U.; Wenschuh, H.; Mollenhauer, M.; Slotta-Huspenina, J.; Boese, J.-H.; Bantscheff, M.; Gerstmair, A.; Faerber, F.; Kuster, B., Mass-Spectrometry-Based Draft of the Human Proteome. *Nature* **2014,** *509*, 582-587.

3. Laatsch, C. N.; Durbin-Johnson, B. P.; Rocke, D. M.; Mukwana, S.; Newland, A. B.; Flagler, M. J.; Davis, M. G.; Eigenheer, R. A.; Phinney, B. S.; Rice, R. H., Human Hair Shaft Proteomic Profiling: Individual Differences, Site Specificity and Cuticle Analysis. *PeerJ* **2014,** *2*, e506.

4. Milan, J. A.; Wu, P.-W.; Salemi, M. R.; Durbin-Johnson, B. P.; Rocke, D. M.; Phinney, B. S.; Rice, R. H.; Parker, G. J., Comparison of Protein Expression Levels and Proteomically-Inferred Genotypes Using Human Hair from Different Body Sites. *Forensic Science International: Genetics* **2019,** *41*, 19-23.

5. Randall, V. A., Androgens and Hair Growth. *Dermatologic Therapy* **2008,** *21* (5), 314-328.

6. Miranda, B. H.; Charlesworth, M. R.; Tobin, D. J.; Sharpe, D. T.; Randall, V. A., Androgens Trigger Different Growth Responses in Genetically Identical Human Hair Follicles in Organ Culture that Reflect Their Epigenetic Diversity in Life. *The FASEB Journal* **2017,** *32* (2), 795-806.

7. Hwang, J.; Mehrani, T.; Millar, S. E.; Morasso, M. I., Dlx3 Is a Crucial Regulator of Hair Follicle Differentiation and Cycling. *Development* **2008,** *135* (18), 3149-3159.

8. Boutet, E.; Liberherr, D.; Tognolli, M.; Schneider, M.; Bairoch, A., UniProtKB/Swiss-Prot. *Methods in Molecular Biology* **2007,** *406*, 89-112.

9. Lek, M.; Karczewski, K. J.; Minikel, E. V.; Samocha, K. E.; Banks, E.; Fennell, T.; O'Donnell-Luria, A. H.; Ware, J. S.; Hill, A. J.; Cummings, B. B.; Tukiainen, T.; Birnbaum, D. P.; Kosmicki, J. A.; Duncan, L. E.; Estrada, K.; Zhao, F.; Zou, J.; Pierce-Hoffman, E.; Berghout, J.; Cooper, D. N.; Deflaux, N.; DePristo, M.; Do, R.; Flannick, J.; Fromer, M.; Gauthier, L.; Goldstein, J.; Gupta, N.; Howrigan, D.; Kiezun, A.; Kurki, M. I.; Moonshine, A. L.; Natarajan, P.; Orozco, L.; Peloso, G. M.; Poplin, R.; Rivas, M. A.; Ruano-Rubio, V.; Rose, S. A.; Ruderfer, D. M.; Shakir, K.; Stenson, P. D.; Stevens, C.; Thomas, B. P.; Tiao, G.; Tusie-Luna, M. T.; Weisburd, B.; Won, H.-H.; Yu, D.; Altshuler, D. M.; Ardissino, D.; Boehnke, M.; Danesh, J.; Donnelly, S.; Elosua, R.; Florez, J. C.; Gabriel, S. B.; Getz, G.;

Glatt, S. J.; Hultman, C. M.; Kathiresan, S.; Laakso, M.; McCarroll, S.; McCarthy, M. I.; McGovern, D.; McPherson, R.; Neale, B. M.; Palotie, A.; Purcell, S. M.; Saleheen, D.; Scharf, J. M.; Sklar, P.; Sullivan, P. F.; Tuomilehto, J.; Tsuang, M. T.; Watkins, H. C.; Wilson, J. G.; Daly, M. J.; MacArthur, D. G.; Exome Aggregation Consortium, Analysis of Protein-Coding Genetic Variation in 60,706 Humans. *Nature* **2016,** *536*, 285.

10. Parker, G. J.; Leppert, T.; Anex, D. S.; Hilmer, J. K.; Matsunami, N.; Baird, L.; Stevens, J.; Parsawar, K.; Durbin-Johnson, B. P.; Rocke, D. M.; Nelson, C.; Fairbanks, D. J.; Wilson, A. S.; Rice, R. H.; Woodward, S. R.; Bothner, B.; Hart, B. R.; Leppert, M., Demonstration of Protein-Based Human Identification Using the Hair Shaft Proteome. *PLoS ONE* **2016,** *11* (9), e0160653.

11. Adav, S. S.; Subbaiaih, R. S.; Kerk, S. K.; Lee, A. Y.; Lai, H. Y.; Ng, K. W.; Sze, S. K.; Schmidtchen, A., Studies on the Proteome of Human Hair - Identification of Histones and Deamidated Keratins. *Scientific Reports* **2018,** *8* (1), 1599.

12. Wang, W.; Zhou, H.; Lin, H.; Roy, S.; Shaler, T. A.; Hill, L. R.; Norton, S.; Kumar, P.; Anderle, M.; Becker, C. H., Quantification of Proteins and Metabolites by Mass Spectrometry Without Isotopic Labeling or Spiked Standards. *Analytical Chemistry* **2003,** *75* (18), 4818-4826.

13. Liu, H.; Sadygov, R. G.; Yates, J. R., A Model for Random Sampling and Estimation of Relative Protein Abundance in Shotgun Proteomics. *Analytical Chemistry* **2004,** *76* (14), 4193-4201.

14. Old, W. M.; Meyer-Arendt, K.; Aveline-Wolf, L.; Pierce, K. G.; Mendoza, A.; Sevinsky, J. R.; Resing, K. A.; Ahn, N. G., Comparison of Label-Free Methods for Quantifying Human Proteins by Shotgun Proteomics. *Molecular & Cellular Proteomics* **2005,** *4* (10), 1487-1502.

15. Chelius, D.; Bondarenko, P. V., Quantitative Profiling of Proteins in Complex Mixtures Using Liquid Chromatography and Mass Spectrometry. *Journal of Proteome Research* **2002,** *1* (4), 317-323.

16. Griffin, N. M.; Yu, J.; Long, F.; Oh, P.; Shore, S.; Li, Y.; Koziol, J. A.; Schnitzer, J. E., Label-Free, Normalized Quantification of Complex Mass Spectrometry Data for Proteomic Analysis. *Nature Biotechnology* **2009,** *28*, 83.

17. Zhang, Y.; Wen, Z.; Washburn, M. P.; Florens, L., Effect of Dynamic Exclusion Duration on Spectral Count Based Quantitative Proteomics. *Analytical Chemistry* **2009,** *81* (15), 6317-6326.

18. Jave-Suarez, L. F.; Langbein, L.; Winter, H.; Praetzel, S.; Rogers, M. A.; Schweizer, J., Androgen Regulation of the Human Hair Follicle: The Type I Hair Keratin hHa7 Is a Direct Target Gene in Trichocytes. *Journal of Investigative Dermatology* **2004,** *122* (3), 555-564.

19. Lee, Y. J.; Rice, R. H.; Lee, Y. M., Proteome Analysis of Human Hair Shaft: From Protein Identification to Posttranslational Modification. *Molecular & Cellular Proteomics* **2006,** *5* (5), 789-800.

20. Tagami, H.; Ray-Gallet, D.; Almouzni, G.; Nakatani, Y., Histone H3.1 and H3.3 Complexes Mediate Nucleosome Assembly Pathways Dependent or Independent of DNA Synthesis. *Cell* **2004,** *116* (1), 51-61.

21. Tabb, D. L.; Vega-Montoto, L.; Rudnick, P. A.; Variyath, A. M.; Ham, A.-J. L.; Bunk, D. M.; Kilpatrick, L. E.; Billheimer, D. D.; Blackman, R. K.; Cardasis, H. L.; Carr, S. A.; Clauser, K. R.; Jaffe, J. D.; Kowalski, K. A.; Neubert, T. A.; Regnier, F. E.; Schilling, B.; Tegeler, T. J.; Wang, M.; Wang, P.; Whiteaker, J. R.; Zimmerman, L. J.; Fisher, S. J.; Gibson, B. W.; Kinsinger, C. R.; Mesri, M.; Rodriguez, H.; Stein, S. E.; Tempst, P.; Paulovich, A. G.; Liebler, D. C.; Spiegelman, C., Repeatability and Reproducibility in Proteomic Identifications by Liquid Chromatography−Tandem Mass Spectrometry. *Journal of Proteome Research* **2010,** *9* (2), 761-776.

22. Bateman, N. W.; Goulding, S. P.; Shulman, N.; Gadok, A. K.; Szumlinski, K. K.; MacCoss, M. J.; Wu, C. C., Maximizing Peptide Identification Events in Proteomic Workflows Utilizing Data-Dependent Acquisition. *Molecular & Cellular Proteomics* **2013,** *13* (1), 329-338.

23. Solazzo, C.; Wadsley, M.; Dyer, J. M.; Clerens, S.; Collins, M. J.; Plowman, J., Characterisation of Novel α-Keratin Peptide Markers for Species Identification in Keratinous Tissues Using Mass Spectrometry. *Rapid Communications in Mass Spectrometry* **2013,** *27* (23), 2685-2698.

24. Zhang, Y.; Alsop, R. J.; Soomro, A.; Yang, F.-C.; Rheinstädter, M. C., Effect of Shampoo, Conditioner and Permanent Waving on the Molecular Structure of Human Hair. *PeerJ* **2015,** *3*, e1296.

25. Schrooyen, P. M. M.; Dijkstra, P. J.; Oberthür, R. C.; Bantjes, A.; Feijen, J., Stabilization of Solutions of Feather Keratins by Sodium Dodecyl Sulfate. *Journal of Colloid and Interface Science* **2001,** *240* (1), 30-39.

26. Langbein, L.; Rogers, M. A.; Winter, H.; Praetzel, S.; Beckhaus, U.; Rackwitz, H.-R.; Schweizer, J., The Catalog of Human Hair Keratins: I. Expression of the Nine Type I Members in the Hair Follicle. *Journal of Biological Chemistry* **1999,** *274* (28), 19874-19884.

27. Langbein, L.; Rogers, M. A.; Winter, H.; Praetzel, S.; Schweizer, J., The Catalog of Human Hair Keratins: II. Expression of the Six Type II Members in the Hair Follicle and the Combined Catalog of Human Type I and II Keratins. *Journal of Biological Chemistry* **2001,** *276* (37), 35123-35132.

28. Reich, D. E.; Cargill, M.; Bolk, S.; Ireland, J.; Sabeti, P. C.; Richter, D. J.; Lavery, T.; Kouyoumjian, R.; Farhadian, S. F.; Ward, R.; Lander, E. S., Linkage Disequilibrium in the Human Genome. *Nature* **2001,** *411*, 199.

29. Stephens, J. C.; Schneider, J. A.; Tanguay, D. A.; Choi, J.; Acharya, T.; Stanley, S. E.; Jiang, R.; Messer, C. J.; Chew, A.; Han, J.-H.; Duan, J.; Carr, J. L.; Lee, M. S.; Koshy, B.; Kumar, A. M.; Zhang, G.; Newell, W. R.; Windemuth, A.; Xu, C.; Kalbfleisch, T. S.; Shaner, S. L.; Arnold, K.; Schulz, V.; Drysdale, C. M.; Nandabalan, K.; Judson, R. S.; Ruaño, G.; Vovis, G. F., Haplotype Variation and Linkage Disequilibrium in 313 Human Genes. *Science* **2001,** *293* (5529), 489.

30. Mason, K. E.; Paul, P. H.; Chu, F.; Anex, D. S.; Hart, B. R., Development of a Protein-Based Human Identification Capability from a Single Hair. *Journal of Forensic Sciences* **2019,** *0* (0).

CHAPTER 4: Effects of Hair Age on Identification of Genetically Variant Peptides

Foreword

Contributions from others to the conduct of the experiments described in this chapter are as follows, in no particular order: T. M. Alfaro quantified mitochondrial DNA (mtDNA) by quantitative real-time polymerase chain reaction and sequenced the Hypervariable Region I in mtDNA, S. A. Malfatti developed protocols for quantifying and sequencing mtDNA and assembled mtDNA single nucleotide variant profiles, B. Rubinfeld and C. L. Strout optimized protocols for the fluorescent peptide assay, P. H. Paul provided nuclear DNA single nucleotide variant lists and individualized mutated protein FASTA files, and K. E. Mason and D. S. Anex acquired the mass spectrometry data.

4.1 Introduction

While the results presented in Chapter 3 demonstrated that variation in hair protein chemistry at different body locations did not affect GVP detection, intrinsic hair protein chemistry may change with other variables. For instance, degradation with increasing hair age, that is, time since biosynthesis of the hair fiber as opposed to the age of the individual, may compromise detection of genetically variant peptides (GVPs), which has yet to be examined. Once hair fibers are synthesized, with daily exposures to both internal and external stimuli, does genetic information in the form of GVPs remain unchanged? In forensic casework, hair specimens recovered at crime scenes are typically shed and not recently synthesized, thus having aged and been exposed to a variety of stimuli. This chapter examines whether the same GVPs are detected in aged hair fibers, quantified as time since their synthesis at the root, as compared to recently synthesized regions, and investigates additional strategies as a supplement to GVP analysis for improvement of discriminative potential.

Over time, with aging after their synthesis, hair fibers experience a variety of weathering processes, that is, external exposures, ranging from daily grooming practices to UV light exposure that contribute to their damage. Additionally, hair experiences natural aging, i.e., degradation of internal components over time from exposure to internal stimuli, such as enzymes and other biomolecules within the hair fiber. Accumulation of damage to hair proteins over time may affect GVP detection, but few studies have even assessed effects of hair age on physicochemical properties. Instead, hair age-related investigations have focused on comparing property differences between older and younger groups of individuals, of which their findings are briefly outlined below.

Knowledge gaps exist regarding physicochemical changes in hair fibers at the molecular level even though these variations may explain the observed macroscopic differences in hair structure with age. Hair age assessments have primarily examined differences in morphological and mechanical properties; in contrast, studies that have probed changes in hair fibers at the molecular level have been limited. Of clinical importance as indicators of human health and symptoms of underlying disorders, hair damage with age includes abrasion and depletion of cuticle layers, resulting in thinning of the hair shaft, graying, and reduction in tensile strength, elasticity, luster, and curl, as measured via light, electron, and atomic force microscopy or by tension tests[1-4]. Additionally, the hair cuticle has received the most attention with respect to examining effects of aging of an individual, including damage to the hair fiber as a result of aging, on physical properties of hair, as it experiences the bulk of weathering, excluding chemical hair treatments such as bleaching. This is owing to proximity of hair cuticle to external exposures, even though hair cortex, which lies beneath the cuticle, comprises the largest component (approximately 90%) of a hair fiber[5-7]. For the present work, variation in intrinsic

hair chemistry with age of the hair fiber since its synthesis, particularly in hair cortex, is most of interest. As hair cortex is the thickest component in hair fiber, the structure likely contains the main source of proteins from which variant peptides can then be identified, and therefore, knowledge of changes to hair physicochemical properties with deeper interrogation within hair shaft would permit more complete evaluation of GVP detection success rates with hair age.

Of note, a few studies have examined differences in intrinsic hair chemistry with age of the individual and not of the hair fiber itself, for indications of degradation. Giesen et al. compared older (> 50 years) to younger individuals (< 25 years), and in the former, observed lower mRNA levels of hair keratins and keratin-associated proteins (KAPs), the two protein classes that dominate hair proteins and localize to hair cortex, in hair follicles that surround the roots of hair fibers and contain DNA, RNA, and protein[8]. Imaging mass spectrometry of hair cortex identified lower abundances of small metabolites dihydrouracil and 3,4-dihydroxymandelic acid, proposed to bind to keratin intermediate filaments for hair structure rigidity, in older individuals[9]. Kuzuhara and co-workers used Raman spectroscopy to quantify disulfide bond crosslinks in hair cortex, and reported fewer crosslinks in virgin black hairs from older subjects compared to the younger group[10]. Together, these findings indicate degradation of internal components that provide stability and rigidity to hair fibers over the course of an individual's lifetime.

However, these studies focused primarily on physical age, i.e., the age of the subject, instead of hair growth age, that is, time since biosynthesis of the hair fiber. Observation of changes in hair physicochemical properties with physical age from 20 to 60 years of age facilitates diagnoses and treatments, but examination of hair fibers over such a long period of time poses little relevance for most forensic analyses. In the forensic context, the emphasis is on

determining the origin of evidence typically recovered from recent crime scenes, such as matching hair evidence to an individual, which involves substantially shorter timeframes, e.g., weeks to a few years. As such, assessment of variation in hair physicochemical properties with growth time, hereafter referred to as hair age, is an important consideration for development of GVP analysis as applicable to forensic investigations.

Notably, Thibaut and co-workers examined changes in hair protein content, among other physicochemical properties, with hair age, and reported substantial degradation of keratin-associated protein (KAP) content over 26 years of hair growth between root and tip (the distal end) for one individual[11]. As individuals rarely permit hair growth to this length, this study was limited to hair samples from one individual. Further specificity in protein degradation rates could not be achieved given the limitation of the analysis technique, 2D gel electrophoresis, and the bulk quantities of hair material used, i.e., 60 cm segments along the hair fiber[11], which correspond to approximately 4 years of hair growth, assuming a growth rate of 1.3 cm per month[12]. 2D gel electrophoresis only permits inference of protein class based upon molecular weight information and the extent of degradation was merely assessed by visual inspection. While this suggests that detection of GVPs in KAPs may be compromised with hair age, knowledge gaps still exist with respect to degradation rates of specific KAPs, as hair KAPs span a large protein class, consisting of 25 families[13] that may differ in degradation rates. Greater specificity in KAP degradation rates than that reported by Thibaut et al.[11] would be useful in guiding inclusion of specific KAP GVPs for human identification profiling. Furthermore, the analysis presented also represents an extreme case of hair growth, as it is unlikely to encounter such a large hair age difference in most forensic analyses. Instead, quantification of peptides from single one-inch (2.54 cm) hair segments using bottom-up proteomics and high-resolution

mass spectrometry will enable elucidation of protein-specific degradation rates with hair age across a time span more relevant for forensic identification.

A second study, conducted by Brandhagen et al., assessed changes in genomic DNA content with increasing hair age in 5-cm segments from single hairs in recently collected samples (i.e., within 4 years) and those collected at least 30 years ago. The investigation found that nuclear DNA, not mtDNA, comprised the majority of detected genomic information at all segments regardless of collection time and hair age; the nuclear DNA had severely degraded to tiny fragments and was therefore of insufficient quality for short tandem repeat profiling, the gold standard of DNA profiling[14]. The tiny nuclear DNA fragments in hair were analyzed via next-generation sequencing, a high-throughput DNA sequencing technique that is suitable for reading short, degraded sequences. Despite comprising 99% of genomic DNA sequence reads, the average nuclear DNA length was 39 bp at the distal end. Successful STR profiling relies on sequencing of longer DNA fragments from more intact nuclear DNA, with sequences sizes spanning at least 60 – 150 bp, which may not be consistently recovered from hair[14]. It has long been believed that minimal amounts of nuclear DNA remain in shed hairs, resulting in failure of STR profiling. However, DNA quality, that is, the extent to which DNA sequences remain intact, usually assessed by the sequence lengths of recovered fragments and not DNA quantity, is the key determining factor in profiling success. Assembling DNA profiles, i.e., SNP profiles, from degraded nuclear DNA in hair fibers requires more sensitive instruments and techniques than those used for STR profiling, such as next-generation sequencing; currently, few forensic laboratories are equipped to perform this analysis. In contrast to nuclear DNA, longer mtDNA fragments were recovered at the distal end of the hair (on average, 91 bp) even though detection of mtDNA sequences made up the minor component (1%) of recoverable DNA in hair; the

longer sequences enabled mtDNA profiling from shed hairs despite a degradation rate of at least 4-fold over approximately 1.6 years of hair growth[14]. Thus, mtDNA may be valuable as a supplement to GVP analyses of hair evidence in forensic investigations for increased specificity of human identification.

Alone, mitochondrial DNA offers limited discriminative power for human identification as its inheritance is exclusively maternal and the mitochondrial genome is much smaller (16,569 bp) for comparing genetic variation. However, this genetic information may still be recoverable in hair[15] for profiling concurrent with GVPs[16] where minimal intact nuclear DNA remains for successful STR profiling. The previous chapter briefly outlined the benefits of incorporating a co-extraction process for proteins and DNA, hereafter referred to as peptide/DNA co-fractionation, with emphasis on identifying SNPs in mtDNA concomitant with GVPs in hair proteins for increased specificity and discriminative power. Initial experiments in Chapter 3, Section 3.3.1 of this dissertation, also presented in Chu et al.[17], demonstrated proof-of-concept and comparable extraction of protein and peptide information to that from hair specimens prepared only for proteomics analysis. In addition to changes in protein chemistry with hair age, the co-extraction process is more systematically examined here by comparing protein and peptide composition in protein digests before and after co-fractionation to determine whether there are adverse effects on GVP identification. These efforts intend to define the extent of improvement in combined discriminative power with inclusion of co-fractionation in single hair sample preparation, and whether SNPs in mtDNA remain invariant in aged hairs.

This chapter aims to first, determine hair proteome variation with increasing hair age in one-inch segments, which provides greater time resolution of protein degradation (2 months) than previously examined, second, assess whether peptide/DNA co-fractionation affects peptide

and GVP detection rates, and third, determine the differentiative potential using GVPs from proteins and SNPs from mtDNA in aged hairs.

4.2 Experimental

4.2.1 Hair Sample Collection and Preparation

Scalp hairs were collected from three individuals under approval from the Institutional Review Board at Lawrence Livermore National Laboratory (Protocol ID# 15-008). Written informed consent for collection and analysis was obtained prior to collection. Specimens were then stored in the dark at room temperature (RT). The length of each single hair fiber was first measured prior to segmenting one-inch samples (Table 4.1). To examine variation in hair protein chemistry over hair growth time, two one-inch hair segments (~2.5 cm) each from the root and distal ends of hair fibers were segmented. These segments are designated root end (R), proximal-to-root (PR), proximal-to-distal (PD), and distal end (D). PR and PD segments represent one-inch samples proximal to the one-inch root and distal end segments, respectively (Figure 4.1). These sites were selected for comparison as the hair proteome may change over hair growth time, and if so, the four sites at the extremes are expected to exhibit the greatest differences in hair protein chemistry. Each one-inch hair segment was further segmented into 4 fragments of approximately 6 mm in length for full immersion of the hair specimen in denaturation solution during protein extraction. Four sets of biological replicates, i.e., 4 different hairs from each of 3 individuals, were sampled in total (n = 48 one-inch hair segments).

Table 4.1. Scalp hair lengths (in inches) for each set of biological replicates.

Individual	Replicate 1	Replicate 2	Replicate 3	Replicate 4
1	7.00	7.75	7.25	6.00
2	13.50	13.50	8.75	13.63
3	5.75	5.75	4.88	5.81

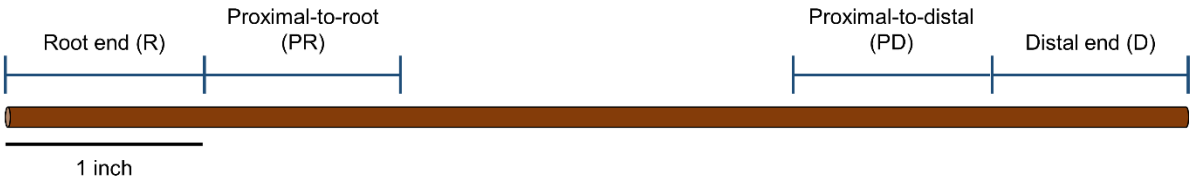


Figure 4.1. One-inch hair segment sampling to examine changes to the hair proteome with increasing hair age, using a 6-in hair fiber as an example.

Proteins were extracted from single one-inch hair segments and alkylated as described in Chu et al.[17] Following protein extraction and alkylation, an overnight protein precipitation with cold acetone was performed after solubilization to remove detergent sodium dodecanoate and concentrate proteins. A 4:1 ratio of cold acetone to aqueous protein extract was allowed to incubate overnight at -20 °C. After centrifugation at 15,000 × *g* for 15 min at RT, supernatants containing detergent were separated from the protein pellets and discarded. Pellets were washed with 400 µL of cold acetone and supernatant was once again discarded after centrifugation. Prior to protein digestion, pellets were resuspended in 50 µL of buffer solution and incubated on a shaker for 1 h at RT. Digestion using 2 µL of 1 µg/µL TPCK-treated trypsin was performed, with overnight incubation at RT accompanied by magnetic stirring. A 10-µL aliquot from each protein digest was then filtered to remove particulates using centrifugal filter tubes (PVDF, 0.1 µm; MilliporeSigma, Burlington, MA) prior to mass spectrometry analysis while peptide/DNA co-fractionation was performed with the remaining volume, which is described in Section 4.2.2.

Buffer solution concentration and composition for reconstitution of the protein pellet slightly differed among the 4 sets of biological replicates to ensure solution compatibility with a fluorescent peptide assay for assessment of total protein concentration in 2 sets of biological replicates. The first two sets of biological replicates were reconstituted to 50 µL of 25 mM ammonium bicarbonate and 0.01% (w/v) ProteaseMAX™ (Promega, Madison, WI) whereas the latter two sets used a buffer containing 50 mM dithiothreitol, 50 mM ammonium bicarbonate,

and 0.01% (w/v) ProteaseMAX™ (Promega, Madison, WI). The weaker concentration was used for solution compatibility with fluorescent peptide assays. More details on the fluorescent peptide assay are elaborated in Appendix Methods. Briefly, 2 2-µL aliquots were obtained from the unfiltered protein digest (40 µL volume) before peptide/DNA co-fractionation for the assay. Based on results of the fluorescent peptide assays, the 10-µL aliquots of protein digest were diluted to concentrations of 0.15 µg/µL and 0.10 µg/µL, respectively, for the first two sets of biological replicates, which represent the lowest concentration values common to each set of replicates. These aliquots were then analyzed via liquid chromatography-tandem mass spectrometry; for the last set of biological replicates, 10-µL aliquots of protein digest were not obtained for analysis.

4.2.2 Peptide/DNA Co-Fractionation

To separate peptides from DNA, aliquots from protein digests were processed using the QIAamp® DNA Micro Kit (Qiagen, Germantown, MD), with a few modifications to manufacturer's instructions. Briefly, 200 µL of 100% ethanol was added to each protein digest and incubated for 5 min at RT after mixing. Digests were then transferred to QIAamp MinElute columns and centrifuged at 6,000 × g for 1 min at RT. Filtrate was collected as the peptide fraction, which was then evaporated to dryness and reconstituted to initial volume with MilliQ water. Reconstituted peptide fractions were then mixed on a shaker for 1 h and filtered using centrifugal filter tubes as above. DNA retained on the MinElute columns was washed with 500 µL each of AW1 and AW2 buffers that were provided in the kit, with centrifugation at 6,000 × g for 1 min at RT following each wash step. Column membranes were then dried via centrifugation at 16,873 × g for 3 min at RT. Each DNA fraction was eluted and collected in a DNA LoBind microcentrifuge tube by addition of 20 µL of AE buffer (10 mM Tris·Cl, 0.5 mM EDTA at pH

9.0) to MinElute columns, incubation for 5 min at RT, and centrifugation at 16,873 × *g* for 1 min at RT; these steps were repeated once for a total volume of 40 µL for the genomic DNA fraction. To match the dilutions performed on protein digest aliquots based on fluorescent peptide assay results, peptide fractions for the first two sets of biological replicates were diluted as above.

4.2.3 Liquid Chromatography-Tandem Mass Spectrometry Analysis

Filtered protein digests, which were aliquoted before peptide/DNA co-fractionation, and peptide fractions, collected and filtered after co-fractionation, were analyzed on an EASY-nLC 1200 system coupled to a Q Exactive Plus Orbitrap mass spectrometer (Thermo Fisher Scientific Inc., Waltham, MA). Samples were injected onto an Acclaim™ PepMap™ 100 C18 trap (75 µm × 20 mm, 3 µm particle size), washed with 6 µL of mobile phase A, and separated on an Easy-Spray™ C18 analytical column (50 µm × 150 mm, 2 µm particle size). For the first two sets of biological replicates, 4- and 6-µL injections were used, respectively, whereas 0.5-µL injections were used for the remaining two sets of replicates; larger injection volumes were used to compensate for the dilutions performed during hair sample preparation. Separations were performed at a flow rate of 300 nL/min using mobile phases A (0.1% formic acid in water) and B (0.1% formic acid in 90% acetonitrile/10% water) over a 107-min gradient: 2 to 3% B in 1 min, 3 to 11% B in 75 min, 11 to 39% B in 15 min, ramped to 100% B in 1 min, and held at 100% B for 15 min. Positive mode nano-electrospray ionization was achieved at a capillary voltage of 1.9 kV. Full MS survey scans were acquired at a resolution of 70,000 over a scan range between *m/z* 380 and 1800, with a maximum ion accumulation time of 30 ms. Data-dependent MS/MS scans were triggered for the 10 most abundant survey scan ions at an intensity threshold of 3.3×10^4 and acquired at a resolution of 17,500, with a maximum ion accumulation time of 60 ms, dynamic exclusion of 24 s, and an isolation window of 2 Da. HCD fragmentation was performed

at a normalized collision energy setting of 27. Singly-charged species and ions with unassigned charge states were excluded from selection for MS/MS scans.

4.2.4 Protein and Peptide Identification

Protein and peptide identifications from mass spectral data were performed using PEAKS Studio 7.5 (Bioinformatics Solutions, Ontario, Canada). Details of the process are elaborated elsewhere[17]. Briefly, *de novo* sequencing was performed to identify peptides, with a precursor mass error tolerance of 20 ppm and fragment ion tolerance of 0.05 Da. Three tryptic missed cleavages were permitted and a total of 3 non-tryptic cleavages were allowed on both ends of peptides. Cysteine carbamidomethylation was selected as a fixed modification while all other post-translational modifications, including asparagine and glutamine deamidation and methionine oxidation, were allowed as variable modifications, with a maximum of three modifications per peptide. Peptides with confidence scores above 15% were then matched to protein sequences from the UniProtKB SwissProt Human database (downloaded September 21, 2017) and from individualized mutated databases that contain amino acid polymorphisms. A 1% false discovery rate was applied to filter peptide-spectrum matches, and only peptides unique to one gene were retained. Non-redundant peptide lists were then exported and further filtered with a 5-ppm mass error tolerance window for genetically variant peptide identification. GVP profiles were assembled as described in Chapter 2, Sections 2.3.6 and 2.3.7, and Chapter 3, Section 3.2.5 of this dissertation.

4.2.5 mtDNA Quantitation and SNP Profiles

Quantification of mtDNA abundance from DNA fractions was accomplished using quantitative real-time polymerase chain reaction (qPCR) and the NovaQUANT™ Human Mitochondrial to Nuclear DNA Ratio Kit (MilliporeSigma, Burlington, MA), with slight

modifications to manufacturer's instructions after obtaining cycle threshold (Ct) values (described below). Additionally, standard curves of Ct value over a range of known genomic DNA amounts (i.e., 0.01, 0.1, and 1 ng) for the human DNA positive control sample provided in the kit were generated for each of the four genes, BECN1, NEB, ND1, and ND6. The first two genes are from nuclear DNA while the latter two derive from mtDNA. These curves were used to quantify mtDNA as an alternative to the metric mitochondrial DNA copy number.

Because nuclear DNA was not recovered in 77% of DNA fractions from one-inch hair segments, mitochondrial DNA could not be quantified uniformly among all hair segments in the dataset using the relative copy number method described in the kit. Instead, mtDNA abundance in each hair segment was determined as a fold difference for a mitochondrial gene relative to genomic DNA in the human DNA positive control sample using the standard curves described above. Appendix Figure S-4.1 depicts the process by which relative fold difference, a proxy for mtDNA abundance, was calculated. The mitochondrial gene ND1 was chosen over ND6 due to its lower variability in qPCR efficiency (Figure S-4.1b). The equations in Figure S-4.1c enable use of ND1 relative fold difference as a proxy for mtDNA abundance.

SNP profiles were assembled from Sanger sequencing of Hypervariable Region I, a mitochondrial control region. To amplify this control region from genomic DNA fractions of one-inch hair segments, amplification via PCR was performed with the primers listed in Table 4.2, in 50 μL reaction volumes with Q5® Hot Start High-Fidelity 2X Master Mix (New England Biolabs, Ipswich, MA), which contains 2.5 μL genomic DNA and 0.2 μM each of forward and reverse primers. The amplification process was carried out on a PTC-200 DNA Engine (MJ Research, Waltham, MA) under the following conditions: 98 °C for 2 min; 15 cycles consisting of 98 °C for 10 s, 58 °C for 30 s, and 72 °C for 30 s; 25 cycles comprising 98 °C for 20 s, 58 °C

for 30 s, 72 °C for 30 s + 10 s/cycle; and a final extension at 72 °C for 2 min. Amplicons were then purified on a 2.0% agarose gel using QIAquick Gel Extraction Kit (Qiagen, Germantown, MD) according to the manufacturer's instructions, with the exception that DNA was eluted with 35 μL EB buffer (10 mM Tris·Cl, pH 8.5). Purified PCR amplicons were visualized following gel electrophoresis on a 2.0% agarose gel and then quantified using a QuBit 2.0 Fluorometer (Thermo Fisher Scientific Inc., Waltham, MA). A Big Dye Terminator v3.1 Cycle Sequencing Kit (Thermo Fisher Scientific Inc., Waltham, MA) and the primers in Table 4.2 were used for DNA amplification using PCR, with the following cycling conditions: 96 °C for 1 min; 30 cycles of 96 °C for 10 s, 50 °C for 5 s, and 60 °C for 2 min. Sequencing reactions were analyzed on an ABI 3500 Genetic Analyzer (Applied Biosystems, Foster City, CA) and *de novo*-assembled using Geneious R9.1.8 (Biomatters Ltd., Auckland, New Zealand). After alignment with all samples in the dataset and the revised Cambridge Reference Sequence (rCRS), SNP profiles were generated to include all loci at which alleles differed within Hypervariable Region I. To ensure sequence data quality, each DNA fraction was amplified and sequenced in duplicate.

Table 4.2. Forward and reverse primers for amplification and sequencing of Hypervariable Region I in mtDNA.

Primer	Sequence
F15975	5'-CTC CAC CAT TAG CAC CCA AA-3'
R16410m	5'-GAG GAT GGT GGT CAA GGG A-3'

4.2.6 Statistical Analysis

All statistical comparisons were performed in R (x64 version 3.4.4). Statistical significance was established at $\alpha = 0.05$. Repeated measures t-tests were performed using the *stats v3.5.0* package; equal variances were not assumed. Non-linear least-squares curve fitting was achieved using the *nls* function within the same package to determine degradation rates over

time since hair biosynthesis, with an assumption of first-order kinetics to model exponential decay, according to the following equations:

$$N(t) = N_0 e^{-\lambda t} \quad \text{Eq. 4.1}$$

and

$$t_{1/2} = \frac{\ln 2}{\lambda}, \quad \text{Eq. 4.2}$$

where $N(t)$ represents the quantity remaining after time t, N_0 is the initial quantity, λ is the exponential decay constant, and $t_{1/2}$ is the half-life, the time in which 50% degradation occurs. A maximum of 250 iterations was selected for curve-fitting. As hair age in this study was measured by distance of the hair segment relative to the root end, half-life was initially measured as distance and then converted to time, which is detailed in Section 4.3.1. All plots were drawn in OriginPro 2018 (OriginLab Corp., Northampton, MA).

4.3 Results and Discussion

Understanding hair proteome variation along the length of hair fibers provides insight into internal structural changes as hair fibers age, both naturally and while experiencing daily weathering. Section 4.3.1 assesses effects of hair age on hair protein chemistry, with the intent to quantify degradation in hair proteins throughout a period of 2 years of hair growth, a relevant timeframe for most forensic analyses, as hair proteome degradation may subsequently affect success rates in GVP detection.

4.3.1 Effects of Hair Age on the Hair Proteome

Hair age affects the proteins that are detected in one-inch hair segments. The numbers of identified proteins and unique peptides decrease over a length of 13 in from the root end (Figure 4.2, indicative of hair proteome degradation over hair growth time. Hair segment distance denoted in the figure represents the distance relative to the root end segment, which is designated

as 0 in, binned to the nearest inch. For example, the distal end segment from a 5-in long hair fiber is 4 in from the root end. On average, 136 ± 60 (mean ± s.d.) proteins and 1,007 ± 530 unique peptides were identified at the recently synthesized root ends, compared to 87 ± 48 proteins and 532 ± 392 peptides detected at the aged distal ends, sampled over a range of 4 to 13 in from the root end due to hair length differences among the 3 individuals. But because hair age at the distal ends varies from individual to individual, direct comparison of the metrics at the root and distal ends to quantify hair proteome degradation with hair age, as presented above, is not appropriate.

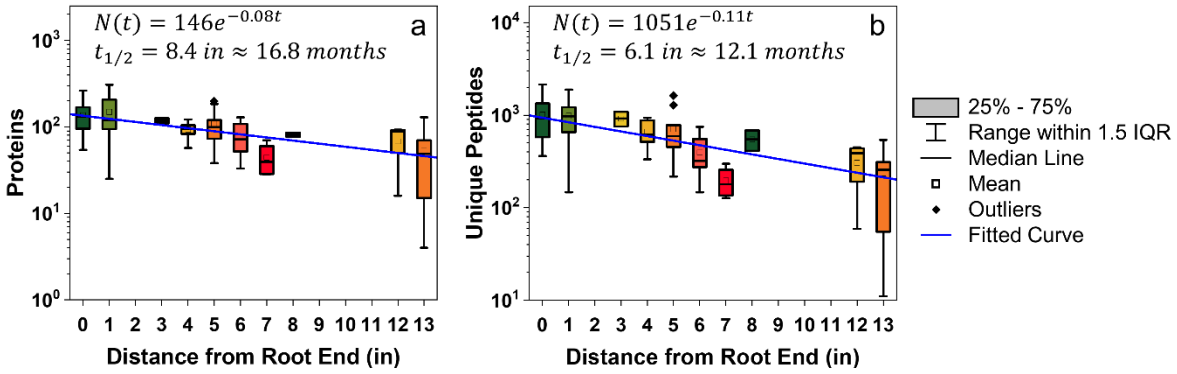


Figure 4.2. Number of identified (a) proteins and (b) unique peptides in single hair segments (n = 84, from 48 one-inch hair samples, with 36 analyzed before and after peptide/DNA co-fractionation) from 3 individuals, with increasing hair age, as measured by hair segment distance from the root end, in inches. Curve fitting was performed using non-linear least-squares methods for half-life determination and the fitted curve is plotted in blue. Half-life $t_{1/2}$ conversion from hair segment distance (in inches) to time (in months) assumes a 0.5-in per month hair growth rate[12]. The numbers of proteins and unique peptides detected in one-inch hair segments decrease with hair age, indicating protein degradation over hair growth time.

Instead, curve fitting to a decay model over the range of hair segment distances allows both a comparison of hair segments grouped by similar hair age and quantification of degradation rate. To quantify degradation rates for each metric, half-lives were determined after curve fitting using non-linear least-squares methods, assuming exponential decay under first-order kinetics and hair segment distance relative to the root end corresponds to time since

biosynthesis, or hair age, of the hair fiber. The dataset includes hair samples analyzed before and after peptide/DNA co-fractionation, totaling 84 samples from 48 one-inch hair segments, with 36 samples analyzed before and after co-fractionation, for more robust curve fitting. Half as many proteins and unique peptides are detected after 16.8 and 12.1 months, respectively, of hair growth (Figure 4.2); half-life conversion from hair segment distance to hair age assumes a growth rate of 0.5-in per month[12, 18]. This observation indicates that with increasing hair age, fewer proteins and peptides are detected; the hair proteome degrades with hair growth spanning 2 years.

Although the hair proteome degrades with hair age, continued detection of albeit smaller populations of proteins and peptides over 2 years of hair growth suggests non-uniform degradation rates among the identified proteins. Differences in protein degradation rates likely influence success of GVP detection, as more stable proteins would survive longer and permit detection of their resultant peptides for longer. To identify the proteins that are more likely to persist and from which GVP detection rates may be stable over 2 years of hair growth time, protein abundances and their degradation rates were compared. Protein abundance was determined as described in Chapter 3, Section 3.2.4 of this dissertation. Decreasing total protein abundance levels along the length of each hair fiber indicate protein degradation (Figure 4.3a), although the protein levels remaining with increasing hair age still permit protein identification. Of the 691 identified proteins, aggregated across the entire dataset, keratins comprise 7%, keratin-associated proteins (KAPs) 8%, and intracellular proteins 85%. 5% of all identified proteins did not show degradation in protein abundance over 2 years of growth time; this population includes 24% keratins, 6% KAPs, and 71% intracellular proteins. For these proteins, the protein abundance data did not fit well with the decay model; a larger sample size or

sampling of hair segments aged beyond 2 years of hair growth could perhaps improve the curve fit but for the purpose of examining degradation, this set of proteins, whose exponential decay constants were < 0, is considered to not have undergone degradation over this length of time. Figure 4.3b presents the distribution of protein half-lives whose abundance was derived by summing over the survey scan chromatographic peak areas for all associated unique peptides. Protein half-lives range from less than 6 months to more than 36 months. This analysis includes only those proteins that were identified in at least 25% of hair samples (n = 84) to have sufficient protein abundance data at the various segment distances for more robust curve fitting; 89 proteins are represented. A comprehensive list of these proteins is found in Appendix Table S-4.1 for keratins and KAPs, and Table S-4.2 for intracellular proteins.

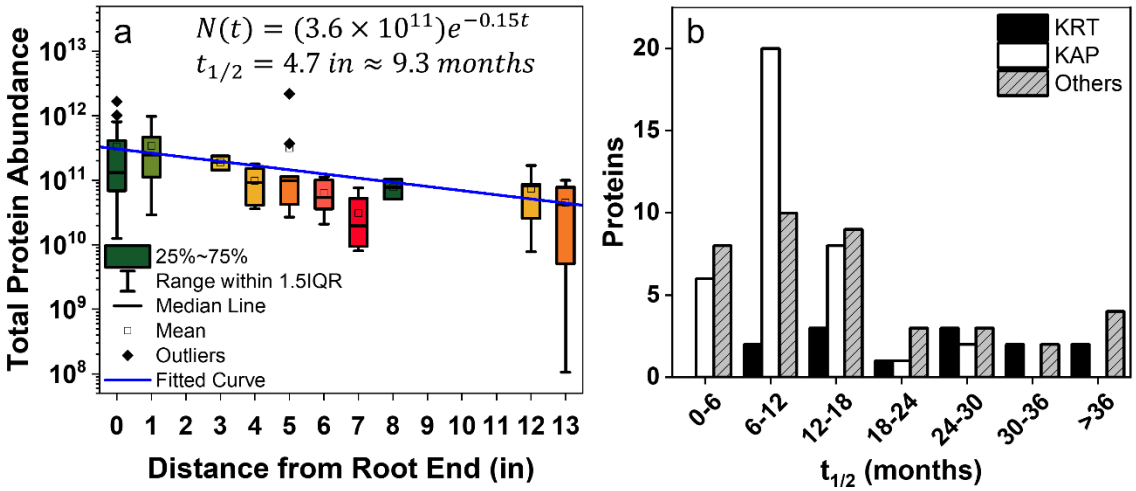


Figure 4.3. (a) Total protein abundance from one-inch hair segments (n = 84, from 48 one-inch hair samples, with 36 analyzed before and after peptide/DNA co-fractionation) from 3 individuals, arranged by increasing distance from root end, which serves as a proxy for hair age. Non-linear least-squares curve fitting was performed and half-life was converted from distance to time by assuming a 0.5-in per month growth rate. The fitted curve is plotted in blue. (b) Distribution of proteins by half-lives and category, including keratins (KRT), keratin-associated proteins (KAP), and intracellular proteins (Others). Proteins in plot represent 13% of all identified proteins. For these proteins, there was at least 25% detection among the hair samples and exponential decay constants (λ) from non-linear least-squares curve fitting were greater than 0. Decay in total protein abundance derives primarily from degradation of intracellular proteins and KAPs, of which the majority exhibit shorter half-lives in comparison to those of keratins.

Despite a wide range of half-lives, the vast majority of intracellular proteins and KAPs exhibit more rapid degradation rates (between 6 and 12 months) than those for keratins, and these shorter half-lives account for the shorter total protein abundance half-life. In contrast, the half-lives for keratins are more evenly distributed, with a few exceeding the hair growth time examined here. It is not surprising that the majority of intracellular proteins have shorter half-lives (Figure 4.3b), given that these proteins become exposed to external environments via degradation of nuclei and cellular organelles[19] as hair keratinocytes begin the differentiation process into corneocytes once emerged from the hair follicle, while simultaneously, keratins are synthesized and crosslinked during this time[20], which provides additional protection for this protein class. And as expected, keratins, the most abundant proteins in hair shaft, remain stable for longest; these proteins are stabilized by crosslinked disulfide bonds and isopeptide bonds[21], which are not easily cleaved without use of reagents such as oxidizing or reducing agents[5]. However, most notable are the degradation rates of KAPs; this class of proteins contains high sulfur content that crosslinks with keratins via disulfide bonds to promote the rigidity of hair shaft[22], and therefore, a similar half-life distribution to that of keratins might have been expected. But in fact, these findings agree with those described in Thibaut et al., who examined protein content along the lengths of scalp hair fibers spanning 2.6 m in length, corresponding to approximately 17 years of hair growth, and visually observed loss of KAP content at the distal ends via 2D gel electrophoresis, though content loss was not quantified[11]. Thibaut and co-workers attributed this degradation to the hair surface, citing inefficient protein extraction from hair cortex[11], but it is more likely a reflection of KAP content loss that includes the cortex, as the single hair preparation method described herein fully solubilizes one-inch hair segments for proteomic analysis and cortical proteins are well-represented in this analysis; keratins K31,

K33A, K33B, K34 – K39, K81, K83, K85, and K86, known to localize to the cortex[23], are detected at high abundance, 2.41×10^9 normalized counts on average, among hair segments, where average protein abundance maximizes at 8.21×10^9.

Further, localization of KAPs to hair fiber structure, either to hair cortex or cuticle, did not delineate differences in degradation rates. The half-life range of 6 – 12 months encompasses the largest group of KAPs including both cortical and cuticular KAPs[22, 24-27] from different KAP families, KAP1, KAP4, and KAP9 – 12 (Appendix Table S-4.1). However, interestingly, cortical KAPs, mostly from the KAP4 family[24], exhibited the fastest degradation rates, i.e., half-lives of under 6 months. In contrast, K40, localized to the cuticle[23], displayed a half-life of 10.5 months, which is the shortest half-life observed among keratins, followed by K82, also differentially expressed in cuticle[23], at 15.8 months (Table S-4.1), though the differences in degradation rates of keratins may not be linked to localization as the decay rates of cuticular keratins are not well-represented here. Of the 4 hair keratins documented to be localized only to hair cuticle[23], degradation rates from only 2 were determined here; K30 was not identified and K32 did not exhibit degradation over 2 years of hair growth. These results suggest that hair damage from daily weathering and natural aging occurs not only on the hair surface, but also affects the internal hair structure. A possible explanation for this may be that KAPs, the small hydrophobic proteins (< 30 kDa) that surround and link to keratin intermediate filaments in the amorphous matrix for stabilization of the hair fiber[28, 29], experience non-specific proteolysis to a greater extent from aging as they are not as protected from external stimuli including water and proteases as keratins, which are organized as intermediate filament (KIF) bundles in the cortex[13].

Hair proteome degradation with hair age is attributed primarily to degradation of KAPs, whose abundances in hair cortex and cuticle decay at a more accelerated rate than previously

found and compared to degradation of keratins. In particular, KAP GVP detection rates along the lengths of hair fibers are evaluated and compared to those from keratins for use as human identification markers in aged hairs, which is discussed in Section 4.3.3. The next section, Section 4.3.2, examines whether sample preparation methods to include mtDNA profiling affect GVP detection, irrespective of hair proteome degradation with hair age, as a viable approach to improve discriminative power from hair evidence.

4.3.2 Effects of Peptide/DNA Co-Fractionation on Peptide and GVP Identification

Although peptide/DNA co-fractionation presents an opportunity to improve discriminative power by enabling detection of GVPs from proteins and SNPs from mtDNA, effects of this technique on protein and peptide identification are not known and must be examined and quantified to determine whether GVP detection is compromised following co-fractionation. Peptide/DNA co-fractionation proceeds through on-column anion exchange, where DNA is retained while other analytes pass through. DNA is then eluted with a buffer containing high salt content. However, other anions and analytes may be inadvertently retained during this process. For example, acidic peptides, whose isoelectric points are below that of the solution pH, are anions and along with DNA, can be captured on the ion exchange column. Additionally, peptides from DNA-binding proteins may be retained with the DNA itself during anion exchange. A comprehensive list of DNA-binding proteins is included in Appendix Table S-4.3. The table includes those with confirmed DNA- and/or RNA-binding functionality *in vivo* as well as those determined to have this capability *in vitro*. As such, the peptide population in after co-fractionation digests may be biased towards fewer acidic peptides and peptides from DNA-binding proteins. Any loss in unique peptides, which is examined here, may subsequently

compromise GVP identification, leading to variable GVP detection rates in protein digests analyzed after co-fractionation.

Effects of co-fractionation on protein and peptide identification were first assessed. Table 4.3 displays the numbers of identified proteins, peptides, and SNPs from major and minor GVPs, aggregated across 36 digested hair samples that were analyzed before and after peptide/DNA co-fractionation. Protein digests yielded fewer unique peptides and SNPs after the co-fractionation process, but the number of identified proteins was greater after co-fractionation. The peptides that were not detected after co-fractionation were likely retained on the anion exchange column. From a total of 679 proteins and 7,589 unique peptides, overlaps of 56% and 42% in protein and peptide population, respectively, suggest that peptide/DNA co-fractionation facilitates detection of somewhat different peptides, perhaps peptides lower in abundance that were not identified in data-dependent analyses of single hairs processed without fractionation owing to coelution of peptides competing for selection for MS/MS analysis. In contrast, the SNPs inferred from major and minor GVPs before and after co-fractionation exhibited greater overlap (Table 4.3); 73% and 65% of SNPs from major and minor GVPs, respectively, overlapped in hair samples analyzed before and after co-fractionation. These findings suggest that despite peptide losses during co-fractionation, there exists a core SNP population that remains detected after peptide/DNA co-fractionation. However, there are SNPs that were only identified in one fraction and not the other; 18 and 8 SNPs were uniquely identified from major GVPs before and after co-fractionation, respectively, and 23 and 4 SNPs, respectively, from minor GVPs were unique to before and after co-fractionation samples (Table 4.3). A decrease in the number of identified peptides in the co-fractionation samples points to peptide retention on the anion exchange column, though peptide composition by protein class and acidity warrants further examination to

assess whether certain peptide populations, namely acidic peptides or those from DNA-/RNA-binding proteins, were preferentially retained on the anion exchange column, and the extent to which this compromises GVP identification and SNP inference, which is discussed below.

Table 4.3. Aggregate number of proteins, peptides, and SNPs from major and minor GVPs identified before and after peptide/DNA co-fractionation from a set of 36 hair samples from 3 individuals.

Metric	Before Co-Fractionation	After Co-Fractionation	Overlap	Total
Proteins	497	560	378	679
Unique Peptides	5,828	4,975	3,214	7,589
SNPs from Major GVPs	87	77	69	95
SNPs from Minor GVPs	74	55	51	78

To determine whether acidic peptides and those from DNA-binding proteins were lost during peptide/DNA co-fractionation, the distributions of peptide populations detected before and after co-fractionation were examined. Figure 4.4a-b displays the number of unique peptides from keratins, KAPs, DNA-/RNA-binding proteins, and other intracellular proteins in protein digest samples before and after co-fractionation, by isoelectric point. A peptide's average theoretical isoelectric point (pI) was determined by averaging across isoelectric points calculated by 17 different methods, which report slight variations on pK_a values for the 7 charged residues Glu, Asp, Cys, Tyr, His, Lys, and Arg, and the N- and C-terminal groups[30]; contributions from the remaining neutral residues to pI were considered negligible. Peptide net charge was determined by summing charge contributions from residues and termini, from imputing the pK_a values into the Henderson-Hasselbalch equation and assuming a certain pH; 6.5 was used as the starting pH. pH was then optimized to yield a peptide net charge of 0; this optimized value is the pI of the peptide[30]. As a theoretical value, the metric only accounts for contributions from the amino acid residues themselves and disregards those from chemical modifications. Because the

digest buffer during co-fractionation was at pH 8, peptides with pI < 8 have a higher chance of being lost due to on-column binding as an anion. Notably, the aggregate peptide population shifts substantially for keratin peptides with pI 4 – 5 (26% decrease) and for peptides from keratin-associated proteins with pI 7 – 8 (40% decrease) as a result of peptide/DNA co-fractionation (Figure 4.4c). Additionally, fewer unique peptides across the range of pIs (pI 3 – 12) and protein categories, including intracellular and DNA-/RNA-binding proteins, also nucleic acid-binding proteins, were also identified after co-fractionation (Figure 4.4d), suggesting that these peptides were also captured in tandem with DNA during anion exchange, even though the absolute differences in number of peptides were not as substantial as those for keratin and KAP peptides at the aforementioned pI ranges.

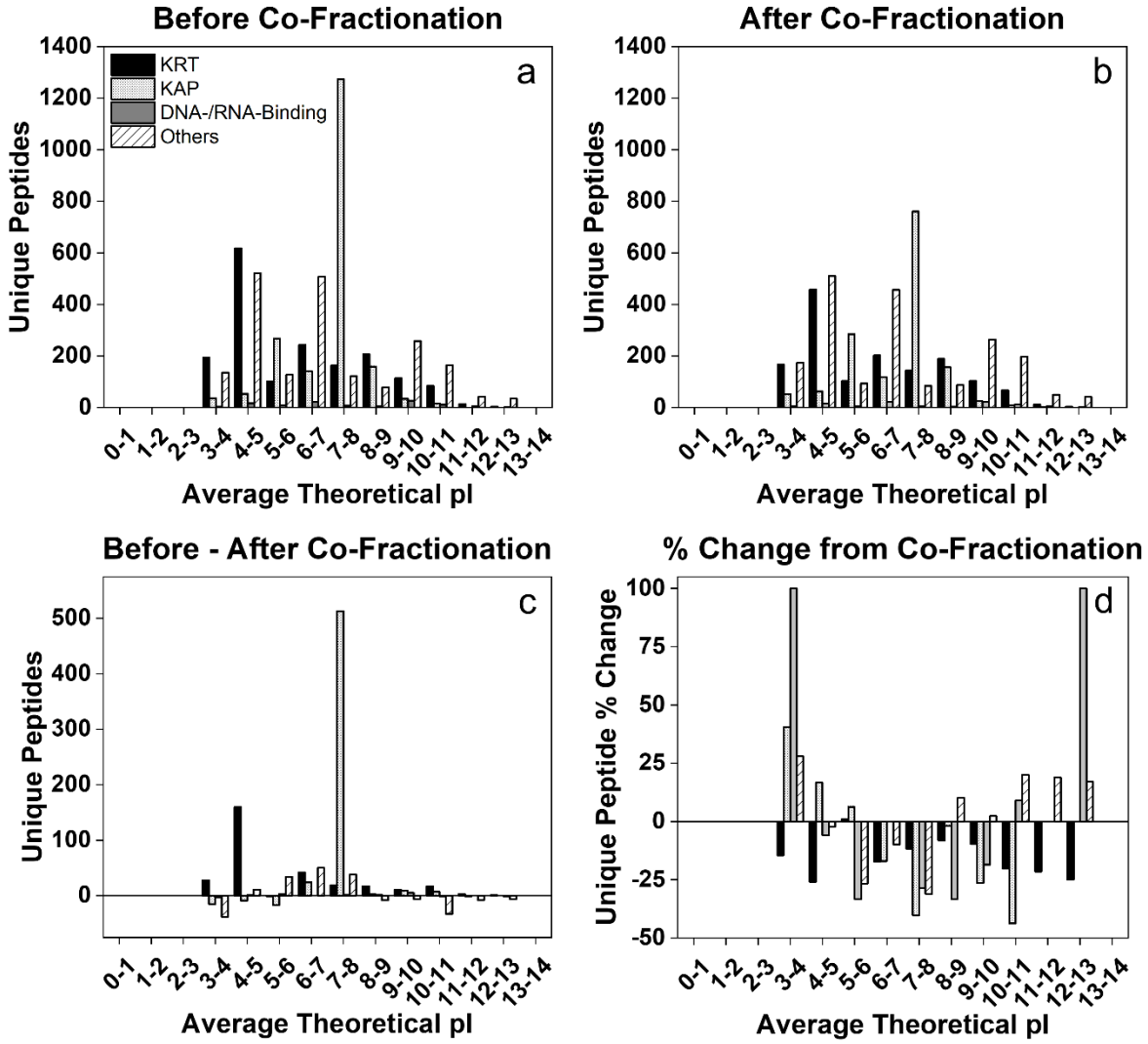


Figure 4.4. Distribution of unique peptides by protein class, keratins (KRT), keratin-associated proteins (KAP), DNA-/RNA-binding proteins, and intracellular proteins (Others), and by average theoretical isoelectric point (pI), identified (a) before and (b) after peptide/DNA co-fractionation, and (c) the difference in number of identified peptides before and after co-fractionation. (d) The percent change in unique peptides as a result of co-fractionation. After co-fractionation, fewer peptides from keratins with pI 4 – 5 and peptides from KAPs with pI 7 – 8 were detected, suggesting that these analytes were retained on the anion exchange column during peptide/DNA co-fractionation. Additionally, percent decreases in unique peptides from all protein classes with pI 3 – 12 indicate that while acidic peptides were retained on the anion exchange column, pI and peptide acidity represent a major but not the sole contributor to peptide loss during co-fractionation.

Direct assessment of GVPs further showed a 15% decrease in variant peptides from keratins with pI 4 – 5 and a decrease of 51% in those from KAPs with pI 7 – 8 after co-

fractionation, indicating that peptide loss from anion exchange during peptide/DNA co-fractionation indeed affects GVP detection. On average, statistically fewer SNPs were identified from major and minor GVPs, respectively, in each hair sample after co-fractionation (20 ± 12 and 11 ± 7 SNPs), compared to the 26 ± 14 and 15 ± 9 SNPs identified before co-fractionation (repeated measures t-test; $p \leq 4.12 \times 10^{-3}$; n = 36 samples per condition). However, these values only represent percent decreases of 23% and 27%, respectively, indicating that the majority, on average, 75%, of SNPs were recovered after co-fractionation.

Interestingly, the proteins in which GVPs are detected in this dataset do not overlap with any of the DNA-/RNA-binding proteins. Because DNA-/RNA-binding proteins are highly basic and are highly attracted to negatively-charged nucleic acids, it was thought that their tryptic peptides would remain bound to nucleic acids during anion exchange and would thus not be recoverable after co-fractionation. If this class of proteins yielded GVPs, these GVPs would potentially be lost during anion exchange. But of the 138 nucleic acid-binding proteins listed in Appendix Table S-4.3, only 30 proteins were identified among the 36 hair samples (n = 36 per co-fractionation condition), on average, 4 proteins identified in a hair segment, and GVPs were not identified from any of these proteins. Low detection rates of these proteins are likely owing to degradation of nuclei and organelles in hair shaft, which would result in exposure and subsequent degradation of these proteins. Further, these proteins were not well-detected in single hair samples even when co-fractionation was not performed (on average, 5 proteins identified in each hair segment), and thus, identification of their GVPs would be challenging even without performing anion exchange to separate DNA from peptides. Therefore, any effects of anion exchange on identification of DNA-/RNA-binding proteins, including GVP identification from this class of proteins, of which there were minimal in this dataset, would be negligible; in

contrast, peptide acidity is a more important factor that influences GVP detection after peptide/DNA co-fractionation.

Of particular concern is the loss of peptides from keratins and KAPs, as the majority of detected hair GVPs originate from the two protein classes. It was unexpected that greater peptide and GVP losses come from the peptides from KAPs than from the far more acidic peptides from keratins, since the KAP peptides are only slightly acidic, i.e., exhibiting pI 7 – 8, with respect to the solution pH 8, and it was thought that peptide losses would increase with acidity owing to more efficient binding of stronger anions, if peptide pI played a major role in peptide retention during anion exchange. However, these results and the loss of peptides with pI > 8 among all protein classes (Figure 4.4d) suggest that while peptide acidity remains an important variable for co-fractionation, mechanisms for peptide retention via anion exchange do not depend on peptide acidity alone and warrant further investigation.

Adjustment of solution pH in protein digests may be necessary to maximize recovery of acidic peptides during co-fractionation. As a preferable alternative, the sample preparation approach taken here could divide separate aliquots for proteomics (before co-fractionation) and mtDNA (after co-fractionation) analyses, and would be expected to minimize loss of acidic peptides while still permitting recovery of peptides and mtDNA from a single hair sample. The potential for increased discriminative power via detection of mtDNA SNPs, enabled by peptide/DNA co-fractionation, is discussed in Section 4.3.3.

4.3.3 Differentiative Potential in Aged Hairs Using GVPs and mtDNA SNPs

This work sought to evaluate success rates in detecting variant markers from proteins and mtDNA, and to quantify differentiative potential in aged hairs. Given their slower degradation rates relative to other proteins as observed in Section 4.3.1, particular attention was paid to GVP

detection variability from keratins, as compared to KAPs. Further, discriminative power from mtDNA SNP profiles was examined with increasing hair age and compared to differentiative potential obtained from GVP profiles.

The main advantage of peptide/DNA co-fractionation to single hair analysis is the increased specificity in discriminative power that may be achieved with a combination of GVPs from proteins and SNPs from mtDNA. From sequencing the Hypervariable Region I within mtDNA for each hair segment sample in this dataset, 10 SNP loci were identified. However, it is also known that mtDNA degrades over time with hair age[14], which may lead to compromised SNP profiles in aged hair samples. Figure 4.5 shows the rate of degradation for mtDNA, using fold difference of ND1 in one-inch hair segments as quantified via qPCR; this mtDNA gene codes for the protein NADH-ubiquinone oxidoreductase chain 1. Fold difference was calculated relative to 1 ng of human DNA positive control samples and serves as a proxy for mtDNA abundance. From this information, the half-life of mtDNA is estimated to be slightly more than 6 months, assuming a hair growth rate of 0.5-inch per month. This represents a more accelerated degradation rate (over 8-fold difference) compared to that reported by Brandhagen et al. (between 4- and 6-fold difference) over a similar period of time (calculated to be 1.6 years hair growth)[14]. Consequently, over a period of 2 years, approximately 12 in of hair growth or 4 half-lives, which is the hair age range of interest, degradation of mtDNA to $\frac{1}{16}$th of its abundance at the root end is expected. To determine the extent to which mtDNA SNP profiles are consistent over this period of time, i.e., the profiles of aged segments at the distal end match those in recently synthesized segments at the root end, intraindividual comparisons of mtDNA SNP profiles were performed.

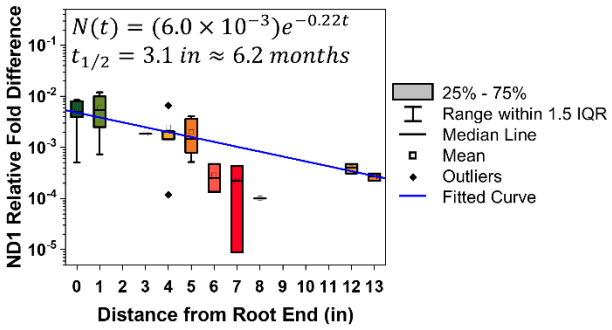
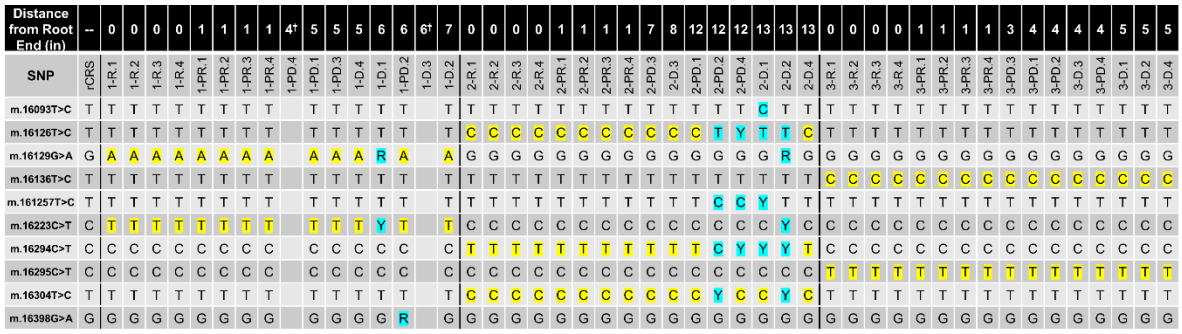


Figure 4.5. ND1 fold difference, relative to 1 ng of human DNA positive control, as a proxy for mtDNA abundance in each one-inch single hair segment (n = 48 hair samples). ND1 is a mtDNA gene. MtDNA degradation rate, as a half-life $t_{1/2}$, was determined by curve fitting the data using non-linear least squares. The fitted curve is plotted in blue. Conversion from segment distance (in inches) to time (in months) assumes 0.5-in per month hair growth rate.

SNP profiles from mtDNA became increasingly variable with hair age, especially at the end of 2 years of hair growth, where different or multiple nucleotide bases were observed, from Sanger sequencing following amplification, compared to those at the root end, likely owing to heteroplasmy, which is explained below. This study examined hair segments at the extremes, that is, 2 in from the root and distal ends of hair fibers. Displayed in Figure 4.6, the SNP profiles for two hair segments at the distal end from Individual 1 differ from that of the individual's consensus profile, established using the profiles from the majority of hair samples. For example, in 1-D.1, a distal end segment located 6 in from the root end, two alleles each are observed at 2 loci. At position 16129, both G and A alleles are present, and both C and T alleles are observed at position 16223, despite observations of only A and T alleles at the two loci, respectively, in the other hair segments. In Individual 2, profile variability also exists at the distal ends of each single hair, which represent 2 years of hair growth, with greater frequency among hair segments and loci. On the other hand, hair segments from Individual 3 remain consistent along the entirety of the length of each single hair, with its distal ends representing hair growth of roughly 10 months. Greater profile variability from mtDNA SNPs with hair age likely results from accumulation of mutations in its genome over time. MtDNA can exhibit heteroplasmy, or co-

existence of mutated and non-mutated alleles within tissue, and the extent of heteroplasmy varies among tissue types[33]; heteroplasmy is known to occur more frequently in hair owing to mtDNA passing through narrow bottlenecks during its development[31]. Hair forms from a small number of stem cells in hair follicle, which limits the number of mtDNA genotypes, creating a bottleneck[34]. Stochastic segregation of mtDNA during each of the numerous cell divisions to form the hair root and shaft introduces genetic drift, resulting in high variability of mtDNA genotypes, i.e., heteroplasmy[34]. This is one proposed mechanism to explain variability in mtDNA SNP profiles produced from distal end segments as compared to root end segments, that is, with increasing hair age. Alternatively, SNP profile variability may arise from PCR amplification or Sanger sequencing errors, though these errors are usually negligible, as mutations must be present at levels higher than 20% to be detected[35] and these levels are unlikely to be attained with single amplification or sequencing errors, when establishing a consensus within each sample. Further, samples were analyzed in duplicate, including an independent amplification process, to ensure data quality. Regardless of the mechanism that leads to profile variability, these results collectively suggest that mtDNA SNP profiles at the distal ends may not be consistent with those at the root ends after a year of hair growth, and more so after 2 years of hair growth.



Distance from Root End (in)	--	0	0	0	0	1	1	1	1	4†	5	5	5	6	6	6†	7	0	0	0	0	1	1	1	1	7	8	12	12	12	13	13	13	0	0	0	0	1	1	1	1	3	4	4	4	4	5	5	5
SNP	rCRS	1-R.1	1-R.2	1-R.3	1-R.4	1-PR.1	1-PR.2	1-PR.3	1-PR.4	1-PD.4	1-PD.1	1-PD.3	1-D.4	1-D.1	1-PD.2	1-D.3	1-D.2	2-R.1	2-R.2	2-R.3	2-R.4	2-PR.1	2-PR.2	2-PR.3	2-PR.4	2-PD.3	2-D.3	2-PD.1	2-PD.2	2-PD.4	2-D.1	2-D.2	2-D.4	3-R.1	3-R.2	3-R.3	3-R.4	3-PR.1	3-PR.2	3-PR.3	3-PR.4	3-PD.3	3-PD.1	3-PD.2	3-D.3	3-PD.4	3-D.1	3-D.2	3-D.4
m.16093T>C	T	T	T	T	T	T	T	T	T		T	T	T	T	T		T	T	T	T	T	T	T	T	T	T	T	T	C	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
m.16126T>C	T	T	T	T	T	T	T	T	T		T	T	T	T	T		T	C	C	C	C	C	C	C	C	C	C	T	Y	T	T	C	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
m.16129G>A	G	A	A	A	A	A	A	A	A		A	A	A	R	A		A	G	G	G	G	G	G	G	G	G	G	G	G	R	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
m.16136T>C	T	T	T	T	T	T	T	T	T		T	T	T	T	T		T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
m.161257T>C	T	T	T	T	T	T	T	T	T		T	T	T	T	T		T	T	T	T	T	T	T	T	T	T	T	C	C	Y	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
m.16223C>T	C	T	T	T	T	T	T	T	T		T	T	T	Y	T		T	C	C	C	C	C	C	C	C	C	C	C	C	Y	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
m.16294C>T	C	C	C	C	C	C	C	C	C		C	C	C	C	C		C	T	T	T	T	T	T	T	T	T	T	C	Y	Y	Y	T	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
m.16295C>T	C	C	C	C	C	C	C	C	C		C	C	C	C	C		C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
m.16304T>C	T	T	T	T	T	T	T	T	T		T	T	T	T	T		T	C	C	C	C	C	C	C	C	C	C	Y	C	C	Y	C	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
m.16398G>A	G	G	G	G	G	G	G	G	G		G	G	G	R			G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G

Figure 4.6. SNP profiles for each one-inch single hair sample (n = 48 hair samples) from the Hypervariable Region I in mtDNA, compared to the revised Cambridge Reference Sequence (rCRS). Profiles for each of the 3 individuals are arranged with increasing hair age. Hair segments are represented by x-y.a, where x is the individual code, y refers to the hair segment location (R: root end, PR: proximal-to-root, PD: proximal-to-distal, and D: distal end), and a is the biological replicate. SNPs are denoted using HGVS notation. Allele differences between the sample and the reference are highlighted in yellow. Each hair segment was analyzed in duplicate, and the allele at each site represents a consensus between the technical duplicates. Alleles highlighted in cyan indicate a difference from the individual's consensus profile, i.e., the allele observed in the majority of hair segments from the individual. Y represents the presence of both C and T alleles, and R represents the observation of both A and G alleles. †No SNP profile generated for hair sample.

Similar to mtDNA SNP profiles, comparison of intraindividual GVP profiles shows increasing profile variability with hair age (Figure 4.7), with variability in GVP profiles stemming from non-detection of one or more variant peptides. Figure 4.7 displays a set of simplified GVP profiles, where observed phenotype frequencies at each SNP locus represent the frequencies associated with detection of the major and/or minor GVPs as described in Chapter 2, Section 2.3.7 of this dissertation. Comprehensive profiles depicting observed phenotypes at each SNP locus for each one-inch segment are found in Appendix Tables S-4.4 – S-4.6 for Individuals 1 – 3, respectively. A one-SNP-per-gene rule was adopted for SNP selection to avoid issues of linkage disequilibrium, or co-inheritance of SNPs; in cases where multiple SNPs were identified from the same gene, SNPs that yielded the most consistent proteomic response in accordance with exome sequence genotypes were chosen. On average, data-dependent mass spectrometry analysis enabled inference of 28 ± 12 and 17 ± 8 SNPs from major and minor GVPs,

respectively, at root end segments, but GVP detection decayed over 2 years of hair growth time; half as many GVPs were detected after 13.7 and 11.7 months, respectively, which correspond to 6 – 7 in of hair growth from the root end (Appendix Figure S-4.2). Among the GVP profiles from Individuals 1 and 2, at proximal-to-distal hair segments located at least 6 in from the root end, fewer SNPs are inferred, compared to the number of SNPs at the root and proximal-to-root ends (Figure 4.7). In contrast, at proximal-to-distal and distal end segments from Individual 3, which are located 4 – 5 in from the root end, similar numbers of SNPs are detected compared to the root end. With increasing hair age, the number of GVPs detected in hair segments decrease, noticeably so beginning at hair segments 6 in from the root end, and as such, at proximal-to-distal and distal end segments, fewer GVPs are likely to be identified than at the root end.

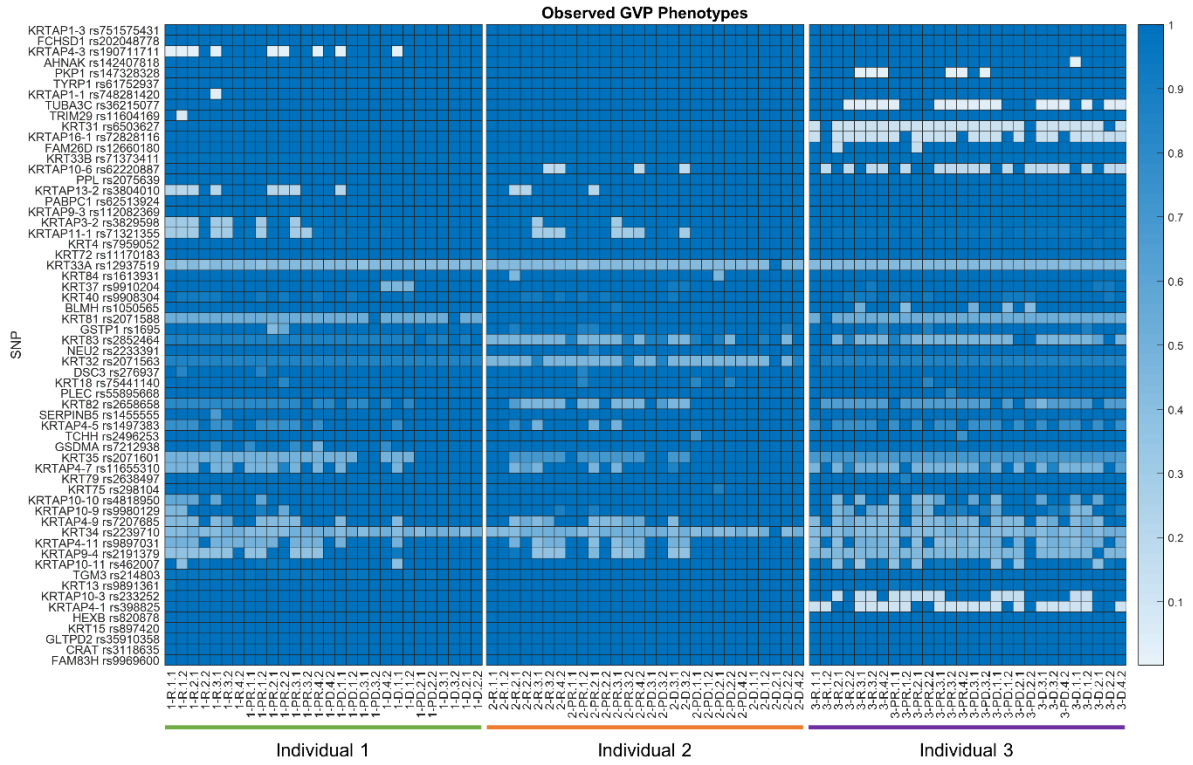


Figure 4.7. GVP profiles for each one-inch single hair sample (n = 84, from 48 hair samples, with 36 segments analyzed before and after peptide/DNA co-fractionation). Profiles for each of the 3 individuals are arranged with increasing hair age. Hair segments are represented by x-y.z.a, where x is the individual code, y refers to the hair segment location (R: root end, PR: proximal-to-root, PD: proximal-to-distal, and D: distal end), z is the biological replicate, and a represents the co-fractionation step (1: before, 2: after). Phenotype frequencies represent detection of a combination of the major and minor GVPs for the corresponding SNP.

Detection rates of GVPs differed between keratins and intracellular proteins, in line with their different hair age-related degradation rates. Notably, successful GVP detection permitted inference to 3 SNPs, rs12937519 from KRT33A, rs71373411 from KRT33B, and rs2239710 from KRT34, across all but 1 hair sample; a distal end segment from Individual 2 located 13 in from the root end (2-D.2.1) yielded no GVPs (Figure 4.7). It is not surprising that the GVPs that yield the most consistent response over 13 in of hair growth derive from keratins, as this class of proteins remains stable for longest of the three protein classes, keratins, KAPs, and intracellular proteins. Indeed, K33A exhibited a half-life of 17.3 months while the rate decay constants for K33B and K34 were not statistically different from 0 (one-sample t-test; $p \geq 0.550$), indicating

negligible evidence of degradation over 2 years of hair growth. On average, inferred SNPs from keratins have higher detection accuracy rates over this period of time, i.e., 50% accuracy in accordance with exome sequence genotypes compared to 45% from KAP SNPs and 35% from intracellular protein SNPs, which include both true positive and negative detection. When accounting solely for true positive responses, SNPs from intracellular proteins merely average 6% detection rate whereas those from keratins and KAPs average 32% and 33%, respectively, indicating that GVPs from intracellular proteins are not consistently detected, even at root end and proximal-to-root (PR) segments, resulting in variable SNP identification rates. However, these proteins are present even in aged hair samples, as intracellular proteins are identified 40% of the time, on average, among one-inch hair segments, based on detection of their unique peptides, though sequence coverage for intracellular proteins remains low, on average, 7% for the intracellular proteins that yield SNPs in this dataset. Many regions of these proteins are not well-covered, including those containing amino acid polymorphisms, which makes GVP identification from this protein class challenging and sporadic. GVP detection at root end segments varied owing to peptide ion competition for MS/MS fragmentation, as the data-dependent approach used in this work only selects the 10 most abundant ions per survey scan for fragmentation and subsequent identification. Given this variability and short half-lives, detection of variant peptides from degraded intracellular proteins at distal end segments becomes more challenging in contrast to keratins, which exhibit greater abundance and longer half-lives. Thus, GVPs from intracellular proteins are not as appealing as markers for human identification using the current analytical scheme; transition to data-independent mass spectrometry approaches may improve GVP detection reproducibility from this protein class.

As for detection of GVPs from KAPs, an average detection rate of 33% for true positive responses across all hair segments parallels that for keratins, which is unexpected as it was demonstrated earlier that KAPs exhibit an accelerated degradation rate with some proteins beginning to degrade within a month of hair growth. This suggests that even as the proteins degrade, likely via non-specific proteolysis from both environmental and internal sources, and abundance levels drop with hair age, their resultant GVPs are still of sufficient abundance for detection at similar rates to GVPs from keratins using the current analytical scheme. To examine abundance levels specifically for GVPs with increasing hair age, Figure 4.8 compares abundances summed over all GVPs corresponding to SNPs from proteins K83 and KAP4-1, from genes KRT83 and KRTAP4-1, respectively, with different degradation half-lives of 25.2 and 9.3 months, along the length of scalp hair. GVP abundances are delineated by major and minor variant. When detected, GVPs, particularly minor variants, exhibit high and similar abundance levels well above the threshold for MS/MS selection (i.e., 3.3×10^4 counts) with increasing hair age, even in hair segments located 12 – 13 in from the root end (Figure 4.8a). Even minor GVPs in KAP4-1 display high abundance levels in hair segments after approximately 16 months of hair growth (at 8 in from the root end), though many more non-detects were observed in aged hair segments (Figure 4.8b). The similarly high GVP abundance levels observed when the variant peptides were detected (i.e., at an abundance greater than the threshold 3.3×10^4) indicate that GVP abundance levels do not decay as rapidly as protein degradation rates suggest, and that variant peptides from keratins and KAPs display adequate abundance for detection in aged hair samples beyond a year of hair growth. GVP non-detection cannot solely be attributed to protein degradation rates with hair age. Instead, peptides from degraded proteins compete for MS/MS fragmentation in the data-dependent approach, as

discussed above, which accounts for much of GVP non-detection. As such, implementation of data-independent, targeted mass spectrometry methods in further development and routine operation of GVP analysis will provide better GVP detection reproducibility, especially in aged hair samples.

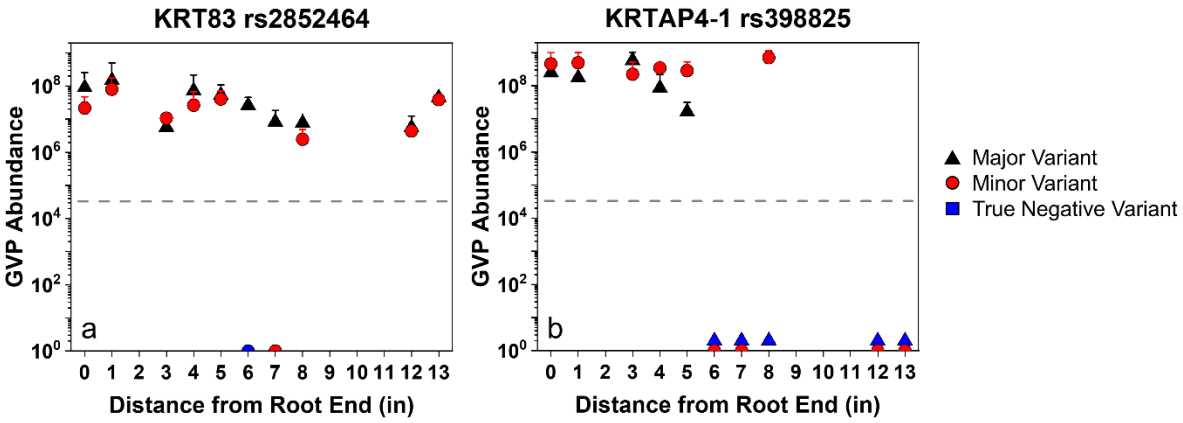


Figure 4.8. Abundances summed over all GVPs corresponding to the SNPs (a) rs2852464 from KRT83 and (b) rs398825 from KRTAP4-1, from each one-inch hair segment, averaged over segments with similar hair age across 3 individuals. Error bars represent standard deviation; only positive error bars are shown. GVP non-detects are indicated in the plot at an abundance of 1 for minor variants and 2 for major variants for completeness. Variants denoted by a blue-colored symbol (either a blue triangle for major variant or blue circle for minor variant) are those that yielded a true negative response, i.e., GVP non-detection in accordance with exome sequence genotype. The dashed line represents the threshold for MS/MS selection, set at 3.3×10^4 counts.

Discriminative power, measured via random match probability here, relies not only on the number of identified SNPs but also on the genotype frequencies associated with SNPs. Both contribute to differences in differentiative potential among individuals (Figure 4.9). Additionally, interindividual variation in hair physicochemical properties and hair age affect the number of identified SNPs. For example, among the 3 individuals, hair segments from Individual 2 yielded identification of the fewest SNPs (15 ± 6 SNPs at root end), resulting in low discriminative power (on average, 1.04×10^{-2} or 1 in 96 at root end) (Figure 4.9b), perhaps owing to difficulties in protein extraction from this hair specimen, as 17% fewer proteins were identified from root end segments compared to those from the other 2 individuals. With increasing hair age up to 2

years of hair growth, differentiative potential for GVP profiles from Individual 2 was quite low (2.77×10^{-1} or 1 in 4 at distal end) owing to many more GVP non-detects at distal end segments. Of the other 2 individuals, despite similar numbers of identified SNPs (21 ± 7 and 23 ± 6 SNPs for Individuals 1 and 2, respectively), profiles from Individual 3 yielded greater discriminative power (up to 1 in 1.85 trillion), indicating that SNP genotypes from this individual occur less frequently in the population (Figure 4.9c). On the other hand, Individual 1 has a more common profile, with associated SNP genotype frequencies closer to 1 at many loci (Figure 4.7). Interindividual variation, in hair type, protein degradation rates, and SNP genotypes, plays a large role in determining the extent of GVP profiling success and discriminative power. Many more GVP non-detects are encountered in aged hairs, especially between 1 and 2 years of hair growth. Though non-detects are expected to be minimized with data-independent approaches and detection of a set list of GVP targets in routine forensic analysis, where a large number of non-detects are observed in GVP profiling, the analysis will find greater utility in exclusionary purposes when applied to aged hair samples.

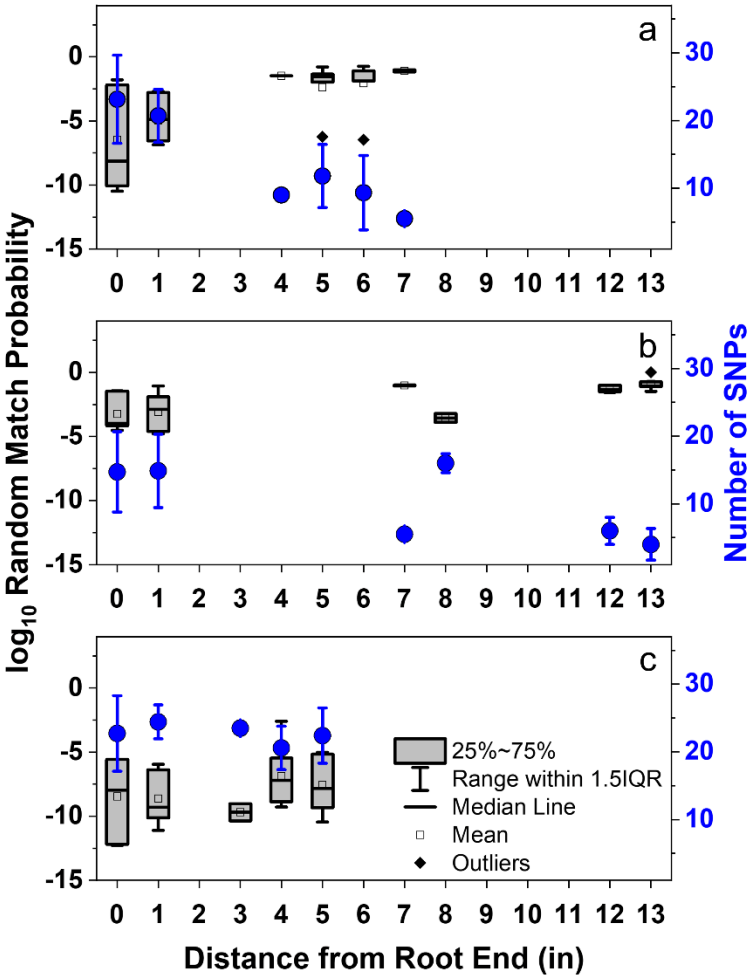


Figure 4.9. Random match probabilities, as a quantitation of discriminative potential, from products of GVP profiles for each one-inch single hair sample (n = 48 hair samples, with 36 segments analyzed before and after peptide/DNA co-fractionation) from Individuals (a) 1, (b) 2, and (c) 3 with increasing hair age. RMPs are plotted with corresponding numbers of SNPs (blue circles; mean ± s.d.), which account for the presence of both major and minor GVPs. Though RMPs improve with a greater number of identified SNPs, which varies with hair age and among individuals, the metric is also dependent on the genotype frequencies for the corresponding SNPs.

As with GVPs, quantification of discriminative potential with mtDNA SNPs using RMPs relies on genotype frequencies at the various loci. Because mtDNA is exclusively inherited from one parent, SNP genotype at each locus is represented by only one allele and allele frequencies can be used as genotype frequencies. Paralleling the use of global population frequencies for determining GVP phenotype frequencies, allele frequencies from all ancestral lineages (African,

Asian, and Eurasian) that are compiled in Mitomap's publicly available database[36] were utilized (downloaded October 16, 2018). Allele frequencies from the reference allele and all alternate alleles were included so that the total allele frequencies sum to 1 at each locus. To account for heteroplasmy at each locus, frequencies for all observed alleles are summed. For example, the hair segment denoted 1-D.1 (i.e., a distal end segment from Individual 1 located 6 in from the root end) exhibits heteroplasmy at position 16129; both A and G alleles are observed. Thus, the allele frequencies for A and G are summed to produce a genotype frequency of 0.994; other alternate alleles not observed in this hair segment exist within the global population captured in Mitomap. This approach is analogous to the calculation of observed phenotype frequencies for GVPs, described in Chapter 2, Section 2.3.7 of this dissertation. Products of genotype frequencies for the 10 SNPs identified in the Hypervariable Region I in mtDNA then permit a comparison of differentiative potential among the 3 individuals from one-inch hair segments.

From the 10 identified mtDNA SNPs, discriminative power maximized at 6.2×10^{-5}, or 1 in 16,197, for Individual 3 (Figure 4.10c). Of the 3 individuals, Individual 1 exhibits the most common profile, resulting in low discriminative power (majority consensus at 1 in 28; Figure 4.10a). Unlike with SNPs inferred from GVPs, 100% allele detection in successfully assembled profiles yielded similar differentiative potential from hair samples within each of the 3 individuals using mtDNA SNPs. But such differentiation became more nebulous with increasing heteroplasmy with hair age, especially after 2 years of hair growth as observed in profiling Individual 2 (Figure 4.10b). On the other hand, GVPs exhibited false negative responses even at root end and proximal-to-root segments, as a result of variable GVP detection associated with data-dependent mass spectrometry, which was discussed above. However, restriction of the analysis to Hypervariable Region I limits discriminative power that can be achieved with

mtDNA SNPs, as there are fewer positions for comparison that enable distinction of individuals relative to variations across the entire mtDNA sequence. From the same dataset, a panel of 60 SNPs inferred from GVPs was identified and assembled, from which RMPs ranged up to 1 in 1.85 trillion, even though marker detection rates varied with this approach. Given the success in identifying mtDNA Hypervariable Region I SNPs concomitant with GVPs from hair proteins, further work can expand to whole mtDNA sequencing. Although conventional forensic analysis of mtDNA utilizes Hypervariable Regions I – III[31], expanding to whole mtDNA sequencing to encompass the entire mitochondrial genome (16,569 bp), from which 11,270 variants (including SNPs) have been documented, would increase specificity and discriminative power from SNP profiles.

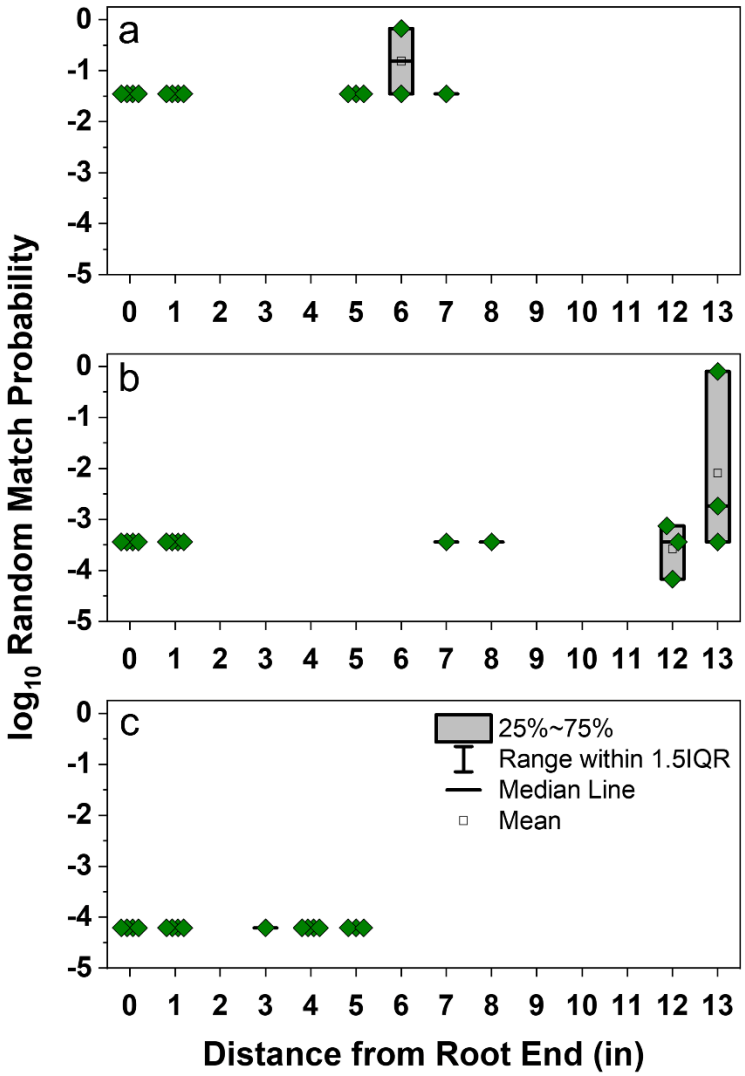


Figure 4.10. Random match probabilities from products of mtDNA SNP profiles for each one-inch single hair sample (n = 48 hair samples) from Individuals (a) 1, (b) 2, and (c) 3 with increasing hair age. Each data point (green diamond) represents a one-inch hair segment. As SNP profiles could not be generated for a hair segment each at 4 and 6 in from the root end for Individual 1, two fewer data points are represented in (a). Unlike RMPs reported from GVP identification, probabilities from mtDNA SNPs more clearly delineate the differentiative potential attained from one-inch hair segments for each individual, though restriction of SNP identification to Hypervariable Region I limits discriminative power.

4.4 Conclusions

This work confirms accelerated degradation rates of KAPs with increasing hair age and is the first to demonstrate this within 2 years of hair growth, providing evidence of hair age-related proteome degradation with greater time resolution. However, GVPs from KAPs remain

sufficiently abundant for detection during this range of hair age. Variability of detection of GVPs from KAPs in aged hairs likely arises from limitations of data-dependent LC-MS/MS when abundances of intact proteins are lower. Use of more sensitive, targeted mass spectrometry methods, for which development is underway, will facilitate detection of lower abundance GVPs and improve detection variability in aged hair samples, enabling application beyond utility as an exclusionary analysis in older hair samples.

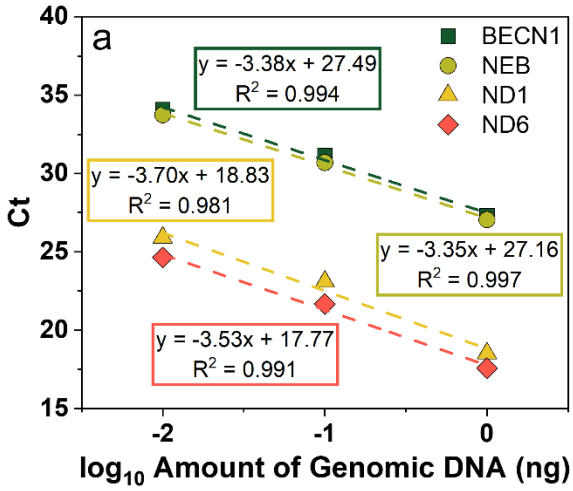
Supplementing GVP profiles with SNP profiles from mtDNA augments specificity of identifications with a greater number of biomarkers. Detection of both types of biomarkers with peptide/DNA co-fractionation makes hair a powerful evidence type. The advantages outweigh some GVP losses by co-fractionation, and these disadvantages could be minimized by processing separate aliquots for GVP and mtDNA SNP analyses. Though there are limitations in profile interpretation when applied to aged hairs, such as mtDNA heteroplasmy and GVP non-detection, combined use of GVPs from hair proteins and SNPs from mtDNA nevertheless offers clear benefits for forensic identification.

APPENDIX

Methods

Fluorescent Peptide Assay

Fluorescent peptide assays were performed to quantify aggregate peptide concentrations as an estimate of protein concentration in single one-inch hairs, modified from manufacturer's instructions in the Pierce™ Quantitative Fluorometric Peptide Assay kit (Thermo Scientific, Waltham, MA). To establish standard curves, 11 peptide standard solutions were prepared at concentrations between 0 and 500 μg/mL in 25 mM ammonium bicarbonate with 0.01% (w/v) ProteaseMAX™. To each 2-μL aliquot of peptide standard or protein digest sample, 14 μL of Assay Buffer followed by 4 μL of Assay Reagent were added and allowed to incubate for 30 min at RT. Fluorescence measurements of 2-μL aliquots were performed at excitation and emission wavelengths of 365 nm and 470 nm, respectively, using a Nanodrop™ 3300 Fluorospectrometer (Thermo Scientific, Waltham, MA). Measurements of peptide standards were obtained in duplicate. To ensure measurement reliability, 2 2-μL aliquots of protein digest samples were prepared and analyzed separately, and measurements averaged. Based on assay measurements and estimated peptide concentration levels in each sample, protein digest samples from Replicate Sets 1 and 2 were diluted to concentrations of 0.15 μg/μL and 0.10 μg/μL, respectively, based on these assays; to load 0.6 μg of peptide mass for liquid chromatography-tandem mass spectrometry analysis, injection volumes of 4 μL and 6 μL, respectively, were chosen.



b

Gene	qPCR Efficiency (%) (mean ± s.d.)	% CV
BECN1	99.5 ± 1.1	1.1
NEB	98.6 ± 6.1	6.2
ND1	91.1 ± 7.5	8.2
ND6	101.7 ± 12.6	12.4

c

$$Amount\ of\ DNA_{sample} = 10^{\frac{Ct_{ND1} - intercept}{slope}} \times Dilution\ Factor$$

$$Amount\ of\ DNA_{ref,1\ ng} = 10^{\frac{Ct_{ND1,1\ ng} - intercept}{slope}} \approx 1$$

$$ND1\ Fold\ Difference = \frac{Amount\ of\ DNA_{sample}}{Amount\ of\ DNA_{ref,1\ ng}}$$

Figure S-4.1. The process by which mitochondrial DNA (mtDNA) abundance was quantified via quantitative real-time PCR (qPCR). (a) Representative standard curve of cycle threshold (Ct) values over a range of known genomic DNA amounts for each gene. BECN1 and NEB, genes from nuclear DNA; ND1 and ND6, genes from mtDNA. Linear curve fitting was performed and R^2 was used to evaluate goodness-of-fit. (b) Evaluation of qPCR efficiency for each gene, using coefficient of variation (% CV). Of the genes from mtDNA, ND1 was selected as a proxy for mtDNA owing to its less variable qPCR amplification efficiency relative to ND6. (c) Equations for determining ND1 fold difference, relative to 1 ng of genomic DNA in the human DNA positive control sample. The dilution factor accounts for DNA lost to aliquots for mass spectrometry and fluorescent peptide assays prior to peptide/DNA co-fractionation; the amount of DNA in a one-inch hair segment is calculated here. ND1 relative fold difference serves as a proxy for mtDNA abundance.

Table S-4.1. Half-lives of keratins and KAPs that degrade over 2 years of hair growth, with citation of their localization in hair fiber as determined from mRNA expression. '--' indicates that expression was not found. *denotes expression in epithelia.

Protein	Detection Frequency (%)	N_0	λ	$t_{1/2}$ (in)	$t_{1/2}$ (months)	Localization	Citation
KAP4-12	31	3.02E+08	2.141	0.3	0.6	cortex	[24]
KAP4-6	65	4.41E+08	1.534	0.5	0.9	cortex	[24]
KAP4-5	56	4.47E+07	0.373	1.9	3.7	cortex	[24]
KAP9-1	51	2.51E+07	0.324	2.1	4.3	cortex	[24]
KAP4-2	70	5.52E+08	0.252	2.8	5.5	cortex	[24]
KAP4-9	61	2.17E+08	0.246	2.8	5.6	cortex	[24]
KAP10-8	30	6.04E+07	0.230	3.0	6.0	cuticle	[25]
KAP9-2	65	9.13E+07	0.212	3.3	6.5	cortex	[24]
KAP12-3	40	1.73E+07	0.200	3.5	6.9	cuticle	[25]
KAP10-11	69	2.80E+08	0.176	3.9	7.9	cuticle	[25]
KAP4-3	73	1.41E+09	0.171	4.1	8.1	cortex	[24]
KAP10-9	55	4.42E+07	0.167	4.1	8.3	cuticle	[25]
KAP9-3	73	1.13E+09	0.162	4.3	8.6	cortex	[24]
KAP10-12	71	4.03E+08	0.160	4.3	8.7	cuticle	[25]
KAP4-11	69	3.67E+08	0.153	4.5	9.0	cortex	[24]
KAP4-1	60	3.27E+08	0.149	4.7	9.3	cortex	[24]
KAP4-4	71	1.38E+09	0.146	4.7	9.5	cortex	[24]
KAP9-6	69	2.39E+08	0.144	4.8	9.6	cortex	[24]
KAP10-3	44	3.18E+06	0.142	4.9	9.8	cuticle	[25]
KAP1-3	61	6.06E+07	0.140	5.0	9.9	cortex	[22]
KAP10-10	65	1.58E+08	0.139	5.0	10.0	cuticle	[25]
KAP9-4	56	4.45E+07	0.136	5.1	10.2	cortex	[24]
KAP9-7	71	3.30E+08	0.132	5.2	10.5	cortex	[24]
KAP9-9	70	1.25E+08	0.132	5.3	10.5	cortex	[24]
KAP11-1	77	1.73E+09	0.128	5.4	10.8	matrix and cortex	[26]
KAP1-1	55	3.43E+07	0.121	5.7	11.4	cortex	[27]
KAP3-3	74	9.20E+08	0.109	6.3	12.7	cortex	[24]
KAP4-8	71	5.63E+08	0.103	6.7	13.5	cortex	[23]
KAP10-6	54	2.42E+07	0.102	6.8	13.6	cuticle	[25]
KAP9-8	68	1.42E+08	0.094	7.4	14.8	cortex	[24]
KAP4-7	56	1.50E+08	0.083	8.3	16.7	cortex	[23]
KAP16-1	76	1.24E+08	0.082	8.4	16.8	--	[24]
KAP3-1	80	4.57E+09	0.080	8.7	17.3	cortex	[24]
KAP3-2	74	1.15E+09	0.080	8.7	17.3	cortex	[24]
KAP24-1	64	1.54E+08	0.060	11.6	23.2	cuticle	[37]
KAP1-5	89	3.24E+08	0.055	12.6	25.2	cortex	[24]
KAP13-2	71	8.43E+07	0.052	13.2	26.5	cortex and cuticle	[26]

Table S-4.1 (cont'd)

Protein	**Detection Frequency (%)**	N_0	λ	$t_{1/2}$ **(in)**	$t_{1/2}$ **(months)**	**Localization**	**Citation**
K40	73	2.48E+07	0.132	5.3	10.5	cuticle	[23]
K79	27	3.10E+07	0.119	5.8	11.7	*	[23]
K82	96	5.85E+08	0.088	7.9	15.8	cuticle	[23]
K36	67	4.45E+07	0.081	8.6	17.1	cortex	[23]
K33A	99	3.23E+09	0.080	8.7	17.3	cortex	[23]
K1	33	1.15E+07	0.071	9.8	19.6	*	[23]
K84	81	1.08E+08	0.057	12.2	24.4	--	[23]
K83	95	2.03E+09	0.055	12.6	25.2	cortex	[23]
K85	100	9.82E+09	0.054	12.8	25.7	cortex and cuticle	[23]
K81	88	2.02E+09	0.043	16.1	32.1	cortex	[23]
K86	99	6.23E+09	0.041	16.7	33.5	cortex	[23]
K35	99	1.11E+09	0.027	25.7	51.4	cortex and cuticle	[23]
K31	100	4.66E+09	0.024	28.7	57.4	cortex	[23]

Table S-4.2. Half-lives of intracellular proteins that degrade over 2 years of hair growth.

Protein	Detection Frequency (%)	N_0	λ	$t_{1/2}$ (in)	$t_{1/2}$ (months)
HSPB1	38	1.42E+07	0.733	0.9	1.9
DSC3	25	8.58E+06	0.530	1.3	2.6
MDH2	26	5.39E+06	0.458	1.5	3.0
VSIG8	94	7.59E+08	0.377	1.8	3.7
GAPDH	42	4.41E+07	0.358	1.9	3.9
FABP4	33	5.43E+06	0.284	2.4	4.9
PPL	40	5.26E+06	0.280	2.5	5.0
MIF	49	3.56E+07	0.232	3.0	6.0
VDAC2	30	8.47E+06	0.215	3.2	6.5
LYG2	46	9.92E+06	0.202	3.4	6.9
S100A3	40	8.78E+07	0.169	4.1	8.2
HEPHL1	45	1.89E+07	0.153	4.5	9.0
CRYBG1	54	6.56E+06	0.151	4.6	9.2
TPI1	30	5.20E+06	0.143	4.8	9.7
EEF2	37	6.03E+06	0.141	4.9	9.8
TGM3	71	1.68E+07	0.134	5.2	10.3
CTNNB1	48	6.05E+06	0.133	5.2	10.4
SFN	61	4.21E+07	0.118	5.9	11.8
HSPA8	31	1.42E+06	0.114	6.1	12.1
LMNA	32	4.78E+06	0.099	7.0	14.0
PLCD1	45	9.43E+06	0.099	7.0	14.1
TRIM29	52	1.43E+07	0.096	7.2	14.5
HIST1H4A	74	3.00E+08	0.092	7.5	15.0
SERPINB5	51	1.37E+07	0.090	7.7	15.4
PKP1	89	5.75E+07	0.088	7.9	15.8
PRDX6	45	7.79E+06	0.087	8.0	15.9
GPNMB	46	2.18E+07	0.087	8.0	16.0
PPIA	49	1.09E+07	0.073	9.4	18.9
GSTP1	45	2.26E+07	0.073	9.5	18.9
HSPA2	49	6.87E+06	0.062	11.1	22.3
RPSA	36	8.68E+06	0.056	12.4	24.8
TUBB4B	29	1.28E+06	0.056	12.5	24.9
FAM26D	54	2.25E+07	0.051	13.7	27.4
HSP90AA1	29	8.85E+06	0.046	15.0	30.1
ATP5B	57	6.02E+06	0.039	17.6	35.1
CHUK	26	2.03E+07	0.036	19.0	38.0
IDH2	32	4.04E+06	0.021	32.5	65.0
PROCR	31	4.41E+06	0.013	52.5	105.0
TUBB2A	43	3.41E+06	0.002	336.7	673.3

Table S-4.3. List of human proteins that bind DNA and/or RNA, including those capable of the functionality *in vitro*, identified by Hudson and Ortlund[38], and histones listed in the UniProtKB SwissProt Human database[39].

Gene	Protein Name
ADAR	Double-stranded RNA-specific adenosine deaminase
AICDA	Single-stranded DNA cytosine deaminase
ANXA2	Annexin A2
APTX	Aprataxin
BAZ2A	Bromodomain adjacent to zinc finger domain protein 2A or Transcription termination factor I-interacting protein 5
BCLAF1	Bcl-2-associated transcription factor 1
BRCA1	Lys-63-specific deubiquitinase BRCC36
C1D	Nuclear nucleic acid-binding protein C1D
CARHSP1	Calcium-regulated heat-stable protein 1
CBX5	Chromobox protein homolog 5 or Heterochromatin protein 1 homolog alpha
CBX7	Chromobox protein homolog 7
CEBPZ	CCAAT/enhancer-binding protein zeta or CCAAT-box-binding transcription factor
CENPC	Centromere protein C
CNBP	Cellular nucleic acid-binding protein or Zinc finger protein 9
CSDE1	Cold shock domain-containing protein E1
CSTF2	Cleavage stimulation factor subunit 2 or CF-1 64 kDa subunit
DDX3X	ATP-dependent RNA helicase DDX3X or DEAD box protein 3, X-chromosomal
DEK	Protein DEK
DHX36	ATP-dependent DNA/RNA helicase DHX36
DLX2	Homeobox protein DLX-2
DNMT1	DNA (cytosine-5)-methyltransferase 1
DNTTIP2	Deoxynucleotidyltransferase terminal-interacting protein 2
DPPA3	Developmental pluripotency-associated protein 3
DUSP11	RNA/RNP complex-1-interacting phosphatase
ENO1	Alpha-enolase
ERCC6	DNA excision repair protein ERCC-6
ESR1	Estrogen receptor
FUBP1	Far upstream element-binding protein 1
FUS	RNA-binding protein FUS

Table S-4.3 (cont'd)

Gene	Protein Name
G3BP1	Ras GTPase-activating protein-binding protein 1
GAPDH	Glyceraldehyde-3-phosphate dehydrogenase
GTF3A	Transcription factor IIIA
H1-0	Histone H1.0
H1FNT	Testis-specific H1 histone
H1FOO	Histone H1oo or Oocyte-specific histone H1
H1FX	Histone H1x
H2AB1	Histone H2A-Bbd type 1
H2AB2 H2AB3	Histone H2A-Bbd type 2/3
H2AFJ	Histone H2A.J
H2AFV	Histone H2A.V
H2AFX	Histone H2AX
H2AFY	Core histone macro-H2A.1
H2AFY2	Core histone macro-H2A.2
H2AZ1	Histone H2A.Z
H2BW2	Histone H2B type F-M
H2BFS	Histone H2B type F-S
H2BFWT	Histone H2B type W-T
H3-3A H3-3B	Histone H3.3
H3F3C	Histone H3.3C or Histone H3.5
GPIHBP1	Glycosylphosphatidylinositol-anchored high density lipoprotein-binding protein 1
HDGF	Hepatoma-derived growth factor
H1-5	Histone H1.5
H3C1 H3C2 H3C3 H3C4 H3C6 H3C7 H3C8 H3C10 H3C11 H3C12	Histone H3.1

Table S-4.3 (cont'd)

Gene	Protein Name
H4C1 H4C2 H4C3 H4C4 H4C5 H4C6 H4C8 H4C9 H4C11 H4C12 H4C13 H4C14 H4C15 H4-16	Histone H4
HIST2H2AB	Histone H2A type 2-B
HIST2H2AC	Histone H2A type 2-C
HIST2H3A HIST2H3C HIST2H3D	Histone H3.2
HIST3H2BB	Histone H2B type 3-B
HMGB1	High mobility group protein B1
HNRNPA1	Heterogeneous nuclear ribonucleoprotein A1
HNRNPC	Heterogeneous nuclear ribonucleoproteins C1/C2
HNRNPD	Heterogeneous nuclear ribonucleoprotein D0
HNRNPK	Heterogeneous nuclear ribonucleoprotein K
HNRNPL	Heterogeneous nuclear ribonucleoprotein L
HNRNPU	Heterogeneous nuclear ribonucleoprotein U
IGHMBP2	DNA-binding protein SMUBP-2
ILF3	Interleukin enhancer-binding factor 3
KHDRBS1	KH domain-containing, RNA-binding, signal transduction-associated protein 1 or GAP-associated tyrosine phosphoprotein p62
KIN	DNA/RNA-binding protein KIN17
LDHA	L-lactate dehydrogenase A chain
LIN28A	Protein lin-28 homolog A or Zinc finger CCHC domain-containing protein 1
LONP1	Lon protease homolog, mitochondrial
LRPPRC	Leucine-rich PPR motif-containing protein, mitochondrial
LRRFIP1	Leucine-rich repeat flightless-interacting protein 1
MECP2	Methyl-CpG-binding protein 2

Table S-4.3 (cont'd)

Gene	Protein Name
NABP1	SOSS complex subunit B2 or Single-stranded DNA-binding protein 2
NACA	Nascent polypeptide-associated complex subunit alpha
NAT10	RNA cytidine acetyltransferase
NCL	Nucleolin
NFKB1	Nuclear factor NF-kappa-B p105 subunit
NFX1	Transcriptional repressor NF-X1
NFYA	Nuclear transcription factor Y subunit alpha
NKRF	NF-kappa-B-repressing factor
NONO	Non-POU domain-containing octamer-binding protein
NR0B1	Nuclear receptor subfamily 0 group B member 1
NR3C1	Glucocorticoid receptor or Nuclear receptor subfamily 3 group C member 1
NR5A1	Steroidogenic factor 1 or Nuclear receptor subfamily 5 group A member 1
NSUN2	RNA cytosine C_5-methyltransferase NSUN2
NUFIP1	Nuclear fragile X mental retardation-interacting protein 1
PARP1	Poly(ADP-ribose) polymerase 1
PCBP1	Poly(rC)-binding protein 1
PGK1	Phosphoglycerate kinase 1
PRNP	Major prion protein
PTBP1	Polypyrimidine tract-binding protein 1
PURA	Transcriptional activator protein Pur-alpha or Purine-rich single-stranded DNA-binding protein alpha
RAD51AP1	RAD51-associated protein 1
RARA	Retinoic acid receptor alpha or Nuclear receptor subfamily 1 group B member 1
RBM3	RNA-binding protein 3
RBMS1	RNA-binding motif, single-stranded-interacting protein 1
RPL7	60S ribosomal protein L7
RTF1	RNA polymerase-associated protein RTF1 homolog
RUNX1	Runt-related transcription factor 1
RUVBL1	RuvB-like 1
SAFB	Scaffold attachment factor B1
SFPQ	Splicing factor, proline- and glutamine-rich or Polypyrimidine tract-binding protein-associated-splicing factor

Table S-4.3 (cont'd)

Gene	Protein Name
SMAD1	Mothers against decapentaplegic homolog 1 or Transforming growth factor-beta-signaling protein 1
SMN1 SMN2	Survival motor neuron protein or Gemin-1
SON	Protein SON
SOX2	Transcription factor SOX-2
SPEN	Msx2-interacting protein
SRSF1	Serine/arginine-rich splicing factor 1
SSBP1	Single-stranded DNA-binding protein, mitochondrial
STAT1	Signal transducer and activator of transcription 1-alpha/beta
SUB1	Activated RNA polymerase II transcriptional coactivator p15
SUPV3L1	ATP-dependent RNA helicase SUPV3L1, mitochondrial
SURF6	Surfeit locus protein 6
SUZ12	Polycomb protein SUZ12
TAF15	TATA-binding protein-associated factor 2N
TARDBP	TAR DNA-binding protein 43
TBX5	T-box transcription factor TBX5
TCF7	Transcription factor 7
TERF2	Telomeric repeat-binding factor 2
THRA	Thyroid hormone receptor alpha
TIA1	Nucleolysin TIA-1 isoform p40
TOP1	DNA topoisomerase 1
TP53	Cellular tumor antigen p53
TSN	Translin
WBP11	WW domain-binding protein 11
WT1	Wilms tumor protein
XRCC5	X-ray repair cross-complementing protein 5
XRCC6	X-ray repair cross-complementing protein 6
YBX1	Y-box-binding protein 1
YY1	Transcriptional repressor protein YY1
ZC3H8	Zinc finger CCCH domain-containing protein 8
ZNF239	Zinc finger protein 239
ZNF638	Zinc finger protein 638

Table S-4.4. Comprehensive GVP profiles for single hair samples from Individual 1, arranged by increasing segment distance from the root end in inches. Sample codes are denoted x-y.z.a, where x is the individual code, y is the hair segment (R: root, PR: proximal-to-root, PD: proximal-to-distal, D:distal), z is the sample replicate number, and a indicates the co-fractionation step (1: before, 2: after). 0 and 1 indicate detection of the major and minor GVP, respectively, and '--' indicates the non-detection of either variant.

Distance from Root End (in)	0	0	0	0	0	0	0	1	1	1	1	1	1	1	4	5	5	5	5	5	6	6	6	6	6	6	7	7
Gene SNP	**1-R1.1**	**1-R1.2**	**1-R2.1**	**1-R2.2**	**1-R3.1**	**1-R3.2**	**1-R4.2**	**1-PR1.1**	**1-PR1.2**	**1-PR2.1**	**1-PR2.2**	**1-PR3.1**	**1-PR3.2**	**1-PR4.2**	**1-PD4.2**	**1-PD1.1**	**1-PD1.2**	**1-PD3.1**	**1-PD3.2**	**1-D4.2**	**1-D1.1**	**1-D1.2**	**1-PD2.1**	**1-PD2.2**	**1-D3.1**	**1-D3.2**	**1-D2.1**	**1-D2.2**
KRTAP1-3 rs751575431	0	0	0	--	0	--	--	0	0	0	0	0	--	0	--	0	0	--	--	--	0	0	--	--	--	--	--	--
FCHSD1 rs202048778	--	--	--	--	--	--	--	--	--	--	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRTAP4-3 rs190711711	0,1	0,1	0,1	--	0,1	0	--	0	0	1	0,1	0	--	0,1	--	1	--	--	--	--	0,1	0	--	--	--	--	--	--
AHNAK rs142407818	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
PKP1 rs147328328	--	0	--	--	0	0	0	--	0	--	--	0	0	0	--	--	--	0	--	--	--	--	--	0	--	--	--	--
TYRP1 rs61752937	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRTAP1-1 rs748281420	0	0	0	--	0,1	0	--	--	0	--	--	0	--	--	--	--	--	--	--	0	--	--	--	--	--	--	--	--
TUBA3C rs36215077	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
TRIM29 rs11604169	--	1	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRT31 rs6503627	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRTAP16-1 rs72828116	0	0	0	0	--	0	--	--	0	0	0	0	--	--	--	0	0	--	--	--	0	0	--	--	--	--	--	--
FAM26D rs12660180	0	--	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRT33B rs71373411	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRTAP10-6 rs62220887	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
PPL rs2075639	0	--	0	--	--	--	--	--	--	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRTAP13-2 rs3804010	1	0,1	1	--	1	--	--	--	--	1	1	1	--	--	--	1	--	--	--	--	--	--	--	--	--	--	--	--
PABPC1 rs62513924	--	0	--	--	--	--	--	--	--	--	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRTAP9-3 rs112082369	0	0	0	0	0	0	--	--	0	0	0	0	0	0	--	--	0	--	--	0	--	--	--	--	--	--	--	--
KRTAP3-2 rs3829598	0,1	0,1	0,1	0	0,1	0,1	0	0	0,1	0	0	0,1	0	0	0	0	0	--	--	0	0	0	--	--	0	--	--	--

Table S-4.4 (cont'd)

Distance from Root End (in)	0	0	0	0	0	0	0	1	1	1	1	1	1	1	4	5	5	5	5	5	6	6	6	6	6	6	7	7
Gene SNP	**1-R.1.1**	**1-R.1.2**	**1-R.2.1**	**1-R.2.2**	**1-R.3.1**	**1-R.3.2**	**1-R.4.2**	**1-PR.1.1**	**1-PR.1.2**	**1-PR.2.1**	**1-PR.2.2**	**1-PR.3.1**	**1-PR.3.2**	**1-PR.4.2**	**1-PD.4.2**	**1-PD.1.1**	**1-PD.1.2**	**1-PD.3.1**	**1-PD.3.2**	**1-D.4.2**	**1-D.1.1**	**1-D.1.2**	**1-PD.2.1**	**1-PD.2.2**	**1-D.3.1**	**1-D.3.2**	**1-D.2.1**	**1-D.2.2**
KRTAP11-1 rs71321355	0,1	0,1	0,1	--	0,1	0,1	--	0	0,1	--	--	0,1	0,1	0	--	0	--	--	--	0	--	--	--	--	--	--	--	--
KRT4 rs7959052	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRT72 rs11170183	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRT33A rs12937519	1	0,1	1	0,1	1	1	1	1	0,1	1	1	1	1	1	1	1	1	0,1	1	1	1	1	1	1	1	1	1	1
KRT84 rs1613931	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRT37 rs9910204	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	1	1	1	--	--	--	--	--	--
KRT40 rs9908304	--	0	0	0	0	--	--	--	0	--	0	--	--	0	--	--	0	--	--	--	--	0	--	--	--	--	--	--
BLMH rs1050565	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRT81 rs2071588	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	--	1	1	1	1	1	1	--	1	1
GSTP1 rs1695	--	--	--	--	--	--	--	--	--	0,1	1	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRT83 rs2852464	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	--	0	0	--	0	--
NEU2 rs2233391	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRT32 rs2071563	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DSC3 rs276937	--	0	--	--	--	--	--	--	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRT18 rs75441140	--	--	--	--	--	--	--	--	--	--	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
PLEC rs55895668	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRT82 rs2658658	0	0	0	0	0	0	0	0	0	0	0	0	--	0	0	--	0	0	--	0	--	0	--	--	--	--	--	--
SERPINB5 rs1455555	--	--	--	--	1	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRTAP4-5 rs1497383	0	0	0	--	0	0	--	--	0	0	0	0	--	0	--	--	--	--	--	--	0	--	--	--	--	--	--	--
TCHH rs2496253	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
GSDMA rs7212938	--	--	--	--	1	--	--	1	--	--	--	1	--	0,1	--	--	--	--	--	1	--	--	--	--	--	--	--	--
KRT35 rs2071601	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	--	--	0,1	0,1	0,1	--	--	--	--	--	--

Table S-4.4 (cont'd)

Distance from Root End (in)	0	0	0	0	0	0	0	1	1	1	1	1	1	1	4	5	5	5	5	5	6	6	6	6	6	6	7	7
Gene SNP	**1-R.1.1**	**1-R.1.2**	**1-R.2.1**	**1-R.2.2**	**1-R.3.1**	**1-R.3.2**	**1-R.4.2**	**1-PR.1.1**	**1-PR.1.2**	**1-PR.2.1**	**1-PR.2.2**	**1-PR.3.1**	**1-PR.3.2**	**1-PR.4.2**	**1-PD.4.2**	**1-PD.1.1**	**1-PD.1.2**	**1-PD.3.1**	**1-PD.3.2**	**1-D.4.2**	**1-D.1.1**	**1-D.1.2**	**1-PD.2.1**	**1-PD.2.2**	**1-D.3.1**	**1-D.3.2**	**1-D.2.1**	**1-D.2.2**
KRTAP4-7 rs11655310	0,1	0,1	0,1	--	0,1	0,1	--	0,1	0,1	0,1	1	0,1	--	0,1	--	0,1	--	--	--	--	0,1	--	--	--	--	--	--	--
KRT79 rs2638497	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRT75 rs298104	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRTAP10-10 rs4818950	0	0	0	--	0	--	--	--	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRTAP10-9 rs9980129	0,1	0,1	--	--	--	--	--	--	1	--	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRTAP4-9 rs7207685	0,1	0,1	0,1	--	0,1	0	--	--	0,1	0,1	0,1	0,1	--	--	--	0,1	--	--	--	--	0,1	--	--	--	--	--	--	--
KRT34 rs2239710	0,1	0,1	0	0,1	0,1	0	0	0,1	0,1	0,1	0,1	0	0	0	0	0,1	0	0,1	0	0	0,1	0	0,1	0	0	0	0,1	0
KRTAP4-11 rs9897031	0	0	0	--	0	0	0	0	0	0	--	0	0	0	--	0	--	--	--	--	0	--	--	--	--	--	--	--
KRTAP9-4 rs2191379	0,1	0,1	0,1	0,1	0,1	0,1	--	0,1	0,1	--	--	0,1	0,1	0,1	--	--	--	--	--	--	1	--	--	--	--	--	--	--
KRTAP10-11 rs462007	--	0	--	1	--	--	--	--	1	1	1	--	--	--	--	--	--	--	--	--	0	--	--	--	--	--	--	--
TGM3 rs214803	1	1	1	1	1	1	--	1	--	1	1	1	1	1	--	--	1	1	--	--	--	1	--	1	--	--	--	--
KRT13 rs9891361	1	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRTAP10-3 rs233252	1	1	--	1	1	1	--	--	1	1	1	--	--	--	--	1	1	--	--	--	1	1	--	--	--	--	--	--
KRTAP4-1 rs398825	1	--	1	1	1	1	1	1	1	--	--	--	1	1	--	--	--	--	--	1	--	--	--	--	--	--	--	--
HEXB rs820878	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRT15 rs897420	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
GLTPD2 rs35910358	1	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
CRAT rs3118635	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
FAM83H rs9969600	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Table S-4.5. Comprehensive GVP profiles for single hair samples from Individual 2, arranged by increasing segment distance from the root end in inches. Sample codes are denoted x-y.z.a, where x is the individual code, y is the hair segment (R: root, PR: proximal-to-root, PD: proximal-to-distal, D:distal), z is the sample replicate number, and a indicates the co-fractionation step (1: before, 2: after). 0 and 1 indicate detection of the major and minor GVP, respectively, and '--' indicates the non-detection of either variant.

Distance from Root End (in)	0	0	0	0	0	0	0	1	1	1	1	1	1	1	7	7	8	8	12	12	12	12	12	13	13	13	13	13
Gene SNP	**2-R.1.1**	**2-R.1.2**	**2-R.2.1**	**2-R.2.2**	**2-R.3.1**	**2-R.3.2**	**2-R.4.2**	**2-PR.1.1**	**2-PR.1.2**	**2-PR.2.1**	**2-PR.2.2**	**2-PR.3.1**	**2-PR.3.2**	**2-PR.4.2**	**2-PD.3.1**	**2-PD.3.2**	**2-D.3.1**	**2-D.3.2**	**2-PD.1.1**	**2-PD.1.2**	**2-PD.2.1**	**2-PD.2.2**	**2-PD.4.2**	**2-D.1.1**	**2-D.1.2**	**2-D.2.1**	**2-D.2.2**	**2-D.4.2**
KRTAP1-3 rs751575431	--	--	0	0	--	--	--	--	--	0	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
FCHSD1 rs202048778	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRTAP4-3 rs190711711	--	--	0	0	0	0	0	--	--	0	0	0	0	0	--	--	0	--	--	--	--	--	--	--	--	--	--	--
AHNAK rs142407818	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
PKP1 rs147328328	0	--	--	--	--	0	0	--	0	0	0	0	0	--	--	--	0	0	0	--	0	--	--	--	--	--	--	--
TYRP1 rs61752937	--	--	--	--	--	--	--	--	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRTAP1-1 rs748281420	--	--	--	--	0	--	--	--	--	--	--	--	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
TUBA3C rs36215077	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
TRIM29 rs11604169	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRT31 rs6503627	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRTAP16-1 rs72828116	--	--	0	0	0	--	0	--	0	0	0	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
FAM26D rs12660180	--	--	0	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRT33B rs71373411	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	--	0	0
KRTAP10-6 rs62220887	--	--	--	--	--	1	1	--	--	--	--	--	--	1	--	--	--	1	--	--	--	--	--	--	--	--	--	--
PPL rs2075639	0	--	0	--	--	--	--	--	--	0	--	--	--	--	--	--	--	--	--	--	--	--	--	0	--	--	--	--
KRTAP13-2 rs3804010	--	--	1	1	--	--	--	--	--	1	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
PABPC1 rs62513924	--	--	--	--	--	--	--	--	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRTAP9-3 rs112082369	--	--	0	--	0	0	0	--	--	0	--	0	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRTAP3-2 rs3829598	--	--	0	0	0,1	0	0	--	0	0	0	0,1	0	0	--	--	0	0	--	--	--	--	--	--	--	--	--	--

Table S-4.5 (cont'd)

Distance from Root End (in)	0	0	0	0	0	0	0	1	1	1	1	1	1	1	7	7	8	8	12	12	12	12	12	13	13	13	13	13
Gene SNP	**2-R.1.1**	**2-R.1.2**	**2-R.2.1**	**2-R.2.2**	**2-R.3.1**	**2-R.3.2**	**2-R.4.2**	**2-PR.1.1**	**2-PR.1.2**	**2-PR.2.1**	**2-PR.2.2**	**2-PR.3.1**	**2-PR.3.2**	**2-PR.4.2**	**2-PD.3.1**	**2-PD.3.2**	**2-D.3.1**	**2-D.3.2**	**2-PD.1.1**	**2-PD.1.2**	**2-PD.2.1**	**2-PD.2.2**	**2-PD.4.2**	**2-D.1.1**	**2-D.1.2**	**2-D.2.1**	**2-D.2.2**	**2-D.4.2**
KRTAP11-1 rs71321355	--	--	--	--	0,1	0,1	0,1	--	--	--	--	0,1	0,1	1	--	--	0	0,1	--	--	--	--	--	--	--	--	--	--
KRT4 rs7959052	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRT72 rs11170183	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRT33A rs12937519	1	1	1	1	1	1	1	1	0,1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	--	1	1
KRT84 rs1613931	--	--	1	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	1	--	--	--	--	--	--	--
KRT37 rs9910204	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	0	0	--	--	--	--	--	--	--	--	--	--
KRT40 rs9908304	--	--	--	0	0	--	0	--	0	--	0	--	--	--	--	--	0	--	--	--	--	--	--	--	--	--	--	--
BLMH rs1050565	--	--	--	--	--	--	--	--	--	--	--	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRT81 rs2071588	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
GSTP1 rs1695	--	--	0	--	--	--	--	--	0	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	0	--
KRT83 rs2852464	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0	0,1	0,1	0,1	0,1	0	0,1	0	0	0,1	0,1	0	--	0	0,1	--	--	0	--	0,1	--
NEU2 rs2233391	--	--	--	--	--	--	--	--	--	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRT32 rs2071563	0,1	0,1	0,1	0,1	0	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0	0,1	0,1	0	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	--	0,1	0
DSC3 rs276937	--	--	--	--	--	--	--	--	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRT18 rs75441140	--	--	--	--	--	--	--	--	0	--	--	--	--	--	--	--	--	--	0	--	--	0	--	--	--	--	--	--
PLEC rs55895668	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRT82 rs2658658	--	--	1	0,1	0,1	0,1	0,1	--	0,1	0,1	--	0,1	0,1	0,1	--	1	0,1	0,1	--	--	--	--	--	--	--	--	--	--
SERPINB5 rs1455555	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRTAP4-5 rs1497383	--	--	0	--	0,1	--	--	--	--	0,1	--	0,1	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
TCHH rs2496253	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	0	--	--	--	--	--	--	--	--	--
GSDMA rs7212938	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRT35 rs2071601	--	--	0	0	0	0	0	--	0	0	0	0	0	0	--	--	0	0	--	--	--	--	--	--	--	--	--	--

Table S-4.5 (cont'd)

Distance from Root End (in)	0	0	0	0	0	0	0	1	1	1	1	1	1	1	7	7	8	8	12	12	12	12	12	13	13	13	13	13
Gene SNP	**2-R.1.1**	**2-R.1.2**	**2-R.2.1**	**2-R.2.2**	**2-R.3.1**	**2-R.3.2**	**2-R.4.2**	**2-PR.1.1**	**2-PR.1.2**	**2-PR.2.1**	**2-PR.2.2**	**2-PR.3.1**	**2-PR.3.2**	**2-PR.4.2**	**2-PD.3.1**	**2-PD.3.2**	**2-D.3.1**	**2-D.3.2**	**2-PD.1.1**	**2-PD.1.2**	**2-PD.2.1**	**2-PD.2.2**	**2-PD.4.2**	**2-D.1.1**	**2-D.1.2**	**2-D.2.1**	**2-D.2.2**	**2-D.4.2**
KRTAP4-7 rs11655310	--	--	0,1	0	0,1	--	--	--	--	0,1	--	0,1	--	--	--	--	1	--	--	--	--	--	--	--	--	--	--	--
KRT79 rs2638497	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRT75 rs298104	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	1	--	--	--	--	--	--	--
KRTAP10-10 rs4818950	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRTAP10-9 rs9980129	--	--	--	--	--	--	1	--	--	1	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRTAP4-9 rs7207685	--	--	0,1	0	0,1	0	--	--	--	0,1	0,1	0,1	0	0	--	--	0,1	--	--	--	--	--	--	--	--	--	--	--
KRT34 rs2239710	0,1	0,1	0,1	0	0	0	0	0	0,1	0,1	0	0	0	0	0	0	0,1	0	0,1	0	0,1	0,1	0	0,1	0	--	0,1	0
KRTAP4-11 rs9897031	--	--	0	--	0,1	0	0	--	--	0,1	--	0,1	0	0	--	--	0	0	--	--	--	--	--	--	--	--	--	--
KRTAP9-4 rs2191379	--	--	1	--	0,1	0,1	0,1	--	--	0	--	0,1	0,1	0	--	--	0,1	0,1	--	--	--	--	--	--	--	--	--	--
KRTAP10-11 rs462007	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
TGM3 rs214803	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRT13 rs9891361	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRTAP10-3 rs233252	--	--	--	1	--	--	--	--	--	1	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRTAP4-1 rs398825	--	--	--	--	1	1	1	--	--	--	--	1	1	1	--	--	1	1	--	--	--	--	--	--	--	--	--	--
HEXB rs820878	--	--	--	1	--	--	--	--	--	--	1	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRT15 rs897420	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
GLTPD2 rs35910358	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
CRAT rs3118635	--	--	--	--	--	--	--	--	1	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
FAM83H rs9969600	--	--	1	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Table S-4.6. Comprehensive GVP profiles for single hair samples from Individual 3, arranged by increasing segment distance from the root end in inches. Sample codes are denoted x-y.z.a, where x is the individual code, y is the hair segment (R: root, PR: proximal-to-root, PD: proximal-to-distal, D:distal), z is the sample replicate number, and a indicates the co-fractionation step (1: before, 2: after). 0 and 1 indicate detection of the major and minor GVP, respectively, and '--' indicates the non-detection of either variant.

Distance from Root End (in)	0	0	0	0	0	0	0	1	1	1	1	1	1	1	3	3	4	4	4	4	4	4	4	5	5	5	5	5
Gene SNP	**3-R.1.1**	**3-R.1.2**	**3-R.2.1**	**3-R.2.2**	**3-R.3.1**	**3-R.3.2**	**3-R.4.2**	**3-PR.1.1**	**3-PR.1.2**	**3-PR.2.1**	**3-PR.2.2**	**3-PR.3.1**	**3-PR.3.2**	**3-PR.4.2**	**3-PD.3.1**	**3-PD.3.2**	**3-PD.1.1**	**3-PD.1.2**	**3-PD.2.1**	**3-PD.2.2**	**3-D.3.1**	**3-D.3.2**	**3-PD.4.2**	**3-D.1.1**	**3-D.1.2**	**3-D.2.1**	**3-D.2.2**	**3-D.4.2**
KRTAP1-3 rs751575431	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
FCHSD1 rs202048778	--	--	--	--	--	--	--	--	--	--	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRTAP4-3 rs190711711	0	--	0	--	0	0	0	0	0	0	0	0	0	0	0	0	0	0	--	--	0	0	0	0	0	--	--	0
AHNAK rs142407818	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	1	--	--	--	--
PKP1 rs147328328	--	--	--	--	0,1	0,1	0,1	--	--	--	--	0	0,1	1	--	0,1	--	--	--	--	--	0	--	0	0	--	--	--
TYRP1 rs61752937	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	0	--	--	--
KRTAP1-1 rs748281420	--	--	0	0	0	0	0	0	--	0	--	0	0	0	0	0	--	--	--	--	0	0	0	0	0	--	--	0
TUBA3C rs36215077	--	--	--	1	1	1	1	1	--	--	--	1	1	1	1	1	1	--	--	--	1	1	1	--	1	--	1	1
TRIM29 rs11604169	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	0	--	--	--	--
KRT31 rs6503627	1	--	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	--	1	1	1	1	1	1	--	1
KRTAP16-1 rs72828116	0,1	--	0,1	0,1	0,1	0,1	0,1	0,1	0	0,1	0,1	0,1	0,1	--	0,1	0,1	0,1	0	0,1	0	0,1	0,1	--	0,1	0,1	0,1	1	1
FAM26D rs12660180	0	--	1	--	--	--	--	0	--	1	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRT33B rs71373411	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRTAP10-6 rs62220887	--	1	--	1	--	1	1	--	1	--	--	1	1	1	1	1	--	1	--	1	1	1	1	--	1	--	1	1
PPL rs2075639	--	--	--	--	--	--	--	--	--	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRTAP13-2 rs3804010	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	0	--	--
PABPC1 rs62513924	--	--	0	--	--	--	--	--	--	--	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRTAP9-3 rs112082369	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	--
KRTAP3-2 rs3829598	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table S-4.6 (cont'd)

Distance from Root End (in)	0	0	0	0	0	0	0	1	1	1	1	1	1	1	3	3	4	4	4	4	4	4	4	5	5	5	5	5
Gene SNP	**3-R.1.1**	**3-R.1.2**	**3-R.2.1**	**3-R.2.2**	**3-R.3.1**	**3-R.3.2**	**3-R.4.2**	**3-PR.1.1**	**3-PR.1.2**	**3-PR.2.1**	**3-PR.2.2**	**3-PR.3.1**	**3-PR.3.2**	**3-PR.4.2**	**3-PD.3.1**	**3-PD.3.2**	**3-PD.1.1**	**3-PD.1.2**	**3-PD.2.1**	**3-PD.2.2**	**3-D.3.1**	**3-D.3.2**	**3-PD.4.2**	**3-D.1.1**	**3-D.1.2**	**3-D.2.1**	**3-D.2.2**	**3-D.4.2**
KRTAP11-1 rs71321355	0	--	0	0	0	0	0	0	--	0	--	0	0	0	0	0	0	--	0	0	0	0	0	0	0	--	0	0
KRT4 rs7959052	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	0	--	--
KRT72 rs11170183	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRT33A rs12937519	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
KRT84 rs1613931	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRT37 rs9910204	--	--	--	--	--	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	0	0	--
KRT40 rs9908304	--	--	--	0	0	0	--	0	0	--	0	--	--	--	--	--	--	0	0	0	--	--	--	--	--	--	0	--
BLMH rs1050565	--	--	--	--	1	--	--	--	0,1	--	--	--	--	--	1	--	--	0,1	--	0,1	--	--	--	--	--	--	--	--
KRT81 rs2071588	1	--	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
GSTP1 rs1695	--	--	--	--	--	--	0	--	--	--	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	0	--	--
KRT83 rs2852464	--	--	0,1	0	0,1	0,1	0,1	0	1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	--	--	0,1	0,1	0,1	0,1	0,1	1	--	0,1
NEU2 rs2233391	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRT32 rs2071563	--	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DSC3 rs276937	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRT18 rs75441140	--	--	--	--	--	--	--	--	--	--	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
PLEC rs55895668	--	--	--	--	--	--	--	--	--	--	--	--	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRT82 rs2658658	--	1	1	1	1	1	1	--	1	1	1	1	1	1	1	1	--	1	--	1	0,1	1	1	1	1	--	1	1
SERPINB5 rs1455555	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRTAP4-5 rs1497383	0	--	0	0	0	0	--	0	--	0	--	0	0	--	0	0	0	--	0	--	0	--	0	0	--	0	--	0
TCHH rs2496253	--	--	--	--	--	--	--	--	--	--	--	--	--	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--
GSDMA rs7212938	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRT35 rs2071601	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table S-4.6 (cont'd)

Distance from Root End (in)	0	0	0	0	0	0	0	1	1	1	1	1	1	1	3	3	4	4	4	4	4	4	4	5	5	5	5	5
Gene SNP	**3-R.1.1**	**3-R.1.2**	**3-R.2.1**	**3-R.2.2**	**3-R.3.1**	**3-R.3.2**	**3-R.4.2**	**3-PR.1.1**	**3-PR.1.2**	**3-PR.2.1**	**3-PR.2.2**	**3-PR.3.1**	**3-PR.3.2**	**3-PR.4.2**	**3-PD.3.1**	**3-PD.3.2**	**3-PD.1.1**	**3-PD.1.2**	**3-PD.2.1**	**3-PD.2.2**	**3-D.3.1**	**3-D.3.2**	**3-PD.4.2**	**3-D.1.1**	**3-D.1.2**	**3-D.2.1**	**3-D.2.2**	**3-D.4.2**
KRTAP4-7 rs11655310	0,1	--	0,1	1	0,1	0,1	0,1	0,1	--	0,1	0,1	0,1	0,1	0,1	0,1	1	0,1	--	0,1	--	0,1	1	0,1	0,1	0,1	0,1	--	0
KRT79 rs2638497	--	--	--	--	--	--	--	--	1	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRT75 rs298104	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRTAP10-10 rs4818950	--	--	0,1	--	0,1	--	--	0,1	--	0,1	0,1	0	--	--	0	--	0,1	--	--	--	0	--	--	0	--	0	--	--
KRTAP10-9 rs9980129	--	--	0,1	0	--	0	0,1	0,1	--	0	0,1	--	1	1	--	1	0	--	--	--	--	1	1	0,1	0,1	--	--	1
KRTAP4-9 rs7207685	0,1	--	0,1	0	0,1	0,1	0,1	0,1	--	0,1	0,1	0,1	0	0,1	0,1	--	0,1	0,1	0,1	--	0,1	0	--	0,1	0	0,1	--	--
KRT34 rs2239710	0,1	0	0,1	0	0,1	0,1	0,1	0,1	0	0,1	0,1	0,1	0	0	0,1	0	0,1	0,1	0,1	0	0,1	0	0,1	0,1	0	0,1	0	0
KRTAP4-11 rs9897031	0,1	--	0,1	--	0	0	0	0,1	--	0,1	--	0,1	0,1	0	0,1	0	0	--	0	--	0,1	0	0	0,1	0	0	--	0
KRTAP9-4 rs2191379	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	--	0	0	0	0	0	--	0	0
KRTAP10-11 rs462007	--	1	0	1	0,1	1	1	0	1	0	--	1	1	--	--	--	0	--	0	--	--	--	--	--	1	0	1	--
TGM3 rs214803	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRT13 rs9891361	--	--	--	--	--	--	--	1	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRTAP10-3 rs233252	--	--	0,1	1	0,1	0,1	--	0,1	0,1	0,1	0,1	0,1	--	1	--	--	0	0,1	0	1	--	--	--	0	0,1	1	--	--
KRTAP4-1 rs398825	0	0	--	--	0,1	0,1	0,1	0,1	0	--	--	0,1	0,1	0,1	0,1	0,1	0,1	--	0	--	0,1	0,1	0,1	0,1	0,1	--	--	0,1
HEXB rs820878	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
KRT15 rs897420	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	1	--	--	--	--	--	--	--	--	--	--	--
GLTPD2 rs35910358	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
CRAT rs3118635	--	1	--	--	--	--	1	--	--	--	--	--	--	1	--	--	--	--	--	--	--	--	--	--	--	--	--	--
FAM83H rs9969600	--	--	--	--	--	--	--	--	--	--	1	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

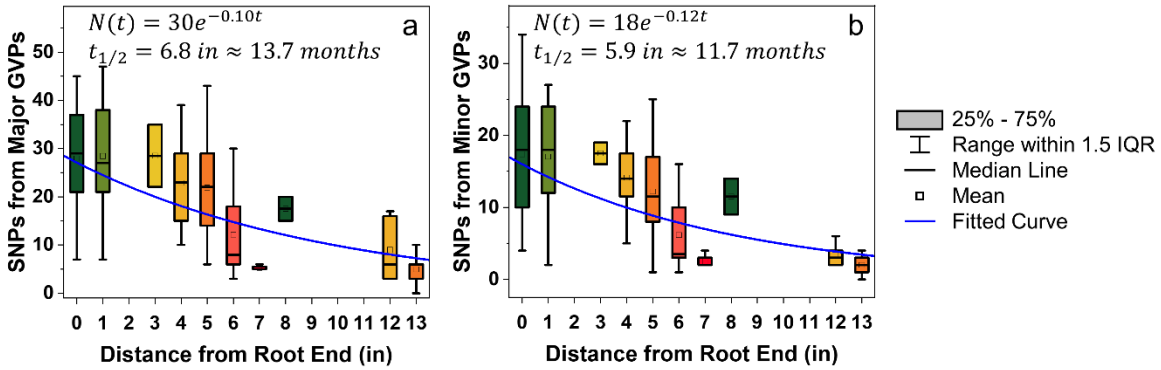


Figure S-4.2. Number of SNPs from (a) major and (b) minor GVPs identified from each one-inch single hair sample (n = 84, from 48 hair samples from 3 individuals, with 36 analyzed before and after peptide/DNA co-fractionation), with increasing hair age. Non-linear least-squares curve fitting was performed and half-life conversion from distance to time (in months) assumes a hair growth rate of 0.5-in per month. The fitted curve is plotted in blue. Not surprisingly, GVP identification of both major and minor variants decays with hair age at a rate similar to that of peptide identification.

REFERENCES

REFERENCES

1. Tang, W.; Zhang, S. G.; Zhang, J. K.; Chen, S.; Zhu, H.; Ge, S. R., Ageing Effects on the Diameter, Nanomechanical Properties and Tactile Perception of Human Hair. *International Journal of Cosmetic Science* **2016,** *38* (2), 155-163.

2. Commo, S.; Gaillard, O.; Bernard, B. A., Human Hair Greying Is Linked to a Specific Depletion of Hair Follicle Melanocytes Affecting Both the Bulb and the Outer Root Sheath. *British Journal of Dermatology* **2004,** *150* (3), 435-443.

3. Kim, S. N.; Lee, S. Y.; Choi, M. H.; Joo, K. M.; Kim, S. H.; Koh, J. S.; Park, W. S., Characteristic Features of Ageing in Korean Women's Hair and Scalp. *British Journal of Dermatology* **2013,** *168* (6), 1215-1223.

4. Jeong, K. H.; Kim, K. S.; Lee, G. J.; Choi, S. J.; Jeong, T. J.; Shin, M.-K.; Park, H. K.; Sim, W. Y.; Lee, M.-H., Investigation of Aging Effects in Human Hair Using Atomic Force Microscopy. *Skin Research and Technology* **2011,** *17* (1), 63-68.

5. Zhang, Y.; Alsop, R. J.; Soomro, A.; Yang, F.-C.; Rheinstädter, M. C., Effect of Shampoo, Conditioner and Permanent Waving on the Molecular Structure of Human Hair. *PeerJ* **2015,** *3*, e1296.

6. McKittrick, J.; Chen, P. Y.; Bodde, S. G.; Yang, W.; Novitskaya, E. E.; Meyers, M. A., The Structure, Functions, and Mechanical Properties of Keratin. *JOM* **2012,** *64* (4), 449-468.

7. Yang, F.-C.; Zhang, Y.; Rheinstädter, M. C., The Structure of People's Hair. *PeerJ* **2014,** *2*, e619-e619.

8. Giesen, M.; Gruedl, S.; Holtkoetter, O.; Fuhrmann, G.; Koerner, A.; Petersohn, D., Ageing Processes Influence Keratin and KAP Expression in Human Hair Follicles. *Experimental Dermatology* **2011,** *20* (9), 759-761.

9. Waki, M. L.; Onoue, K.; Takahashi, T.; Goto, K.; Saito, Y.; Inami, K.; Makita, I.; Angata, Y.; Suzuki, T.; Yamashita, M.; Sato, N.; Nakamura, S.; Yuki, D.; Sugiura, Y.; Zaima, N.; Goto-Inoue, N.; Hayasaka, T.; Shimomura, Y.; Setou, M., Investigation by Imaging Mass Spectrometry of Biomarker Candidates for Aging in the Hair Cortex. *PloS ONE* **2011,** *6* (10), e26721-e26721.

10. Kuzuhara, A.; Fujiwara, N.; Hori, T., Analysis of Internal Structure Changes in Black Human Hair Keratin Fibers with Aging Using Raman Spectroscopy. *Biopolymers* **2007,** *87* (2-3), 134-140.

11. Thibaut, S.; De Becker, E.; Bernard, B. A.; Huart, M.; Fiat, F.; Baghdadli, N.; Luengo, G. S.; Leroy, F.; Angevin, P.; Kermoal, A. M.; Muller, S.; Peron, M.; Provot, G.; Kravtchenko, S.; Saint-Léger, D.; Desbois, G.; Gauchet, L.; Nowbuth, K.; Galliano, A.;

Kempf, J. Y.; Silberzan, I., Chronological Ageing of Human Hair Keratin Fibres. *International Journal of Cosmetic Science* **2010,** *32* (6), 422-434.

12. Miyazawa, N.; Uematsu, T., Analysis of Ofloxacin in Hair as a Measure of Hair Growth and as a Time Marker for Hair Analysis. *Therapeutic Drug Monitoring* **1992,** *14* (6), 525-528.

13. Rogers, M. A.; Langbein, L.; Praetzel-Wunder, S.; Giehl, K., Characterization and Expression Analysis of the Hair Keratin Associated Protein KAP26.1. *British Journal of Dermatology* **2008,** *159* (3), 725-729.

14. Brandhagen, D. M.; Loreille, O.; Irwin, A. J., Fragmented Nuclear DNA Is the Predominant Genetic Material in Human Hair Shafts. *Genes* **2018,** *9* (12).

15. Melton, T.; Dimick, G.; Higgins, B.; Lindstrom, L.; Nelson, K., Forensic Mitochondrial DNA Analysis of 691 Casework Hairs. *Journal of Forensic Sciences* **2005,** *50* (1), JFS2004230-8.

16. Catlin, L. A.; Chou, R. M.; Goecker, Z. C.; Mullins, L. A.; Silva, D.; Spurbeck, R. R.; Parker, G. J.; Bartling, C. M., Demonstration of a Mitochondrial DNA-Compatible Workflow for Genetically Variant Peptide Identification from Human Hair Samples. *Forensic Science International: Genetics* **2019,** *43*, 102148.

17. Chu, F.; Mason, K. E.; Anex, D. S.; Jones, A. D.; Hart, B. R., Hair Proteome Variation at Different Body Locations on Genetically Variant Peptide Detection for Protein-Based Human Identification. *Scientific Reports* **2019,** *9* (1), 7641.

18. Loussouarn, G.; El Rawadi, C.; Genain, G., Diversity of Hair Growth Profiles. *International Journal of Dermatology* **2005,** *44* (s1), 6-9.

19. Bengtsson, C. F.; Olsen, M. E.; Brandt, L. Ø.; Bertelsen, M. F.; Willerslev, E.; Tobin, D. J.; Wilson, A. S.; Gilbert, M. T. P., DNA from Keratinous Tissue. Part I: Hair and Nail. *Annals of Anatomy - Anatomischer Anzeiger* **2012,** *194* (1), 17-25.

20. Morioka, K., A Guide to Hair Follicle Analysis by Transmission Electron Microscopy: Technique and Practice. *Experimental Dermatology* **2009,** *18* (7), 577-582.

21. Cruz, C. F.; Azoia, N. G.; Matamá, T.; Cavaco-Paulo, A., Peptide—Protein Interactions Within Human Hair Keratins. *International Journal of Biological Macromolecules* **2017,** *101*, 805-814.

22. Shimomura, Y.; Aoki, N.; Ito, M.; Rogers, M. A.; Langbein, L.; Schweizer, J., Characterization of Human Keratin-Associated Protein 1 Family Members. *Journal of Investigative Dermatology Symposium Proceedings* **2003,** *8* (1), 96-99.

23. Moll, R.; Divo, M.; Langbein, L., The Human Keratins: Biology and Pathology. *Histochemistry and Cell Biology* **2008,** *129* (6), 705.

24. Rogers, M. A.; Langbein, L.; Winter, H.; Ehmann, C.; Praetzel, S.; Korn, B.; Schweizer, J., Characterization of a Cluster of Human High/Ultrahigh Sulfur Keratin-Associated Protein Genes Embedded in the Type I Keratin Gene Domain on Chromosome 17q12-21. *Journal of Biological Chemistry* **2001,** *276* (22), 19440-19451.

25. Rogers, M. A.; Winter, H.; Beckmann, I.; Schweizer, J.; Langbein, L.; Praetzel, S., Hair Keratin Associated Proteins: Characterization of a Second High Sulfur KAP Gene Domain on Human Chromosome 21. *Journal of Investigative Dermatology* **2004,** *122* (1), 147-158.

26. Rogers, M. A.; Langbein, L.; Winter, H.; Ehmann, C.; Praetzel, S.; Schweizer, J., Characterization of a First Domain of Human High Glycine-Tyrosine and High Sulfur Keratin-Associated Protein (KAP) Genes on Chromosome 21q22.1. *Journal of Biological Chemistry* **2002,** *277* (50), 48993-49002.

27. Shimomura, Y.; Aoki, N.; Ito, M.; Rogers, M. A.; Langbein, L.; Schweizer, J., hKAP1.6 and hKAP1.7, Two Novel Human High Sulfur Keratin-Associated Proteins Are Expressed in the Hair Follicle Cortex. *Journal of Investigative Dermatology* **2002,** *118* (2), 226-231.

28. Shimomura, Y.; Ito, M., Human Hair Keratin-Associated Proteins. *Journal of Investigative Dermatology Symposium Proceedings* **2005,** *10* (3), 230-233.

29. Rogers, M. A.; Schweizer, J., Human KAP Genes, Only the Half of It? Extensive Size Polymorphisms in Hair Keratin-Associated Protein Genes. *Journal of Investigative Dermatology* **2005,** *124* (6), vii-ix.

30. Kozlowski, L. P., IPC – Isoelectric Point Calculator. *Biology Direct* **2016,** *11* (1), 55.

31. Parson, W.; Gusmão, L.; Hares, D. R.; Irwin, J. A.; Mayr, W. R.; Morling, N.; Pokorak, E.; Prinz, M.; Salas, A.; Schneider, P. M.; Parsons, T. J., DNA Commission of the International Society for Forensic Genetics: Revised and Extended Guidelines for Mitochondrial DNA Typing. *Forensic Science International: Genetics* **2014,** *13*, 134-142.

32. Melton, T., Mitochondrial DNA Heteroplasmy. *Forensic Science Review* **2004,** *16* (1), 1-20.

33. Li, M.; Schröder, R.; Ni, S.; Madea, B.; Stoneking, M., Extensive Tissue-Related and Allele-Related MtDNA Heteroplasmy Suggests Positive Selection for Somatic Mutations. *Proceedings of the National Academy of Sciences* **2015,** *112* (8), 2491-2496.

34. Barrett, A.; Arbeithuber, B.; Zaidi, A.; Wilton, P.; Paul, I. M.; Nielsen, R.; Makova, K. D., Pronounced Somatic Bottleneck in Mitochondrial DNA of Human Hair. *Philosophical Transactions of the Royal Society B: Biological Sciences* **2020,** *375* (1790), 20190175.

35. Salas, A.; Lareu, M. V.; Carracedo, A., Heteroplasmy in MtDNA and the Weight of Evidence in Forensic MtDNA Analysis: A Case Report. *International Journal of Legal Medicine* **2001,** *114* (3), 186-190.

36. Lott, M. T.; Leipzig, J. N.; Derbeneva, O.; Xie, H. M.; Chalkia, D.; Sarmady, M.; Procaccio, V.; Wallace, D. C., MtDNA Variation and Analysis Using Mitomap and Mitomaster. *Current Protocols in Bioinformatics* **2013,** *44* (123), 1.23.1-1.23.26.

37. Rogers, M. A.; Winter, H.; Langbein, L.; Wollschläger, A.; Praetzel-Wunder, S.; Jave-Suarez, L. F.; Schweizer, J., Characterization of Human KAP24.1, a Cuticular Hair Keratin-Associated Protein with Unusual Amino-Acid Composition and Repeat Structure. *Journal of Investigative Dermatology* **2007,** *127* (5), 1197-1204.

38. Hudson, W. H.; Ortlund, E. A., The Structure, Function and Evolution of Proteins that Bind DNA and RNA. *Nature Reviews Molecular Cell Biology* **2014,** *15* (11), 749-760.

39. Boutet, E.; Liberherr, D.; Tognolli, M.; Schneider, M.; Bairoch, A., UniProtKB/Swiss-Prot. *Methods in Molecular Biology* **2007,** *406*, 89-112.

CHAPTER 5: Characterization of Mechanical Hair Damage and Effects on Genetically Variant Peptide Identification

Foreword

This chapter describes work that has been adapted from both a published paper[1] and a submitted manuscript (Chu et al., submitted). Contributions from others to the conduct of the experiments described in this chapter are as follows, in no particular order: Z. Dai assisted with acquisition of scanning electron micrographs, P. H. Paul provided single nucleotide variant lists and individualized mutated protein FASTA files, K. E. Mason acquired the mass spectrometry data, and D. S. Anex recovered the hair samples post-blast.

5.1 Introduction

Previous chapters have examined effects of intrinsic hair chemistry, including variation in the hair proteome by body location and hair age, directly enabled by the development and optimization of single hair analysis. However, effects of environmental exposures on the hair proteome for protein-based human identification have not been studied. In particular, hair protein chemistry in damaged single hairs recovered after an explosion has not yet been characterized.

Evidence recovery where explosions have occurred is challenging. In hair, only minimal nuclear DNA remains intact even in the absence of harsh conditions[2, 3]; it can reasonably be expected that recovered hair evidence, a matrix that is sufficiently robust to survive most explosive blasts, contains even less intact nuclear DNA for profiling. Such limitations may further reduce chances of obtaining full short-tandem repeat profiles, making the technique less reliable for identification. In contrast, protein content in hair is likely to survive, though hair proteins may sustain damage owing to heat exposure from explosive blasts, which in turn, may compromise detection of GVPs. There are knowledge gaps regarding the relationships between

the extent of damage to hair protein chemistry from an explosive blast and the success of hair protein-based human identification.

Hair damage has primarily been assessed at the morphological level, but these works often describe qualitative observations. Health disorders such as lamellar ichthyosis and trichothiodystrophy are commonly diagnosed by examining morphology, predominantly via light microscopy[4, 5]. Previous studies have also described morphological hair damage from physical and chemical weathering, including applications of detergent, brushing, and bleach, which cause holes, cracks, cuticle lift-ups, and exposure of the cortex with increasing severity in morphological damage[6]. Further, damage such as axial cracks can be observed in micrographs from hair subjected to hot air-drying, even following exposure during drying temperature as low as 61 °C[7]. Exposure to higher temperatures results in the hair matrix transitioning to a strengthened state with an increase in fiber crystallinity, which occurs when heated to temperatures between 130 °C and 170 °C[8, 9], followed by denaturation of the crystalline phase at 233 °C[8]. With longer heating times, such as 1 h, destabilization of the α-helical regions of wool, a matrix similar in structure to human hair, as measured by rate of fiber extension and contraction times, has been observed[9].

Few studies have probed the chemical changes underlying thermal hair damage, much less damage induced by an explosion, instead characterizing damage from other sources, including chemical oxidative damage from bleach[10-13] and photodegradation from ultraviolet radiation[14-16]. Notably, McMullen and Jachowitz investigated effects of heating with a curling iron on tryptophan decomposition in hair using fluorescence spectroscopy, showing degradation of not only tryptophan, but also its oxidation products kynurenine and N-formylkynurenine with more severe thermal damage[17]. In addition to tryptophan oxidation, decompositions of cysteine

and tyrosine residues in keratins are thought to contribute to hair fiber yellowing, as these processes are proposed pathways for formation of yellow-colored chromophores[18]. Richena and Rezende observed an increase in conversion of cysteine to cysteine sulfonic acid in hair damaged after irradiation with UV light over 500 h, although the spectroscopic analysis was restricted to the hair cuticle[14]. Furthermore, compound effects of solar radiation and heat exposure on hair were examined specifically in aromatic and sulfur-containing amino acids, but localization of degradation to the hair cuticle or cortex was not established and effects of heat treatments alone were not assessed[19]. Recovered hair specimens from an explosive blast may sustain damage similar to that from heat exposure, as elevated temperatures may be attained in an explosion, yet little is known about the details of how hair protein chemistry may be affected.

This research aims to examine effects of an explosive blast, specifically via morphological and chemical analysis, on the hair proteome and GVP identification, evaluate morphological assessments of hair damage as a predictor of proteomic profiling success, and quantify the differentiative potential of individuals using recovered single hairs. Through comparison between exploded and undamaged control hair specimens, minimal effect of an explosive blast on hair protein chemistry and GVP identification were found, with proteomic profiling success independent from morphological hair damage and similar differentiative potential of individuals regardless of hair damage. While proteins localized external to the cortex and medulla can serve as markers for hair cuticular damage under explosion conditions, GVP identification remains independent of affected proteins, which demonstrates a path forward for application of GVP technology to recovered hair evidence from an explosion for forensic identification.

5.2 Experimental

5.2.1 Hair Sampling and Collection

Scalp hair specimens were collected from individuals under approval by the Institutional Review Board at Lawrence Livermore National Laboratory (Protocol ID# 15-008 and 16-002) and in accordance with the Declaration of Helsinki. Written informed consent for specimen collection and analysis was obtained prior to collection. Hair fibers from three individuals were used in this study. After assembly of an experimental explosive device using commercial materials as part of a training exercise hosted by the Bureau of Alcohol, Tobacco, Firearms, and Explosives National Center for Explosives Training and Research at the Redstone Arsenal in Huntsville, AL, hairs (< 5 cm) were taped onto the internal and external regions of the device. The device was then detonated in a spherical total containment vessel with a diameter of 48 in (1.2 m) using 2 inches (5 cm) of dynamite. Remnants of the device and hair fibers (exploded hairs were designated with sample identifier Ex) were collected from the total containment vessel after the explosion and stored in the dark at room temperature. Undamaged control hairs were classified as either travel blanks (identifier Tr) or control blanks (identifier Ctrl); the former hair fibers were stored in an envelope and brought to the site of the explosion but were not exposed to an explosive blast, and the latter remained stored in the laboratory.

5.2.2 Scanning Electron Microscopic Analysis

Recovered hair fibers were isolated and segmented; one inch (~2.5 cm) each was allotted for protein analysis described in Section 5.2.3, and the remainder (approximately 1 cm) was used for scanning electron microscopic (SEM) analysis. Three exploded and two control hairs were randomly selected for analysis and image acquisition. Each segmented hair was fixed onto a stub prior to carbon coating; under vacuum, a carbon layer of approximately 10 nm was deposited

onto each specimen after heating for approximately 5 s. Secondary electron images were acquired along the length of each hair fiber using an Inspect F (FEI Company, Hillsboro, OR) scanning electron microscope, at an acceleration voltage of 5 kV, a dwell time of 3 µs, and a working distance of 7 mm over a range of magnifications. Brightness and contrast were automatically adjusted for each image. In total, 58 digital SEM images (8-bit) were acquired from five hair segments, and all were then processed using ImageJ 1.52k software in replicates of n = 5.

5.2.3 Hair Sample Preparation for Mass Spectrometry Analysis

Proteins were extracted from single one-inch hairs and alkylated as described in Chapter 3, Section 3.2.1 of this dissertation. To remove detergent sodium dodecanoate and concentrate proteins, an overnight protein precipitation with cold acetone was performed. A 4:1 ratio of cold acetone to aqueous solution was allowed to incubate overnight at -20 °C. After centrifugation at 15,000 × *g* for 15 min at RT, supernatants containing detergent were separated from the protein pellets and discarded. Pellets were washed with 400 µL of cold acetone and supernatant was once again discarded after centrifugation. Prior to protein digestion, pellets were resuspended in 50 µL of 50 mM dithiothreitol, 50 mM ammonium bicarbonate, and 0.01% (w/v) ProteaseMAX™ (Promega, Madison, WI) and incubated on a shaker for 1 h at RT. Digestion using 2 µL of 1 µg/µL TPCK-treated trypsin was performed, with overnight incubation at RT accompanied by magnetic stirring. Protein digests were then filtered for particulates using centrifugal filter tubes (PVDF, 0.1 µm; MilliporeSigma, Burlington, MA).

5.2.4 Liquid Chromatography-Tandem Mass Spectrometry Analysis and Protein and Peptide Identification

Filtered protein digests were analyzed on an EASY-nLC 1200 system coupled to a Q Exactive Plus Orbitrap mass spectrometer (Thermo Fisher Scientific Inc., Waltham, MA). Injection volumes of 0.5 µL were loaded onto an Acclaim™ PepMap™ 100 C18 trap (75 µm × 20 mm, 3 µm particle size), washed, and separated on an Easy-Spray™ C18 analytical column (50 µm × 150 mm, 2 µm particle size). Separations were performed at a flow rate of 300 nL/min using mobile phases A (0.1% formic acid in water) and B (0.1% formic acid in 90% acetonitrile/10% water) over a 107-min gradient: 2 to 3% B in 1 min, 3 to 11% B in 75 min, 11 to 39% B in 15 min, ramped to 100% B in 1 min, and held at 100% B for 15 min. Positive mode nano-electrospray ionization was achieved at a voltage of 1.9 kV. Full MS survey scans were acquired at a resolution of 70,000 over a scan range between *m/z* 380 and 1800, with a maximum ion accumulation time of 30 ms. Data-dependent MS/MS scans were triggered for the 10 most abundant survey scan ions at an intensity threshold of 3.3×10^4 and acquired at a resolution of 17,500, with a maximum ion accumulation time of 60 ms, dynamic exclusion of 24 s, and an isolation window of 2 Da. HCD fragmentation was performed at a normalized collision energy setting of 27. Singly-charged species and ions with unassigned charge states were excluded from MS/MS.

Protein and peptide identifications from mass spectral data were performed using PEAKS Studio 7.5 (Bioinformatics Solutions, Ontario, Canada). Details of the process are elaborated elsewhere[20]. Briefly, *de novo* sequencing was performed to identify peptides, with a precursor mass error tolerance of 20 ppm and fragment ion tolerance of 0.05 Da, and permitting 3 missed tryptic cleavages. On both ends of each peptide, a total of 3 non-tryptic cleavages were allowed.

Cysteine carbamidomethylation was selected as a fixed modification while all other post-translational modifications, including asparagine and glutamine deamidation, and methionine oxidation, were allowed as variable modifications, with a maximum of three per peptide. Peptides with confidence scores above 15% were then matched to protein sequences from the UniProtKB SwissProt Human database (downloaded September 21, 2017) and from individualized mutated databases that contain amino acid polymorphisms. A 1% false discovery rate was applied to filter peptide-spectrum matches, and only peptides unique to one gene were retained. Non-redundant peptide lists were then exported and further filtered with a 5-ppm mass error tolerance window for genetically variant peptide identification.

5.2.5 Statistical Analysis

All statistical comparisons were performed in R (x64 version 3.4.4). Statistical significance was established at $\alpha = 0.05$. Two-sample t-tests were performed using the *stats v3.5.0* package; equal variances were not assumed. The asymptotic test[21] was used for comparison of coefficients of variation from morphological analysis metrics, using the *cvequality v0.2.0* package. Pearson product-moment correlations, Spearman's rank correlations, and one-sample t-tests of correlation coefficients were performed to determine statistical significance of the correlations from morphological and proteomic analyses using the *cor.test* function in the *stats v3.5.0* package. Training and test sets for a k-Nearest Neighbor Classification (kNN) model of the morphological analysis data were established by randomization, each comprising 50% of the dataset and contained the same number of SEM images from exploded and undamaged hairs. The model was developed using the *knn* function in the *class v7.3-15* package, with $k = 3$ nearest neighbors determined by Euclidean distance. Non-parametric two-sample Wilcoxon Rank Sum test was performed to compare intra- and interindividual GVP profile differences using the

wilcox.test function in the *stats v3.5.0* package. All plots were drawn in OriginPro 2018 (OriginLab Corp., Northampton, MA).

5.3 Results and Discussion

Using single hairs recovered post-explosion and undamaged control hairs, hair damage was characterized by SEM and by untargeted mass spectrometry. Morphological analysis via microscopy serves to establish a baseline for comparison to previous works describing thermally damaged hair fibers. Comparison to findings by proteomic analysis then allows a more complete characterization of damage to the single hair fibers. However, while there exist quantitative metrics for proteomic analysis, microscopy has relied primarily on qualitative observations. Thus, to enable a comparison of results from the two approaches, a metric must be developed for quantifying damage in recovered hair fibers via morphological analysis. Section 5.3.1 describes such a development for a quantitative and objective metric.

5.3.1 Characterization of Hair Damage via Scanning Electron Microscopy

In clinical settings, microscopic analysis can be used as a tool to assess hair damage as an indicator of health status[4, 5, 22], but analyses are performed in a qualitative manner through identification of morphological features of hair damage. Few studies have quantified the extent of hair damage based on morphological features[6, 23, 24]. Notably, Kim et al. developed a classification system with five damage grades for characterizing hair surface damage from weathering[6], which was then expanded upon to a twelve-point scale by Lee and colleagues[24]. However, the grading systems were dependent on visual scoring of scanning electron microscopic (SEM) images based on subjective evaluations of severity in the irregularity of hair cuticular structure. Microscopic analysis remains a predominantly qualitative technique via

visual assessments; little emphasis has been placed on developing more objective metrics and even less so for quantitation of hair damage severity.

Digital image analysis has been underutilized for classification of hair fibers from various microscopic methods, despite offering potential for more objective detection and comparison of image features. Of these studies, the majority concentrated on morphological features detected by light microscopy and analyzed using commercial software[25-29], even though other microscopic techniques such as atomic force microscopy (AFM) and SEM permit more extensive hair structure analysis for comparison at higher spatial resolution[30, 31]. In particular, Gurden and co-workers assessed hair structural damage from bleaching, and differentiated root and distal ends affected by chemical treatment with cuticular structure measurements such as surface roughness from AFM images[30], though the extent of damage was not graded or quantified. There is a need for development of an objective scoring system to characterize the extent of hair damage using automated analysis of higher resolution microscopic images.

5.3.1.1 Development of an Automated Image Normalization Procedure

Hair-to-hair variation and image acquisition differences obfuscate characterization and scoring of hair surface damage from microscopic images, which rely on feature detection and pairwise comparisons. Structural differences between two hair segments (e.g., width, curvature), even along the length of a hair, and automatic setting of brightness and contrast parameters for optimal SEM image acquisition make feature detection and hair segment comparison in image analysis challenging. Briefly, SEM imaging to interrogate specimen surface topography is achieved as an electron beam scans over the specimen, and interactions between primary electrons and accessible atoms from the specimen lead to emission of secondary electrons[32]. The number of secondary electrons that reach the detector manifests as pixel grayscale brightness in

an image[33]. Pixel brightness contrast within an image provides topographical information about the specimen surface, which is most accessible to the primary electrons, and is affected by, among other factors, the ease with which secondary electrons escape the surface of the specimen once formed[32]. However, detection of morphological features on the surface for image comparison may be complicated when brightness contrast varies within and between images. For example, SEM images of segments from two different hairs (Hair Samples 4 and 2), shown in Figure 5.1a-b, respectively, can display vastly different brightness levels, even within the same image along the width of the hair segment, owing to hair fiber positions and stage tilt angles. Furthermore, the tubular structure of hair creates different incident angles for the electron beam, which affects formation and detection of secondary electrons. Coupled with the orientation of the electron beam and the detector with respect to the hair fiber, detection of an abundance of secondary electrons formed from contact with hair segment edges manifests as abnormally bright bands along the edges of the hair segment that obscure image features entirely, even after carbon coating. To facilitate feature identification and enable direct comparison of different images, such artifacts must be removed or addressed.

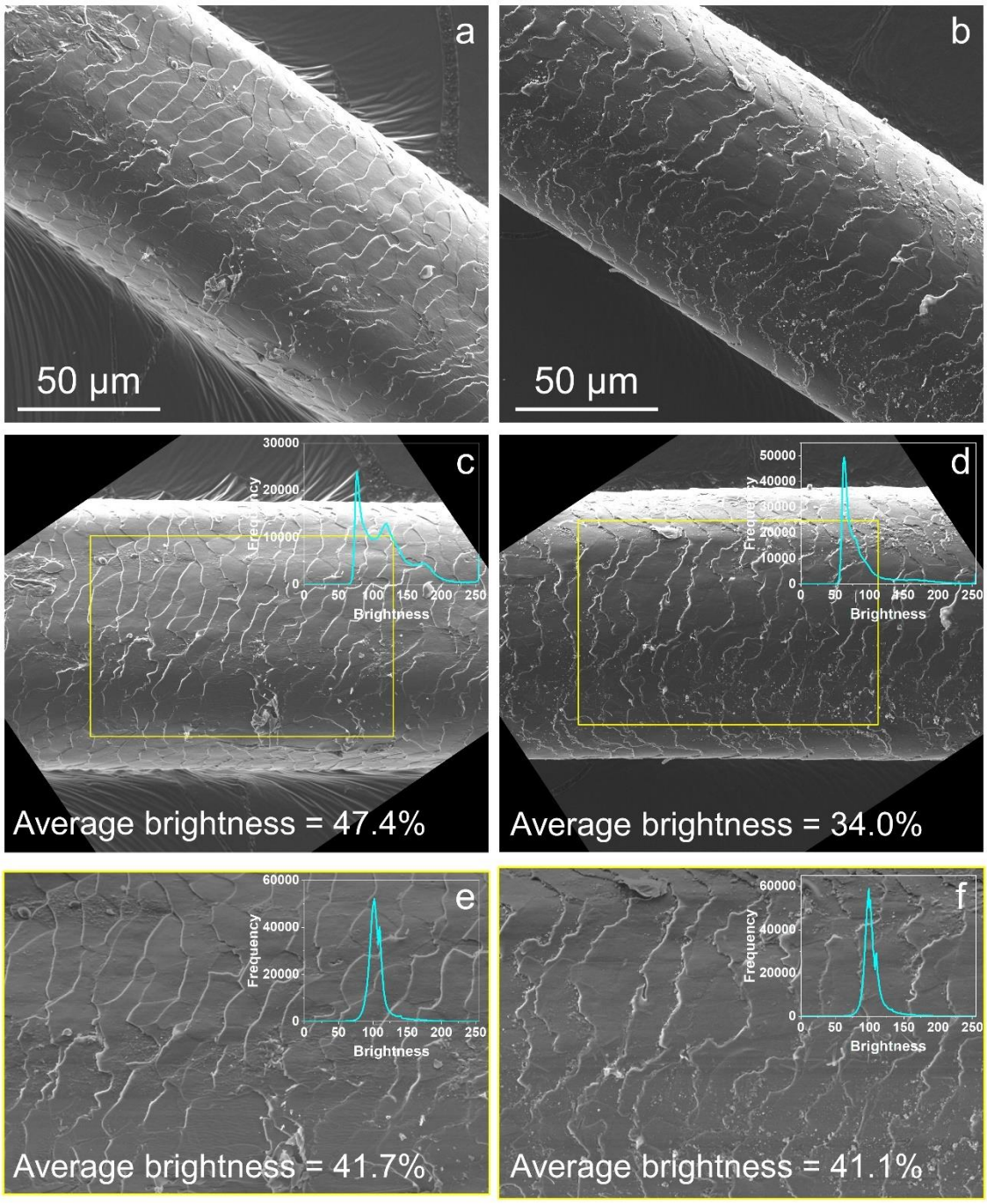


Figure 5.1. SEM images of (a) and (b) hair segments as original raw images, (c) and (d) rotated segments with selected region of interest (yellow rectangle) and image brightness histogram of region, and (e) and (f) regions of interest after normalization and corresponding image brightness histograms, from Hair Samples 4 and 2, respectively. Original images (c) and (d) show different brightness levels due to hair-to-hair variation and image acquisition differences. The described normalization procedure minimized brightness differences within and between images, as displayed in (e) and (f).

While many normalization methods have been implemented to remove image artifacts such as brightness variation, procedures used to process digital images often focus on contrast enhancement. These include variations of histogram equalization, gamma intensity correction (GIC), and wavelet-based methods for applications such as feature detection in retinal and magnetic resonance images for disease diagnosis[34, 35] and in digital images for facial recognition[36, 37]. However, these methods necessitate user inputs and parameter optimization, such as the gamma value in GIC, and are used to enhance features for detection in an image. Variations in both parameters are not conducive to comparative image analysis with a scoring system. Instead, desirable normalization procedures require minimal user-defined inputs, are computationally inexpensive, and preserve pixel brightness information in an image while reducing hair-to-hair and image acquisition variation.

This section sought to facilitate identification of microscopic features characteristic of hair damage and image comparison by developing and applying a simple and automated normalization procedure and evaluate metrics for representing hair damage for comparison to proteomic profiling results. Development and automation of image analysis for assessment of hair surface damage directly enables correlations of the effects of an explosive blast on morphological hair damage with alterations in chemical composition of hair, which is discussed in Section 5.3.3. Using open source image visualization program ImageJ[38, 39], morphological features unique to hairs subjected to explosive blast conditions were identified after normalization.

Prior to normalization, a region of interest (ROI) was computationally defined in each image to ensure that regions exhibiting abnormally high brightness on the edges of hair segments were excluded from the area processed by image analysis. The original raw image was first

rotated using a user-defined line input along the length of the hair segment so that the length of the hair segment was oriented along the horizontal axis of the ROI. Empirical evaluations of a few hair segment images showed that up to 10 μm of hair surface along the vertical axis from either edge were prone to abnormally high brightness (Figure 5.1a-b), approximately 20% of the width of hair segment. To uniformly define the ROI bounds between images yet exclude abnormally bright regions, 75% of the hair segment width and length equidistant from the image center were included in the ROI. The bounds, length, and width of the hair segment were then defined (in pixels) using a user-defined diagonal line, with coordinates (x_1, y_1) and (x_2, y_2), that spanned two corners of the segment. From the diagonal line, the upper left-hand corner coordinates (x_{lh}, y_{lh}), length, and width of the ROI were defined according to Equations 5.1 – 5.3:

$$(x_{lh}, y_{lh}) = \left(\frac{7x_{min}+x_{max}}{8}, \frac{7y_{min}+y_{max}}{8}\right), \quad \text{Eq. 5.1}$$

$$l = \frac{3}{4}(x_{max} - x_{min}), \quad \text{Eq. 5.2}$$

and

$$w = \frac{3}{4}(y_{max} - y_{min}), \quad \text{Eq. 5.3}$$

where x_{min}, x_{max}, y_{min}, and y_{max} represent the minima and maxima of x and y, respectively, extracted from the diagonal line described by coordinates (x_1, y_1) and (x_2, y_2). Centering the ROI to encompass $\frac{9}{16}$th (or 56%) of the hair segment area ($l \times w$) allows most of the segment to be included for image analysis while excluding edge regions where features are entirely obscured due to abnormal pixel brightness for reasons discussed above (Figure 5.1c-d).

After ROI definition, brightness within an image was equalized by normalizing to the average brightness per row of pixels followed by centering the average at a value of 109 and rescaling. The value of 109 was selected empirically by considering two hair segment images

and calculating the average image brightness within the ROIs from the two images and then averaging the obtained results. The resultant centering value of 109 is equivalent to 43% of the maximum brightness (from a scale ranging between 0 and 255) and represents a dark gray pixel. This pixel brightness value corresponds to undamaged regions of hair segments, which make up the majority of the pixels in the image. To preserve pixel brightness ratios with respect to the average brightness per row of pixels but ensure that average image brightness centers around 109 and pixel brightness maximizes at 255, normalization was performed for each pixel within the ROI according to Equations 5.4 and 5.5:

$$I_{i,j,norm} = \frac{I_{i,j}}{\frac{1}{l}\sum_{n=1}^{l} I_{n,j}} \cdot 109 \qquad \text{Eq. 5.4}$$

and

$$I_{i,j,norm,scale} = \begin{cases} \frac{146 \cdot (I_{i,j,norm} - 109)}{I_{j,norm,max} - 109} + 109, & I_{i,j,norm} > 109 \\ I_{i,j,norm}, & I_{i,j,norm} \leq 109 \end{cases}, \qquad \text{Eq. 5.5}$$

where $I_{i,j}$, $I_{i,j,norm}$, and $I_{i,j,norm,scale}$ represent the raw, normalized, and rescaled brightness of a pixel at image position i, j, respectively, $I_{n,j}$ is the brightness of a pixel at image position n, j from 1 to ROI length l, $I_{j,norm,max}$ is defined as the maximum normalized brightness at the *j*th row, and 146 represents the difference between maximum pixel brightness 255 and brightness value 109. Equation 5.5 is based on min-max normalization, a common score normalization approach[40]. After normalization, ROIs exhibited less variance within an image and similar average brightness values (Figure 5.1e-f).

5.3.1.2 Identification of Microscopic Features of Hair Damage

Single hairs recovered after an explosive blast sustained damage comparable to that from physical and chemical weathering, as similar morphological features were identified via scanning

electron microscopic (SEM) analysis in this study. Visual inspection of microscopic images of damaged hairs after normalization enabled identification of holes, cracks, lifting and tearing of the cuticle, and partial exposure of the cortex (Figure 5.2a-b). Images were scored based on qualitative presence or absence of features, as described in the SEM damage grade system proposed by Kim et al., where overlapped cuticles represent the lowest degree of hair surface damage (Scu 1, or Grade 1 damage assessed in SEM images of hair cuticle), apart from intact virgin hair, while the most severe hair damage (Scu 4) is characterized by the complete absence of cuticle and full exposure of cortex[6]. In particular, severe lifting of the cuticle edges in Figure 5.2b is attributed to scorching of the hair surface during the explosion, with the more intense scorching creating concavities in the edges of the cuticles along the center of the hair fiber; cuticle lifting is observed even when hair is exposed to 61 °C temperatures from hair-drying[7]. Heat from the explosion also likely stressed the hair fibers, resulting in axial cracks on the surface along the hair length as the first indication of thermal damage, likely from cortical swelling in the fiber[7, 41]. These features were also observed in hair exposed to physical and chemical weathering, such as frequent washing with detergent and exposures to bleach and UV light[6, 23]. The extent of damage in recovered hairs varied along each hair. Regions in which damage consisted only of overlapped cuticles, attributed predominantly to daily weathering, were observed in exploded hairs, although the majority of SEM images containing this damage feature belonged to control hairs (Figure 5.2d-e).

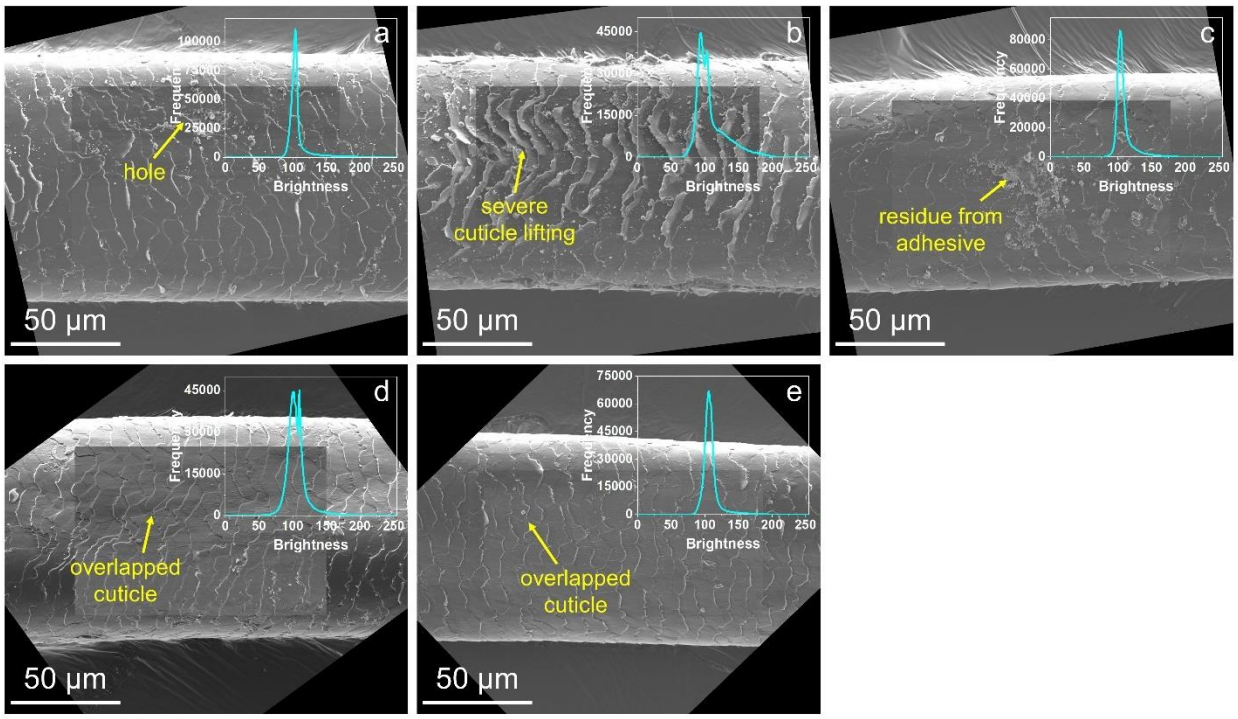


Figure 5.2. Representative rotated SEM images with overlays of normalized regions of interest and corresponding brightness histograms from Hair Samples 1 – 5, respectively. Features are labeled and denoted by yellow arrows. In addition to debris and particulates on the hair surface, features characteristic of damage from an explosion induced by an explosive device include (a) holes exposing layers of cuticle, (b) severe lifting of the cuticle edges and large cracks leading to partial exposure of cortex, and (c) localized non-specific cuticle lifting with residue from adhesive tape. Undamaged control hairs (d) and (e) predominantly display overlapped cuticles from daily weathering, illustrating substantially less severe hair surface damage compared to exploded hairs.

In addition to the above features, exploded hairs contained features not typically observed from physical and chemical weathering alone; embedded debris and particulates and cuticle lifting with adhered amorphous residue further characterized exploded hairs. Even without washing hair specimens after sample collection, control hair samples 4 and 5 were debris-free (Figure 5.2d-e), indicating that the presence of embedded particulates is characteristic of physical contact with the explosive or remnants of the device. Furthermore, amorphous residue adhered to lifted cuticles (Figure 5.2c) likely resulted during the hair fiber isolation process. Hairs

previously attached to the experimental device via adhesive tape were isolated with forceps; detachment of hair fibers led to cuticle lifting, with residual adhesive bonded to the cuticle.

Normalization ensured that undamaged regions of hair fibers remain uniform in pixel brightness (intensity scale ranging 0 – 255) as a dark gray while physical features of hair surface damage appear as clusters of brighter pixels, ranging from light gray to white. For example, cuticle lifting is characterized by a cluster of brighter pixels, bounded by white pixels along the cuticle edges as distinct from the cuticle layer underneath, which is the result of elevation differences from the hair surface (Figure 5.2b). In contrast, a depressed feature such as a hole manifests as alternating rings of light and dark pixels, as the light pixels delineate the edges of each exposed cuticle layer that is represented by darker pixels, down into the cortex (Figure 5.2a). Similarly, many microscopic features of hair surface damage identified herein are characterized by pixel brightness differences that can be further exploited in image analysis.

5.3.1.3 Evaluation of Image Parameters and Metrics for Scoring Hair Surface Damage and Image Comparison

Automated image analysis for scoring hair surface damage and image comparison requires a reliable metric that not only represents the microscopic features identified above but can also be calculated from the image. As discussed above, many features characteristic of damage manifest as brighter pixels in images, compared to the uniform dark gray of smoother undamaged regions. Thus, the potential of using pixel brightness to score hair surface damage was investigated. The first representation of pixel brightness examined was average image brightness $\overline{I_{norm}}$, defined for an ROI using the equation:

$$\overline{I_{norm}} = \frac{1}{l \cdot w} \sum_{j=1}^{w} \sum_{i=1}^{l} I_{i,j} \ , \qquad \text{Eq. 5.6}$$

where $I_{i,j}$ is the brightness of a pixel at image position i, j, and l and w represent the length and width of the ROI (in pixels), respectively.

Although the simplest representation of brightness is its average, the metric was excluded from consideration due to its role in the normalization procedure and its correlation with magnification. The normalization procedure re-centered the average row brightness at a value of 109, which corresponds to 43% of the maximum brightness. This method not only readjusted brightness within images, but also effectively equalized brightness of undamaged regions between images for direct comparison. Therefore, average image brightness cannot be used to capture brightness differences between images from microscopic features of damage. Additionally, a significant positive correlation ($r = 0.595$; $p = 8.26 \times 10^{-7}$; Figure 5.3a) between average image brightness and magnification indicates that images acquired at vastly different magnifications (between 1000X and 7000X) cannot be directly compared. Images exhibit greater average pixel brightness at higher magnifications, most likely from automatic brightness and contrast settings that were not modified when changing from low to high magnifications in the same region of the hair fiber, which result in higher overall brightness and low contrast for a smaller image area. Because of this correlation, only images acquired between 1000X and 4000X magnification were retained ($r = 0.059$; $p = 0.750$) for scoring and comparison.

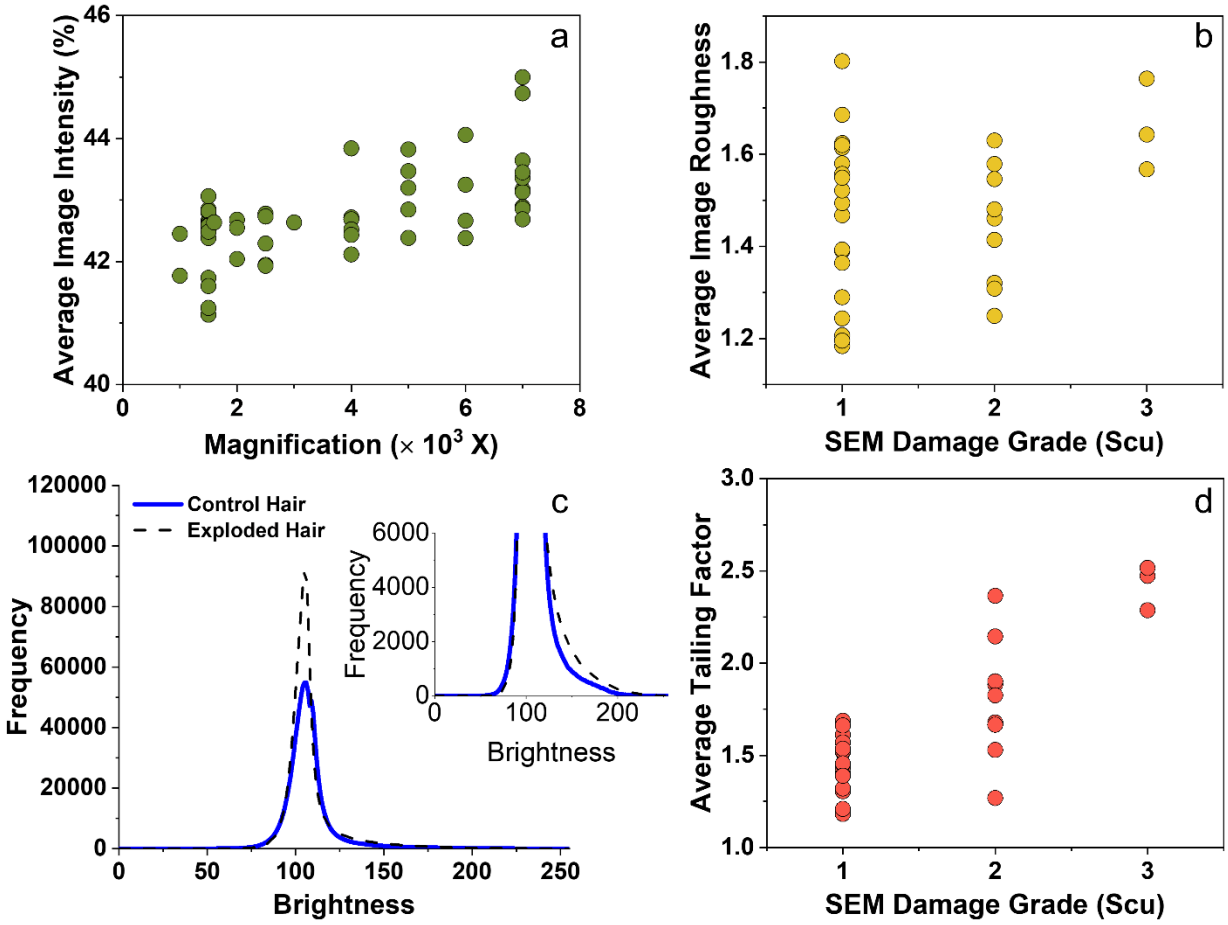


Figure 5.3. Image metrics and parameters for characterization of hair surface damage in SEM images. (a) Correlation between average image brightness and magnification after normalization. A moderate correlation (Pearson product-moment correlation (PPMC) coefficient r = 0.595, p = 8.26×10^{-7}, df = 56) between image brightness and magnification indicates that images acquired with vastly different magnifications cannot be compared without an alternative normalization scheme. Thus, only images with magnification ≤ 4000X were considered for damage scoring. (b) Correlation of average image roughness, as calculated using Equations 7 – 9, with SEM damage grade (PPMC coefficient r = 0.259, p = 0.153, df = 30). Roughness was calculated for 100 sections along the horizontal axis. As evidenced by the wide range of roughness measurements for images designated as sustaining Scu 1 damage, average image roughness does not sufficiently represent hair surface damage. (c) Average image brightness histograms for exploded and control SEM hair images with inset. Inset shows pronounced peak tailing in histograms of exploded hairs compared to control hair image brightness histograms, which can be further exploited to describe hair damage. (d) Correlation of tailing factor with SEM damage grade (PPMC coefficient r = 0.823, p = 7.31×10^{-9}, df = 30). Tailing factor, a measure of peak tailing, was calculated at 2% of the peak height maximum. Compared to image roughness, tailing factor better captures the extent of hair surface damage in SEM images.

Metrics for assessing image roughness as an indication of hair surface damage were chosen for investigation. Physical surface roughness formed by elevations and depressions from the cuticle surface creates variations in adjacent pixel brightness that deviate from the average owing to differences in secondary electron trajectories from surface to detector. Roughness was previously used in conjunction with other metrics to profile morphological damage in the cuticular structure of human hair from images acquired via atomic force microscopy (AFM) in contact mode; using these metrics and multivariate statistics, Gurden et al. reported 86% accuracy to classifying hair segments as bleached vs. untreated and from root or distal end[30]. Roughness was calculated from surface profiles over the length of the profile since AFM measures surface height, with a completely flat profile having a roughness defined as 1. But because height information is not directly obtained from SEM images, hair surface roughness determination was modified using the summation of pixel brightness differences between image sections over the length of the region of interest. Through adaptation of the distance formula for application to SEM images, average image roughness $\bar{r}$ was determined for n sections along length l of the ROI using Equations 5.7 – 5.9:

$$s = \left\lfloor \frac{l}{n} \right\rfloor, \ 1 \leq s \leq l \ , \qquad \text{Eq. 5.7}$$

$$n_{actual} = \begin{cases} n, \ \left\{\frac{l}{s}\right\} = 0 \\ \lceil n \rceil, \ \left\{\frac{l}{s}\right\} > 0 \end{cases}, \qquad \text{Eq. 5.8}$$

and

$$\bar{r} = \frac{1}{w}\sum_{j=1}^{w}\left[\frac{1}{l}\left(\left(\sum_{i=1}^{n_{actual}-1}\sqrt{\left(I_{si+1,j} - I_{s\cdot(i-1)+1,j}\right)^2 + (s)^2}\right) + \sqrt{\left(I_{l,j} - I_{s\cdot(n_{actual}-1)+1,j}\right)^2 + \left(l - s(n_{actual}-1)\right)^2}\right)\right], \quad \text{Eq. 5.9}$$

where *s* is the section width, n_{actual} is the total number of sections after accounting for dividends when sectioning *l* by *s* pixels, *w* represents the width of the ROI, and $I_{si+1,j}$, $I_{s\cdot(i-1)+1,j}$, and $I_{s\cdot(n_{actual}-1)+1,j}$ are the pixel brightness values at image positions $si+1, j$, $s\cdot(i-1)+1, j$, and $s\cdot(n_{actual}-1)+1, j$, respectively, designated by the *i*th section.

However, image roughness failed to characterize the extent of hair surface damage in SEM images, as the metric does not sufficiently correlate surface roughness with variation in pixel brightness. Average image roughness, optimized with summation of brightness differences in 100 sections, yielded a correlation of only 0.259 with SEM damage grade[6] ($p = 0.153$; Figure 5.3b), after evaluation over a range of 10, 20, 50, 100, and pixel-by-pixel sections. For example, two images exhibited similar average image roughness, as calculated using Equations 5.7 – 5.9, despite showing substantially different extents of surface damage, assessed using the SEM damage grade system; the mostly undamaged hair was even associated with a greater average roughness than one displaying holes in the cuticle and partial exposure of the cortex, likely from overrepresentation of overlapped cuticles in the former (Appendix Figure S-5.1b) and underrepresentation of holes and partial cortex exposure in the latter image (Appendix Figure S-5.1a). It is obvious that average image roughness does not adequately represent morphological features of damage, and thus, is not an appropriate metric for scoring hair surface damage.

A third metric, tailing factor, was then examined for representing hair surface damage, since brighter morphological features of damage manifest as peak lag tailing in pixel brightness

histograms. Tailing factor of an image brightness histogram, adapted from the USP measurement of chromatographic peak tailing[42], was determined in two steps: the peak apex was first redefined, bounded by the full width at half-maximum, to remove histogram skew created by the presence of multiple peaks, and then tailing factor was calculated at a fraction of the peak height maximum. Conventionally, the metric is used to characterize peak shape in chromatographic separations and is calculated at 5% of the peak height maximum[42], though fraction *f* was optimized between 1 and 10% of the peak height maximum for this application. Peak apex brightness I_h and tailing factor t_{fH} were determined using Equations 5.10 – 5.12:

$$\bar{I_H} = \frac{\sum_{i=I_{lead,0.5H}}^{I_{lag,0.5H}} ic_i}{\sum_{i=I_{lead,0.5H}}^{I_{lag,0.5H}} c_i}, \quad \text{Eq. 5.10}$$

$$I_h = \begin{cases} I_H, & |\bar{I_H} - I_H| \leq 3 \\ \bar{I_H}, & |\bar{I_H} - I_H| > 3 \end{cases}, \quad \text{Eq. 5.11}$$

and

$$t_{fH} = \frac{I_{lag,fH} - I_{lead,fH}}{2(I_h - I_{lead,fH})}, \quad \text{Eq. 5.12}$$

where $I_{lead,0.5H}$ and $I_{lag,0.5H}$ represent the peak lead and lag brightness, respectively, at 50% of the brightness profile peak height maximum *H*, c_i is the frequency of pixel brightness *i*, I_H is the brightness value at *H*, and $I_{lead,fH}$ and $I_{lag,fH}$ represent the peak lead and lag brightness at fraction *f* of the peak height maximum.

Analysis of pixel brightness histograms in a manner similar to chromatographic peak tailing more effectively captured roughness associated with hair surface damage. Average pixel brightness histograms showed pronounced peak lag tailing for SEM images of exploded hairs compared to controls (Figure 5.3c), linked to the pixel brightness of damage features, and thus, to roughness. As damage features accumulate in an image, surface roughness increases along

with a higher proportion of brighter pixels, thereby positively skewing the image brightness profile and creating a tailing effect. A tailing factor of 1 indicates a symmetrical peak and thus, the absence of tailing, while values greater than 1 indicate peak lag tailing. For quantification of hair surface roughness, the tailing factor for pixel brightness histograms yielded maximum statistically significant correlation with SEM damage grade when determined at 2% of the height maximum ($r = 0.823$; $p = 7.31 \times 10^{-9}$; Figure 5.3d). In contrast to average image roughness, comparison of SEM images in Appendix Figure S-5.1a and S-5.1b demonstrated good agreement between SEM damage grade and tailing factor; small holes, lifting of the cuticle edges, and peeling of a cuticle layer partially exposing the cortex were features in the exploded hair that contributed to a tailing factor of 2.473, compared to tailing factor 1.451 in Figure S-5.1b for a control hair.

Tailing factor further represents more generalized features of hair surface damage and requires no predefined characteristics for scoring of damage. Given the strong correlation, a kNN model was developed and tested for prediction of SEM damage grade using tailing factor; with 3 nearest neighbors, 81% classification accuracy was achieved (Table 5.1), reiterating the success of capturing the same features defined by the SEM damage grade system. However, three misclassified images highlight the limitations of a classification system based on specific microscopic features. For example, a higher damage grade was predicted for Appendix Figure S-5.2a, an exploded hair, initially classified as having Scu 2 damage from lift-up of the cuticle and presence of holes. But the presence of embedded particulates and residue remaining after removal from adhesive were ignored as they were not specified features in the damage grade criteria, though these features contribute prominently to surface roughness and damage in the image. On the other hand, tailing factor enabled prediction of a relatively undamaged control hair

to Scu 1, though classified as sustaining Scu 2 damage due to the presence of a hole and a few cuticle lift-ups (Appendix Figure S-5.2b). Classification systems based on presence or absence of observer-defined features do not provide quantitative scoring for images based on extent of damage. Tailing factor overcomes limitations of hair damage classification systems, as it is intrinsically linked to the magnitude of surface damage and it enables successful scoring of images without prior identification of specific features.

Table 5.1. Predicted microscopy damage grade and probability of prediction for SEM hair images in test set from kNN model with k = 3 based on tailing factor calculated at 2% of peak height maximum.

Hair Sample	Sample Damage Classification	SEM Damage Grade	Tailing Factor	Predicted Damage Grade	Probability
4	Control	Scu 1	1.182	Scu 1	1
3	Exploded	Scu 1	1.197	Scu 1	1
4	Control	Scu 1	1.209	Scu 1	1
5	Control	Scu 2	1.269	Scu 1*	1
3	Exploded	Scu 1	1.306	Scu 1	1
3	Exploded	Scu 1	1.403	Scu 1	1
5	Control	Scu 1	1.434	Scu 1	1
4	Control	Scu 1	1.451	Scu 1	1
2	Exploded	Scu 1	1.521	Scu 1	0.667
1	Exploded	Scu 1	1.570	Scu 1	0.667
2	Exploded	Scu 1	1.687	Scu 2*	0.667
3	Exploded	Scu 2	1.825	Scu 2	0.667
1	Exploded	Scu 2	1.883	Scu 2	0.667
1	Exploded	Scu 2	1.901	Scu 2	0.667
1	Exploded	Scu 2	2.365	Scu 3*	0.667
1	Exploded	Scu 3	2.473	Scu 3	0.667

*Incorrectly predicted damage grade

Applied to each hair specimen, average tailing factor and its range across images describe hair damage severity more completely. For example, compared to Hair Sample 1, tailing factors for images of different regions along Hair Sample 3 are smaller (Table 5.1), thus indicating less severe cuticular damage even though both are exploded hairs. Indeed, some tailing factors for Hair Sample 3 images in the test set are similar to those for Hair Sample 4, an undamaged

control hair. However, when accounting for all of the tailing factors from SEM images of different regions along Hair Sample 3, including those from the training set, a larger average tailing factor is attained (1.545 ± 0.363 (mean ± s.d.)), with a wider tailing factor range (minimum = 1.197, maximum = 2.145), as compared to Hair Sample 4 (1.337 ± 0.153; minimum = 1.182, maximum = 1.535). Clearly, more damage has been sustained by Hair Sample 3, given the larger average tailing factor. The wider tailing factor range in the exploded hair further indicates the presence of both damaged and undamaged regions, whereas Hair Sample 4 is primarily undamaged. This large dispersion in tailing factor extends to other exploded hairs. Compared to undamaged control hairs, the tailing factor coefficient of variation (CV) is 2.5 times greater in exploded hairs (CV = 21.5%; asymptotic test; p = 0.011; Figure 5.4), signifying non-uniform severity in cuticular damage along each exploded hair. Collectively, the magnitude and range of tailing factors for each hair quantify the severity of cuticular damage in a more comprehensive manner.

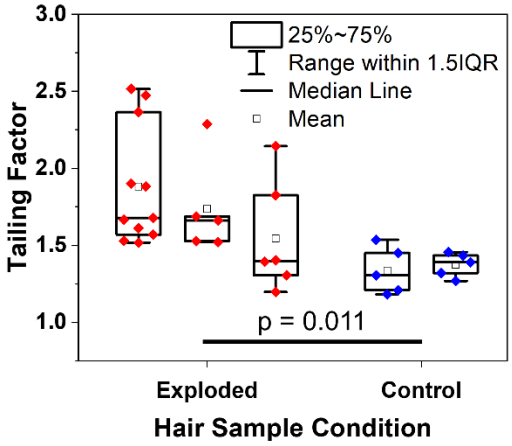


Figure 5.4. SEM image pixel brightness tailing factor as a proxy of hair surface damage in 5 single exploded and undamaged control hairs (approximately 1 cm). 32 SEM images across 5 single hairs were used in this analysis, with at least 5 imaged regions along each exploded and control hair. Each data point represents the tailing factor for a specific imaged region within the hair sample. The severity in morphological damage within each exploded hair varies significantly more than control hairs (asymptotic test; p = 0.011), indicating non-uniform mechanical and thermal damage along each single hair recovered after an explosion.

Contrary to expectations, all morphological damage in exploded hairs was localized to the cuticle, as captured and quantified via tailing factor as a proxy for hair surface damage in SEM images. The next section, Section 5.3.2, describes the effects of an explosive blast on the hair proteome via peptide populations identified by mass spectrometry and whether these findings corroborate those observed via morphological analysis.

5.3.2 Effects of Mechanical Damage on the Hair Proteome

This section aimed to evaluate effects of an explosive blast on the hair proteome, towards an assessment of success rates in identifying genetically variant peptides (GVPs) for differentiation of individuals in damaged single hairs. Comparison of proteins and unique peptides, i.e., peptides assigned to a single protein, between exploded and undamaged control hairs revealed no significant effect of damage from an explosive blast on the numbers of identified proteins and peptides. Similar numbers of proteins and unique peptides were identified between the two populations; 104 ± 39 (mean ± s.d.) proteins and 998 ± 502 unique peptide sequences were identified in exploded hairs (n = 3 replicates from a single individual), not statistically different from 106 ± 19 proteins and 971 ± 294 peptides from control hair samples (n = 5 from 3 different individuals; two-sample t-test; $p \geq 0.940$; Figure 5.5a-b). This initial observation suggests that protein degradation is minimal from the explosive blast. However, further examination of proteins and unique peptides annotated for both sample sets is needed to assess the extent of degradation in individual proteins resulting from explosion conditions, as an aggregate measure of protein damage may not adequately reflect effects of protein damage on SNP detection.

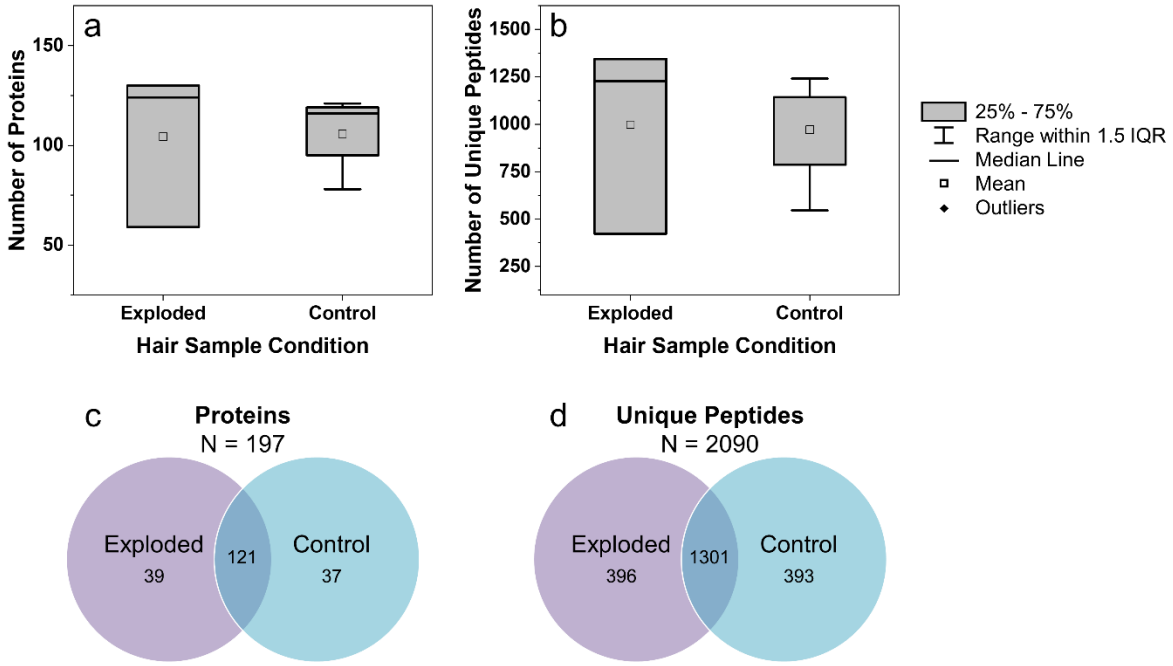


Figure 5.5. Comparison of numbers of identified (a) proteins and (b) unique peptides between exploded and undamaged control hairs (n = 8), and the overlap in composition of (c) proteins and (d) unique peptides between exploded and control single hairs from Individual 1 (n = 6). Exploded hairs yield similar numbers of proteins and unique peptides as compared to undamaged control hairs (two-sample t-test; $p \geq 0.940$). Composition overlap was examined by comparing pooled populations of proteins and unique peptides identified within each sample set. The majority of proteins (61%) and unique peptides (62%) overlap, indicating detection of similar populations of proteins and unique peptides between exploded and undamaged control hair fibers.

To compare protein and unique peptide populations, the overlap in composition of identified proteins and peptides between exploded hairs and their control counterparts (n = 3 per condition) was assessed. Only control and exploded hair samples collected from a single individual were considered to remove biases owing to interindividual protein abundance variation. One of the exploded hairs selected for analysis was unpigmented (i.e., a gray hair), while the other two hair specimens were black. As such, one gray and two black single undamaged control hair fibers were selected for analysis to match the pigments of exploded hair samples; none of the hair samples were dyed. The protein overlap between exploded and

undamaged control hairs represents 61% of identified proteins from the 6 hair specimens (Figure 5.5c). Of the 160 proteins pooled from the three exploded hairs, 121 proteins were also detected among the 158 aggregate proteins from the control hair sample set. Shared proteins consist of keratins (19%), keratin-associated proteins (34%), and cellular proteins (47%) such as V-set and immunoglobulin domain-containing protein 8 (VSIG8) and leucine-rich repeat-containing protein 15 (LRRC15), both known to be present in hair and contain GVPs[20, 43]. Likewise, the majority (62%) of unique peptides are shared between exploded and control hairs (Figure 5.5d), indicating detection of similar compositions of proteins and unique peptides among single hair samples, regardless of damage from the explosive blast.

Though the same protein populations are identified in hair fibers after an explosive blast, protein abundances may differ in single one-inch hairs from the same individual as a result of mechanical or thermal damage in hair fibers from the explosion. Resultant statistically different protein abundances potentially affect GVP identification, thus necessitating removal of potentially unreliable GVPs from consideration for human identification. Further, degradation of specific hair proteins from explosion conditions may also offer insight into protein localization within the hair fiber structure. Once correlated with the sites of morphological damage in exploded hairs, a lower abundance of certain proteins suggests their predominant expression in damaged regions of exploded hairs among the various layers in hair shaft. Proteins were quantified by summing integrated extracted ion chromatograms of identified unique peptides from MS1 survey scans and normalized to the total chromatographic peak area of all identified peptides, including shared peptides. Of 197 proteins identified in six single one-inch hair samples (n = 3 per condition) from the same individual, only two (1.0%) showed statistically lower abundance in exploded hairs: K75 and KAP4-6 (two-sample t-test; $p \leq 0.024$; Figure 5.6a-

b). Results of statistical testing of abundances from all identified proteins are presented in Appendix Table S-5.1. Large variances accompany the mean abundances of many proteins, producing non-statistically different protein abundances between groups. Surprisingly, K33B exhibited statistically higher abundance in exploded hairs (Figure 5.6c). As the protein is predominantly found in the cortex of the hair shaft[44], abundance of type I keratin K33B was not expected to be greater in exploded hairs; perhaps damage from the event facilitated accessibility to a few hair proteins including K33B for digestion, thus resulting in a higher abundance of the corresponding tryptic peptides, though such effects on peptides from other proteins were not statistically significant. The keratin-associated protein KAP4-6 is a member of the ultra-high sulfur KAP family (> 30 mol % cysteine content)[45] and facilitates keratin crosslinking[46]. Localization of its mRNA expression pattern has not been specifically elucidated, though it is thought to reside in the cortex; analysis of the closely related family member KAP4-3 demonstrated high expression in the cortex[45], but perhaps KAP4-6 is peripherally expressed and participates in crosslinking in the hair cuticle or proximal to the cuticle, thereby exhibiting an abundance decrease in exploded hairs. On the other hand, K75 is an abundant cytoskeletal keratin expressed primarily in the companion layer of scalp hair follicles[47, 48], with lower cysteine content than typical α-keratins in hair shaft (e.g., K81)[48]. Localization of K75 to the companion layer, which is the innermost layer of the outer root sheath of the hair follicle[49] that lies exterior to the cuticle layer[50] may have increased its susceptibility to decomposition from the explosion. Nevertheless, with the exception of K75 and KAP4-6 degradation, which is examined in later sections, few differences in aggregate protein abundances indicate minimal protein degradation resulting from an explosive blast, suggesting little effect on GVP identification.

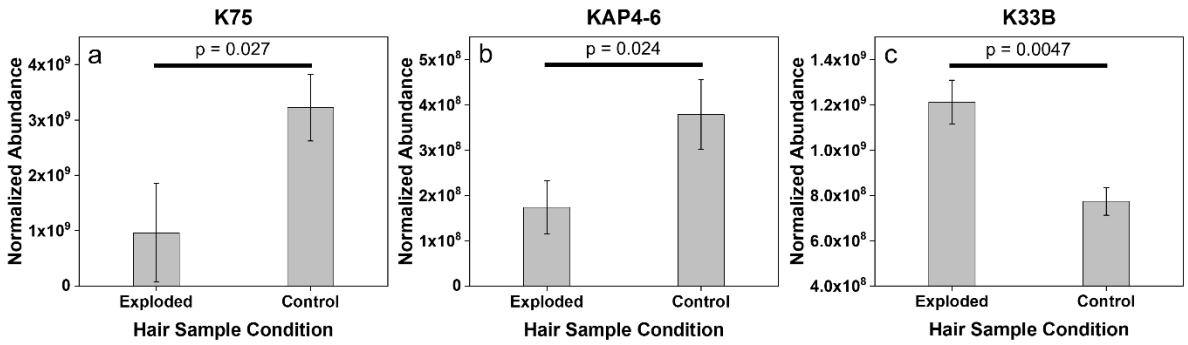


Figure 5.6. Proteins with statistically different abundances from exploded and undamaged control hairs from Individual 1 (two-sample t-test). Error bars represent the standard deviation (n = 3 per condition). Interestingly, a greater abundance of K33B was found in exploded hairs, though this result is likely unrelated to the explosion event. Only 1% of all identified proteins yielded statistically lower abundances in exploded hairs, consistent with minimal degradation of protein resulting from an explosive blast.

More subtle signs of hair proteome degradation may be found in the form of chemical modifications, particularly oxidation of amino acids cysteine, tyrosine, and tryptophan, as decomposition of these specific amino acids have been postulated to induce yellowing in thermally damaged hair[17, 18]. Chemical modifications were grouped by the affected amino acid and carbamidomethylation-related modifications were excluded from analysis, as these were intentionally created during hair sample preparation. Figure 5.7 displays aggregate frequencies of chemical modifications by amino acid. Statistical comparisons of accumulated modifications to amino acids showed that total chemical modifications occur with similar frequency in unique peptides between exploded and control hairs from Individual 1 (two-sample t-test; $p \geq 0.079$). Furthermore, when comparing each chemical modification for individual amino acids, modification frequencies also did not differ between exploded and undamaged control hairs (two-sample t-test; $p \geq 0.056$). The 10 most abundant modifications and their frequencies in exploded and control hairs are listed in Table 5.2, which encompass 69% of identified modifications in each single hair sample. As expected, deamidation in glutamine and asparagine

residues and methionine oxidation were the most prevalent modifications, though it is surprising that a greater extent of oxidation in exploded hairs was not observed.

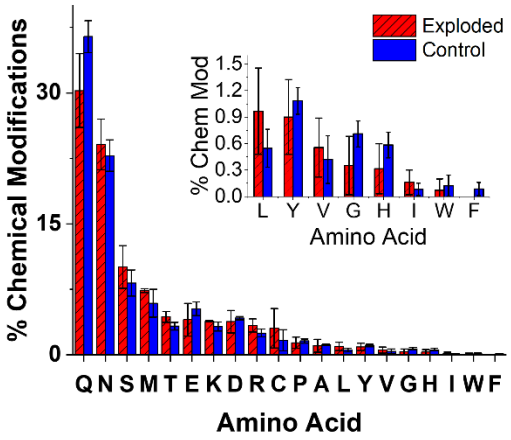


Figure 5.7. Frequency of all chemical modifications for individual amino acids in identified unique peptides, comparing exploded and undamaged control hairs from Individual 1. Inset expands those with ≤ 1% chemical modification. Carbamidomethylation modifications were excluded. Error bars represent standard deviations (n = 3 per condition). The frequencies of total modifications for each amino acid were not statistically different between exploded and control hairs (two-sample t-test; $p \geq 0.079$). Even when comparing each chemical modification for individual amino acids (data not shown), frequencies were similar between exploded and control hairs (two-sample t-test; $p \geq 0.056$), consistent with minimal protein modification resulting from an explosive blast.

Table 5.2. Chemical modification frequencies in exploded and undamaged control hairs for the 10 most abundant modifications, which account for 69% of chemical modifications identified in each hair sample, excluding carbamidomethylation-related modifications. Frequencies of chemical modifications were not statistically different between exploded and control hair samples, indicating little evidence of hair proteome degradation in exploded hairs via induction of chemical modifications.

Amino Acid	Chemical Modification	Exploded Hair Frequency (%, M ± SD)	Control Hair Frequency (%, M ± SD)
Q	Deamidation	29.4 ± 4.1	35.5 ± 2.0
N	Deamidation	23.2 ± 3.4	21.6 ± 1.9
M	Oxidation	5.3 ± 1.4	4.9 ± 1.5
S	Phosphorylation	3.1 ± 0.6	2.8 ± 0.4
E	Methylation	1.1 ± 0.6	1.2 ± 0.2
S	Acetylation (Protein N-term)	1.5 ± 1.2	0.8 ± 0.2
K	Formylation	0.9 ± 0.4	0.8 ± 0.3
T	Acetylation (Protein N-term)	1.0 ± 0.8	0.7 ± 0.1
R	Deamidation	0.7 ± 0.2	1.0 ± 0.1
C	Acetylation (Protein N-term)	1.0 ± 0.4	0.6 ± 0.4

In contrast to the prevalence of oxidized methionine residues, which has been linked to age-related decomposition[51], oxidative modifications to other residues were sparse. Notably, modifications to tryptophan were rare, with addition of one oxygen (+15.9949 Da) and further oxidative conversion to kynurenine (+3.9949 Da), respectively, observed once each in an exploded and control hair specimen (0.1% frequency, respectively) (Figure 5.7). Unlike McMullen and Jachowitz's accounts of tryptophan degradation in thermally-damaged hair[17], 99% of tryptophan residues, detected in approximately 7 – 8% of unique peptides in each hair sample, remained unmodified. Similarly, chemical modifications to tyrosine residues also occurred infrequently (on average, 1%). More relevant oxidative events include hydroxylation of Tyr (+15.9949 Da) to dihydroxyphenylalanine (on average, 0.04%), proposed to indicate oxidative stress[52], and oxidation (+31.9898 Da; 0.16%), annotated as trihydroxyphenylalanine, but neither modification differed in frequency among exploded and undamaged control hairs.

Few modifications to cysteine were detected, despite its sensitivity to reactive oxygen species[53] and Richena and Rezende's observations of increased cysteine sulfonic acid in photodamaged hair[14], likely as a result of enforcing cysteine carbamidomethylation as a fixed modification. Oxidation to cysteine sulfinic acid comprised, on average, 0.6 ± 0.2 % and 0.7 ± 0.5 % of all chemical modifications excluding carbamidomethylation in exploded and control hairs, respectively, and further oxidation to cysteine sulfonic acid was not observed. However, frequencies of detection of cysteine sulfinic and cysteine sulfonic acids did not differ between exploded and control hairs even when data were processed using a variable carbamidomethylation of Cys. Contrary to expectations, few oxidative or degradative modifications were identified in cysteine, tryptophan, and tyrosine residues, and accumulation of these oxidative modifications occurred with similar frequency between exploded and control hairs. These results show no evidence of sustained oxidative damage from explosion conditions, perhaps due to the transient nature of the event.

In sum, with minimal protein abundance differences and no difference in chemical modification accumulation, the hair proteome remained largely unaffected by the explosive blast in this investigation, which allows identification of unique proteins and peptides to the same extent from exploded hairs and indicates minimal effect on GVP identification. Furthermore, proteome analysis identified potential markers of hair cuticular damage within the minimal observed differences in protein abundance, which corroborates with the hair surface damage assessment via morphological analysis. With the results from both morphological and proteomic analyses, Section 5.3.3 focuses on correlating these findings to determine whether morphological analysis can be used as a rapid predictor of proteomic profiling success in damaged hairs.

5.3.3 Comparison of Morphological and Proteomic Profiling of Mechanical Hair Damage

Morphological assessment of hair damage from an explosive blast via scanning electron microscopy was investigated as a quick and inexpensive orthogonal technique that can serve as an indicator of proteomic profiling success to improve analysis throughput. Hair fibers recovered from the site of an explosion may exhibit graded levels of mechanical damage owing to proximity to the explosive and its proteome potentially degraded, which would subsequently affect GVP identification. Microscopic analysis of damaged hairs would be an effective predictor of proteomic profiling success if statistically significant correlations occurred between morphological hair damage and changes in chemical composition. Thus, using tailing factor as a quantitative measure to score morphological hair damage as described in Section 5.3.1, correlations with proteomics metrics were performed.

Minimal correlation between morphological damage and hair proteome degradation demonstrates proteomic profiling independence from hair cuticular damage. Figure 5.8 displays correlations between tailing factor, as a measure of hair morphological damage, and proteomics metrics indicative of protein degradation. For correlations between tailing factor and protein abundance, only those proteins detected in the majority ($\geq 60\%$) of hair samples were considered, but none correlated significantly with tailing factor ($p \geq 0.083$). However, the abundances of proteins K75, K80, K40, and KAP10-11 showed strong negative correlation with tailing factor, where single hairs with more severe cuticular damage exhibited a lower abundance of these three cytoskeletal keratins and one keratin-associated protein (Spearman's rank correlation; $\rho \leq -0.82$; Figure 5.8c-f). This observation further supports the discussion above regarding localization of K75 and its presence in the hair companion layer. Similarly, K80 resides in the companion layer, though *in situ* hybridization studies indicate that the protein

exhibits more promiscuous behavior than that of K75, including expression in the hair cuticle and all the layers in the inner root sheath in scalp hair[54]. As such, it is not surprising that K80 abundance decreases with larger tailing factors measured using SEM, indicative of increasing severity of hair cuticular damage from the explosive blast. Of note, K80 exhibited three orders of magnitude lower abundance compared to K75; although both are cytoskeletal keratins, K80 expression is weaker in scalp hairs, but it is highly expressed in medullated beard hair[54]. K40 is yet another cytoskeletal keratin, though a type I keratin, and is localized to the cuticle[48]. The keratin-associated protein KAP10-11, a member of the high sulfur KAP family (< 30 mol % cysteine content), also exhibits predominant mRNA expression in the hair fiber cuticle. Indeed, all members of the KAP10 family concentrate in a narrow region within the hair cuticle, lying approximately 20 cell layers above the dermal papilla[46]. Damage to the hair cuticle reasonably explains the observed negative correlations between tailing factor and abundances of proteins K75, K80, K40, and KAP10-11, known for their localization to the cuticle and exterior. SEM image analysis provides corroborating evidence of localized cuticular damage as a result of an explosive blast. Proteins with predominant expression external to the cortex and medulla experience greater susceptibility to explosion conditions and as such, serve not only as biomarkers of proteome degradation in single hairs recovered post-blast, but also broadly as indicators of hair cuticular damage. However, as these correlated localized proteins make up the minority of all identified proteins (2%) with minimal degradation in the hair proteome, successful proteomic profiling can be achieved in recovered single hairs regardless of the extent of morphological hair cuticular damage when the cortex and medulla remain intact.

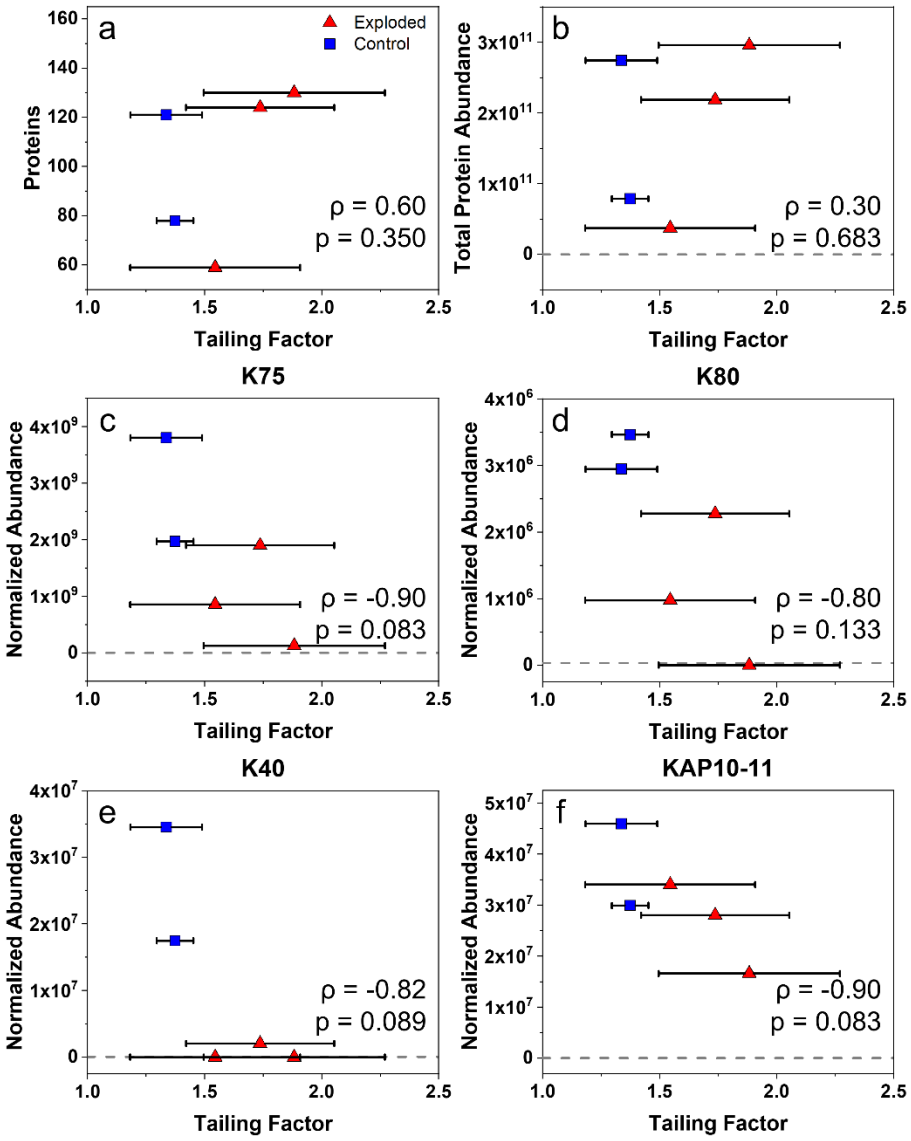


Figure 5.8. Relationships between SEM image tailing factor proxy of morphological damage and proteome analysis results of (a) number of proteins identified, (b) total protein abundance, and normalized abundances of (c) K75, (d) K80, (e) K40, and (f) KAP10-11 for each of 5 SEM-imaged exploded (red triangle) and control (blue square) hairs. Spearman's rank correlations (ρ) show strong negative correlations of damage with these individual protein abundances but are not significant at the $\alpha = 0.05$ level. Error bars represent standard deviations for tailing factor, of which at least 5 regions along each single hair were imaged, from a total of 32 SEM images. The dashed lines in (b) – (f) represent the threshold for selecting peptide precursor ions for MS/MS fragmentation, set at 3.3×10^4 counts.

5.3.4 Potential to Differentiate Individuals Using Exploded Hairs

Minimal differences in protein abundance and accumulation of chemical modifications from an explosive blast suggest little effect of sustained hair cuticular damage on GVP identification. To evaluate success rates in detecting GVPs for SNP inference in exploded hairs, overlaps in identified SNPs and GVP profiles were examined, with particular attention to GVPs in K75, K80, K40, and KAP10-11 because these proteins degrade with more severe hair surface damage induced by the explosion.

Detection of a similar number of SNPs from both major and minor GVPs and large overlap of SNPs identified by both exploded and undamaged control hairs indicate that GVP identification has not been hindered by explosion-related hair damage. Statistical comparison of SNPs from both major and minor GVPs showed no difference between exploded and control hairs (two-sample t-test; $p \geq 0.713$; Figure 5.9a-b). Further, SNPs inferred from major and minor GVPs exhibit substantial overlap (79% and 65%, respectively) (Figure 5.9c-d), signifying that the same SNPs are identified from major GVPs regardless of hair damage.

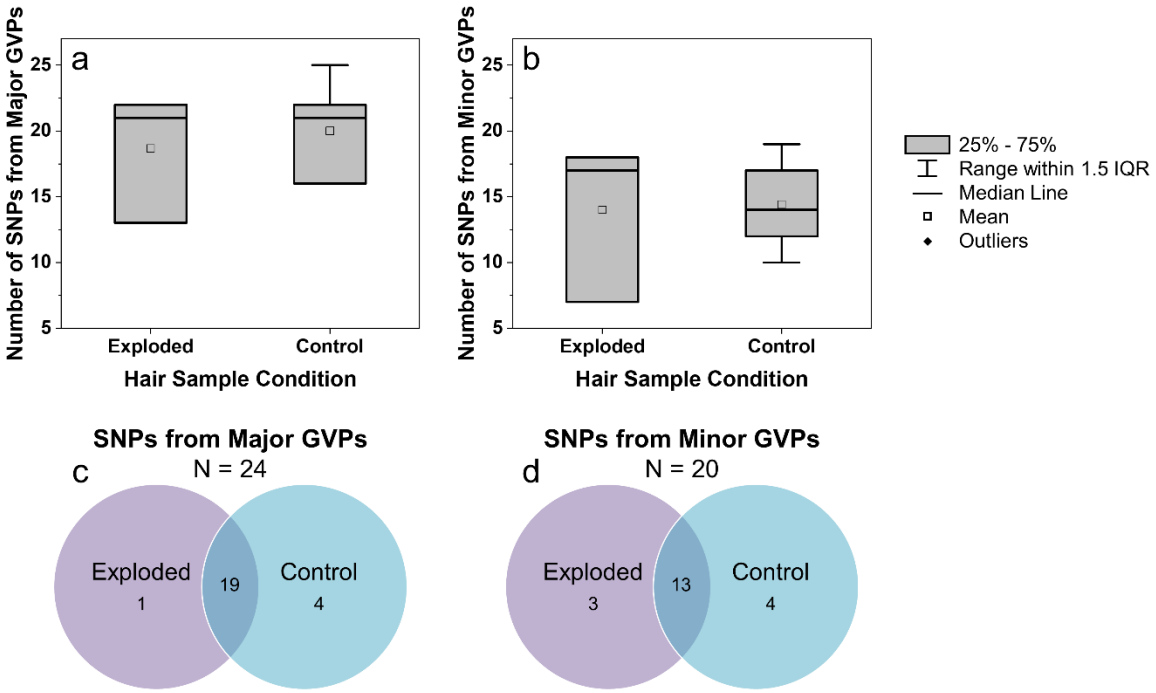


Figure 5.9. Comparison of numbers of SNPs from (a) major and (b) minor GVPs identified in digests of exploded (n = 3 hair samples from the same individual) and undamaged control hairs (n = 5 hair samples from 3 individuals). Exploded hairs yield similar numbers of SNPs as compared to undamaged control hairs (two-sample t-test; $p \geq 0.713$). Overlap in SNPs from (c) major and (d) minor GVPs in exploded and control single hairs from Individual 1 (n = 6 samples), from aggregate SNPs identified within each of the two populations. SNPs identified from major and minor GVPs substantially overlap between the two populations (79% and 65%, respectively).

Analysis of GVP profiles showed that one SNP each was inferred from GVPs in K40 and KAP10-11, proteins associated with the cuticle and surface. GVP profiles were established for each single hair sample based on the presence or non-detection of major and minor GVPs that enable inference to the corresponding SNP (Appendix Table S-5.2). A simplified set of profiles are shown in Figure 5.10a, where the locus frequency based on the major and minor GVP responses is represented for each SNP. Locus frequency was established as described in a previous work,[20] as the sum of the heterozygote population frequency and homozygote frequency for the major or minor phenotype in the absence of the major or minor GVP, respectively. In cases where multiple SNPs from the same gene were inferred, a one-SNP-per-

gene rule was adopted, in which the SNP yielding the most consistent response among hair samples and with the expected genotype was selected. For SNP rs9908304 in KRT40, major and minor GVPs were expected in two of the three individuals, respectively, based on exome sequence information. Examination of the GVP profiles showed that only the major GVP was identified, once in control hair specimen 1-Ctrl.B (Figure 5.10a). The GVP YFNTIEDLQQKILC**T**KAENSR (mutation site denoted in larger, bold red text) was detected at a signal abundance of 6.42×10^5, somewhat above the ion threshold of 3.3×10^4, and contains two missed cleavage sites; the shortest tryptic GVP does not meet the minimum length for peptide identification and requires the preceding amino acid sequence for confirmation as a peptide from K40. Similarly, the shortest tryptic minor GVP requires either the preceding or succeeding sequence for peptide identification, but reliable detection of peptides containing missed cleavage sites, which depends upon variable protein digestion efficiency, cannot be expected, as demonstrated by the lower abundance of this GVP. Both detected major and minor GVPs enable inference of SNP rs462007 in KRTAP10-11, although their successful detection among undamaged control hairs is variable at best (67% and 50% detection, respectively), owing to their length. Typical of peptides in KAPs, which contain few tryptic cleavage sites, these GVPs are defined by long sequences amid repeated arginine-proline units, with the major GVP 48 amino acids in length and a version of the minor GVP with 38 amino acids, both well above the observed average tryptic peptide length of 22 ± 12 amino acids. Longer peptides present a challenge for peptide identification, as their inclusion necessitates confident identification of their fragment ions during *de novo* sequencing and matching to peptides in the protein database, which is more difficult to attain due to the need for a greater number of matched fragment ions to achieve a higher confidence score. Thus, while degradation of K40 and KAP10-11 in hair cuticle

may play a role in the non-detection of these GVPs in exploded hairs, the difficulties involved with detecting the longer GVPs, potentially with missed cleavage sites, among undamaged control hairs detract from their appeal as robust candidate markers for a GVP panel.

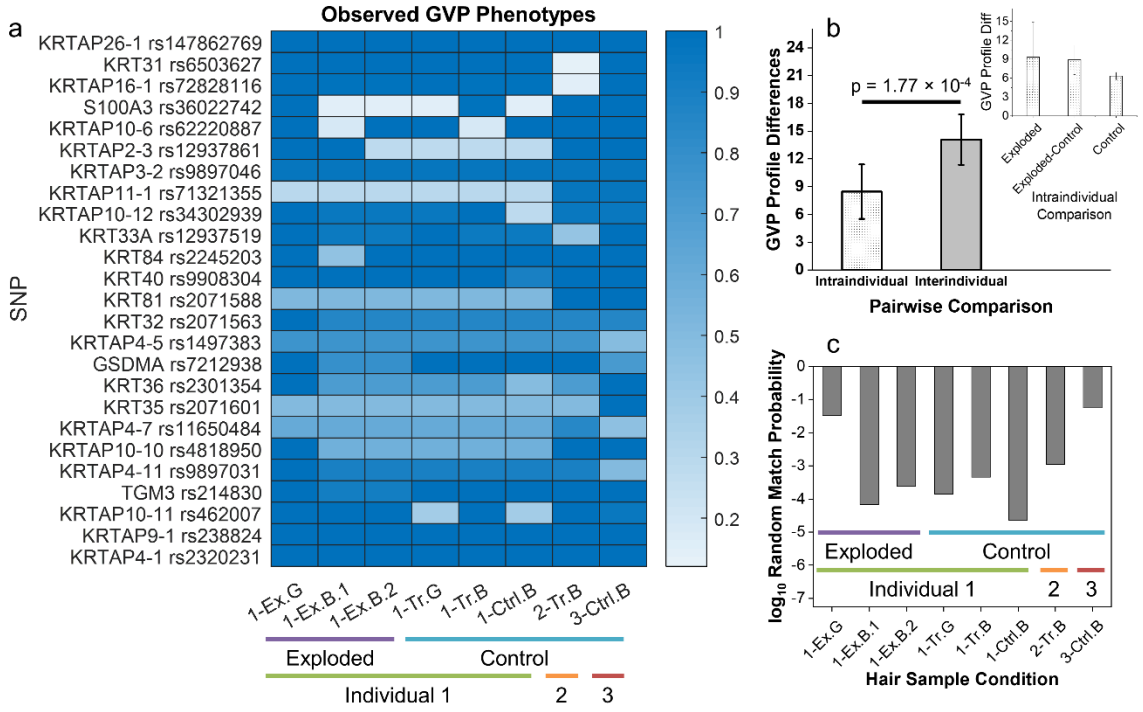


Figure 5.10. (a) GVP profiles, (b) number of GVP profile differences among pairwise comparison groups, and (c) resulting random match probabilities of undamaged control and exploded hairs. Samples are denoted x-y.z.a, where x is the individual, y indicates the sample condition of exploded (Ex), travel blank (Tr), or control (Ctrl) blank, z represents hair pigmentation as black (B) or gray (G), and a is the sample replicate, i.e., different hairs from an individual. Error bars in (b) represent the standard deviation. SNPs were not detected in K75 nor K80 in both populations, and only sporadically detected in K40 and KAP10-11 among control hair samples, suggesting that degradation of proteins at the hair surface has minimal effect on GVP identification. GVP profiles from exploded hairs were similar to control hairs from Individual 1, as evidenced by a similar number of intraindividual profile differences, which shows statistically fewer differences compared to interindividual comparisons (Wilcoxon Rank Sum test; $p = 1.77 \times 10^{-4}$). Quantification of differentiative potential via random match probability demonstrates that individuals can be distinguished to a similar extent regardless of hair damage induced by an explosive blast.

Closer examination of the SNPs inferred from detected major and minor GVPs among exploded and undamaged control hair samples from Individual 1 established the non-detection of GVPs from K75 and K80. Instead, the majority of keratins in which SNPs are inferred belong to

the hard α-keratin class, including both types I and II, that dominate hair shaft (Figure 5.10a). Exome sequences of the three individuals were then analyzed to determine whether GVPs were expected from these proteins. Of the 269 missense SNPs from the gene KRT80 annotated in the SNP reference database (dbSNP)[55], none were identified in the exome sequences of this set of subjects. This accounts for the absence of minor GVPs from K80; it is likely that major GVPs from this protein, along with other tryptic peptides, exhibited abundances below the MS/MS fragmentation threshold, contributing to an average sequence coverage of 5% among hair samples. On the other hand, two SNPs from KRT75 were detected across the three individuals, and as such, minor GVPs from SNPs rs298104 and rs298109 were expected. To confirm that minor GVPs corresponding to the two SNPs would be detected in a proteomics experiment barring any influence from low variant peptide abundances, an *in silico* tryptic digest of the canonical and mutated K75 sequences was performed; Table 5.3 summarizes detection feasibility for the shortest tryptic GVPs. Among the four variant peptides, only the minor GVP from SNP rs298104 can theoretically and feasibly be detected using the current analytical scheme; however, an average sequence coverage of 5% for shorter tryptic peptides well upstream of this variant peptide likely resulted in the absence of this GVP. In contrast, average sequence coverage of the seven cuticular keratins that yield GVPs is 37%, based only on unique peptides; average protein sequence coverage increases to 63% when including shared peptides. These examples not only highlight dependence of GVP identification on sequence coverage, but also, and more importantly, coupled with the observations in K40 and KAP10-11 above, demonstrate through their non-detection in undamaged control hairs, minimal effect of explosion-related hair cuticular damage on GVP identification.

Table 5.3. Detection feasibility for the shortest tryptic major and minor GVPs corresponding to two SNPs identified in the exome sequences of the three individuals, resulting from an *in silico* tryptic digest of the canonical and mutated K75 sequences. Locations for the amino acid polymorphisms for each SNP are denoted in larger, bold red text.

SNP	Variant Type	GVP	Detection Feasibility	Comment
rs298104	Major	LSGEGVSPVNISVVTSTLSS GYG**S**GSSIGGGNLGLGGGSG YSFTTSGGHSLGAGLGGSGF SATSNR	Not likely	Longer than the average 22-amino acid peptide (66 amino acids)
rs298104	Minor	LSGEGVSPVNISVVTSTLSSGYG**R**	Feasible	--
rs298109	Major	ASN**R**	Not feasible	Shorter than the minimum length for identification (6 amino acids) and common to many proteins
rs298109	Minor	ASN**G**FGVNSGFGYGGGVGGG FSGPSFPVCPPGGIQEVTVN QSLLTPLHLQIDPTIQR	Not likely	Longer than the average 22-amino acid peptide (57 amino acids)

Pairwise comparison of GVP profiles yielded a similar number of profile differences and differentiative potential among exploded and control hair samples from Individual 1. Not surprisingly, most consistent GVP detection and SNP inference arise from keratins owing to their dominance in hair shaft and greater sequence coverage compared to non-keratinous proteins. GVP profile differences, quantified through pairwise comparisons, between exploded and control hair samples (9 ± 2 differences) are similar to those comparisons among exploded hairs (9 ± 6 differences) and also to those among undamaged control hairs (6 ± 1 differences) (Figure 5.10b inset). The slightly greater variation in GVP profile differences among exploded hairs originates from pairwise comparisons involving the profile of a single exploded gray hair (1-Ex.G), which exhibited many more non-detects in GVP responses than any other hair sample from Individual 1, including responses from keratins (Figure 5.10a). The sparsity in GVP responses is linked to

low peptide yields from the protein digest of this gray hair sample (5.0% peptide-spectrum matches as opposed to an average of 16.9% PSMs from other hair samples belonging to Individual 1), perhaps as a consequence of its hair follicle differentiation mechanism, which differs from pigmented hairs. However, the small number of unpigmented hair samples in this study limited further exploration of the role of pigmentation on GVP detection. Except for 1-Ex.G, similar differentiative potential was achieved between exploded and undamaged control hairs from the same individual (Figure 5.10c), quantified via random match probability as the product of SNP loci frequencies for each sample. Therefore, single-inch pigmented hairs recovered after an explosion display probative value equivalent to undamaged hairs from the same individual for forensic identification.

Despite the profile variation in intraindividual pairwise comparisons, statistically greater interindividual differences in GVP profiles (14 ± 3 differences) illustrate differentiation among the three individuals based on identified GVPs (Wilcoxon Rank Sum test; $p = 1.77 \times 10^{-4}$; Figure 5.10b). For example, the presence of SNPs rs6503627 from KRT31, rs12937519 from KRT33A, and rs72828116 from KRTAP16-1 distinguishes Individual 2 from the other two subjects, and successful detection of their minor GVPs enhances discriminative power of profiles generated from hair samples belonging to the individual. However, the GVP profile from Individual 3 does not exhibit distinctive or as many SNPs as identified from the other two individuals, resulting in a more common profile with lower discriminative power. Of the three individuals, GVP profiles from Individual 1 present the highest differentiative potential with a greater number of inferred SNPs (i.e., between 1 in 2,180 and 1 in 43,725, excluding 1-Ex.G). Further development of more sensitive mass spectrometry approaches including data-independent analyses or multiple analyses that employ exclusion of known peptides is expected to enable deeper interrogation of

the hair proteome to increase sequence coverage and improve GVP identification and RMPs for greater discriminative power. Nonetheless, the current analytical approach successfully demonstrates minimal hair proteome degradation following explosion, with equal ability to identify GVPs in recovered single hairs that have sustained explosion-related hair cuticular damage.

5.4 Conclusions

This research describes minimal effects of explosive blast on proteins in damaged hair evidence, of which this is the first report of combining morphological and chemical analyses to examine mechanical and thermal hair damage. Addressing a need for more objective methods in microscopic analysis, the novel metric described herein, tailing factor, quantified hair surface damage severity by exploiting pixel brightness in elevated and depressed morphological features of damage in SEM images, as a proxy for surface roughness. This work also directly enabled an investigation into the correlations of morphological damage and chemical composition changes in exploded hairs. Comparison to protein abundance measurements of exploded hair specimens confirmed localization of a subset of proteins to hair structural components and corroborated an intact hair proteome within the cortex, with non-uniform damage along each exploded hair restricted to the hair cuticle and exterior, and identifies protein markers of hair cuticular damage.

These findings build upon previous work to demonstrate equivalent success of detecting peptide biomarkers from hairs following harsh exposures compared to undamaged hair. Additionally, these results provide a foundation for selection of targets for inclusion in a GVP panel as part of data-independent approaches for application to forensic casework. Development of a targeted method to detect specific GVP markers that parallels the well-established detection of short-tandem repeats in nuclear DNA, which is currently underway, minimizes false negatives

in complex matrices by enhancing sensitivity and enables a more confident and reliable GVP identification process, which are critically important to forensic analyses.

Successful characterization of morphological features unique to exploded hairs demonstrates the ability for tailing factor to accommodate a diverse set of features as a broad metric to probe surface topography, which may find utility such as in the clinical, forensic, and material sciences to provide quantitative microscopic analyses of mechanical damage in hair and other materials. Effective extraction and quantification of chemical information in damaged matrices suggests applicability of protein-based human identification to hair evidence damaged under other mechanically and thermally harsh conditions, such as from fires and vehicular crashes. Further, this work extends beyond human identification and forensics, including species identification from damaged hairs in archaeology, and in the agricultural and medical fields for evaluation of damage such as from radiation.

APPENDIX

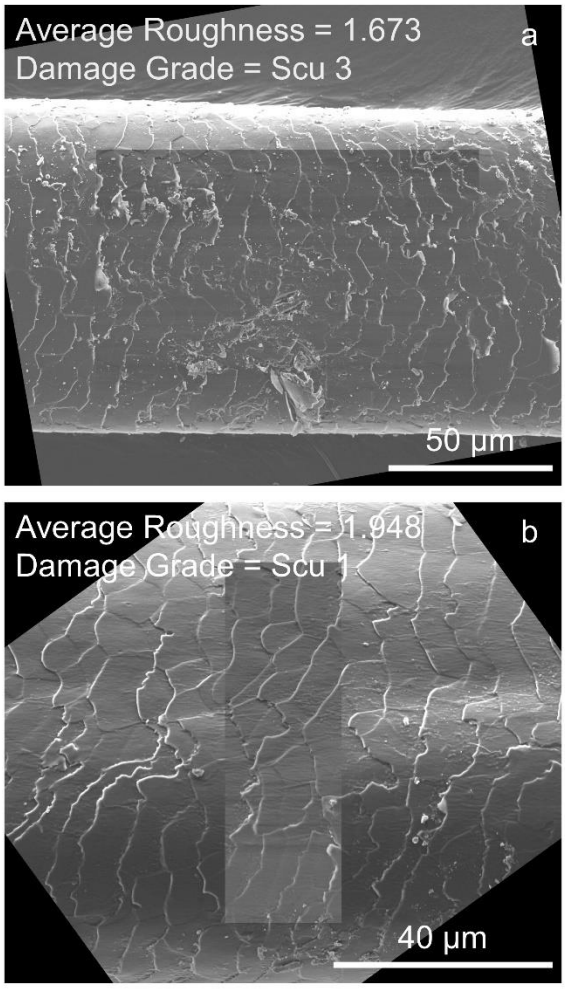


Figure S-5.1. Example rotated SEM images with overlays of normalized regions of interest from (a) Hair Sample 1 and (b) Hair Sample 4 that exhibit similar average image roughness but vastly different extents of damage, as assessed with the SEM damage grade system. Average image roughness fails to effectively characterize hair surface damage.

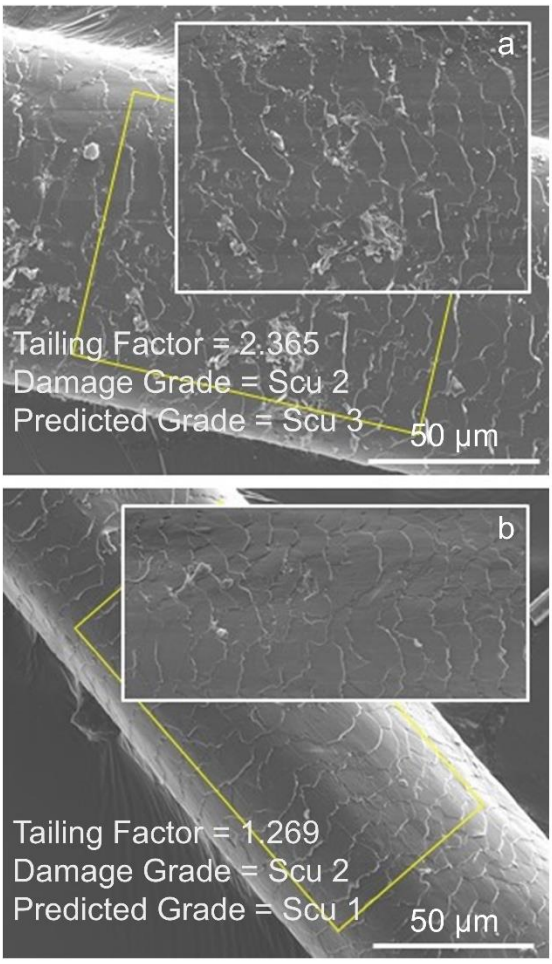


Figure S-5.2. Example SEM images of hair segments with normalized regions of interest from (a) Hair Sample 1 (exploded) and (b) Hair Sample 5 (control) whose damage grades were incorrectly predicted by the kNN model.

Table S-5.1. Average protein abundances from extracted ion chromatographic peak areas in exploded and undamaged control hairs from Individual 1 (n = 3 hairs per condition). Statistical significance from two-sample t-tests are reported.

Gene	Exploded Hair Abundance		Control Hair Abundance		p-Value
	Mean	S.D.	Mean	S.D.	
KRT85	6.46E+09	2.89E+09	5.97E+09	1.26E+09	0.810
KRT86	5.19E+09	5.84E+08	5.32E+09	1.48E+09	0.901
KRT31	4.96E+09	8.50E+08	3.30E+09	7.25E+08	0.063
KRTAP3-1	3.16E+09	4.06E+08	2.39E+09	1.70E+09	0.521
KRT34	3.72E+09	2.63E+09	2.73E+09	4.43E+08	0.581
KRT83	1.59E+09	1.39E+09	3.12E+09	1.00E+09	0.203
KRTAP4-4	2.01E+09	1.36E+09	2.26E+09	1.07E+09	0.815
KRT75	9.61E+08	8.90E+08	3.22E+09	6.02E+08	0.027
KRT81	2.43E+09	2.27E+09	1.73E+09	5.34E+08	0.649
KRTAP11-1	1.84E+09	4.65E+08	1.28E+09	7.96E+08	0.367
KRT32	9.54E+08	8.04E+08	2.55E+09	6.61E+08	0.058
KRT33B	1.21E+09	9.65E+07	7.74E+08	6.14E+07	0.005
KRTAP4-1	1.39E+09	1.64E+09	7.98E+08	3.05E+08	0.598
KRT33A	1.13E+09	4.20E+08	1.12E+09	3.07E+08	0.990
KRTAP3-2	1.03E+09	8.60E+08	1.24E+08	3.06E+07	0.209
KRTAP4-3	9.55E+08	2.33E+08	6.74E+08	2.57E+08	0.233
KRTAP9-3	1.19E+09	6.80E+08	8.19E+08	1.35E+08	0.441
KRTAP1-5	9.07E+08	7.42E+08	4.96E+08	1.62E+08	0.440
KRTAP3-3	4.73E+08	4.46E+08	1.70E+08	1.41E+08	0.362
KRTAP4-8	6.11E+08	5.61E+08	5.15E+08	2.80E+08	0.808
KRTAP4-2	4.12E+08	4.21E+07	5.84E+08	2.37E+08	0.334
KRTAP9-6	5.32E+08	2.29E+08	3.45E+08	9.83E+07	0.294
KRT35	4.06E+08	1.54E+08	3.16E+08	1.73E+08	0.541
KRTAP10-8	1.48E+08	2.57E+08	1.68E+08	2.89E+08	0.936
KRTAP4-5	3.91E+08	4.35E+08	2.80E+08	7.02E+07	0.705
KRTAP4-9	2.59E+08	2.56E+08	4.05E+08	8.27E+07	0.430
KRTAP4-11	2.55E+08	5.96E+07	1.85E+08	8.15E+07	0.303
KRTAP4-7	1.96E+08	9.97E+07	3.47E+08	2.26E+08	0.375
KRTAP4-6	1.74E+08	5.86E+07	3.79E+08	7.65E+07	0.024
KRT39	1.34E+08	7.44E+07	2.18E+08	1.76E+08	0.509
KRT82	1.25E+08	9.69E+07	1.34E+08	3.48E+07	0.895
VSIG8	1.15E+08	7.17E+07	1.11E+08	3.75E+07	0.927
KRT72	2.40E+08	2.08E+08	4.68E+07	4.12E+07	0.247
KRTAP9-9	1.25E+08	9.24E+07	1.43E+08	9.07E+07	0.822
KRTAP1-3	1.72E+08	1.51E+08	1.05E+08	4.26E+07	0.523
KRTAP9-2	1.12E+08	3.45E+07	1.55E+08	7.33E+07	0.429
KRT36	2.16E+07	1.99E+07	2.56E+08	4.22E+08	0.438
KRTAP4-12	5.33E+07	1.48E+07	1.36E+08	1.38E+08	0.408
DSP	8.91E+07	7.74E+07	6.95E+07	1.59E+07	0.707
KRTAP9-4	2.63E+07	2.44E+07	3.01E+07	3.22E+06	0.815

Table S-5.1 cont'd.

Gene	Exploded Hair Abundance		Control Hair Abundance		p-Value
	Mean	S.D.	Mean	S.D.	
KRTAP1-1	9.05E+07	1.53E+07	6.97E+07	1.87E+07	0.213
KRTAP9-7	4.80E+07	3.61E+07	5.28E+07	4.48E+07	0.891
DSG4	6.26E+07	6.09E+07	5.45E+07	5.38E+06	0.839
KRTAP9-1	2.35E+07	2.98E+07	2.08E+07	1.50E+07	0.897
KRT84	9.91E+07	1.43E+08	2.57E+07	1.85E+07	0.468
DUSP14	2.17E+07	1.89E+07	4.24E+07	4.32E+06	0.194
KRTAP10-11	2.62E+07	8.86E+06	4.21E+07	1.14E+07	0.134
KRTAP10-12	1.96E+07	1.71E+07	1.42E+07	5.05E+06	0.646
KRTAP10-9	1.88E+07	1.94E+07	1.49E+07	1.70E+07	0.810
DES	0.00E+00	0.00E+00	4.57E+07	4.02E+07	0.188
KRTAP10-10	2.34E+07	6.40E+06	1.81E+07	7.12E+06	0.390
CALML3	2.04E+07	1.77E+07	2.06E+07	6.71E+06	0.990
SELENBP1	2.82E+07	2.44E+07	1.76E+07	7.97E+06	0.536
LGALS7	1.89E+07	1.75E+07	1.59E+07	3.60E+06	0.798
KRTAP13-2	4.66E+06	4.64E+06	1.47E+07	1.48E+07	0.362
KRTAP24-1	1.07E+07	1.04E+07	1.42E+07	7.57E+06	0.665
PKP1	1.48E+07	1.72E+07	1.51E+07	5.81E+06	0.977
PRSS1	2.85E+06	4.93E+06	3.54E+05	6.14E+05	0.474
JUP	1.31E+07	1.29E+07	9.12E+06	4.71E+06	0.656
KRT40	6.82E+05	1.18E+06	1.42E+07	1.79E+07	0.322
HIST3H2BB	1.99E+07	3.45E+07	0.00E+00	0.00E+00	0.423
WNK3	1.97E+07	3.42E+07	0.00E+00	0.00E+00	0.423
LGALS3	1.17E+07	1.08E+07	4.90E+06	4.66E+06	0.399
KRT7	3.98E+05	6.90E+05	1.81E+07	2.85E+07	0.395
LRRC15	7.42E+06	8.58E+06	2.39E+06	2.09E+06	0.418
S100A3	1.28E+07	1.59E+07	3.91E+06	3.81E+06	0.437
KRTAP9-8	1.41E+06	2.44E+06	1.41E+07	2.26E+07	0.434
HSPA2	1.09E+07	1.06E+07	4.24E+06	7.35E+06	0.425
PKD2	2.29E+06	3.97E+06	4.90E+06	4.54E+06	0.496
KRTAP10-6	0.00E+00	0.00E+00	1.47E+06	2.31E+06	0.387
GPNMB	3.57E+06	6.18E+06	6.06E+06	5.28E+06	0.624
LDB3	1.29E+07	2.23E+07	0.00E+00	0.00E+00	0.423
RPSA	5.46E+06	5.11E+06	5.80E+06	5.01E+05	0.919
FABP5	1.10E+07	1.91E+07	0.00E+00	0.00E+00	0.423
KRT1	1.19E+05	2.07E+05	6.36E+06	6.80E+06	0.253
TRIM29	7.17E+06	1.24E+07	2.98E+06	3.62E+06	0.624
CTNNB1	6.12E+06	5.47E+06	3.56E+06	7.03E+05	0.503
FAM26D	4.12E+06	6.84E+06	4.87E+06	2.66E+06	0.872
KRTAP16-1	4.21E+06	5.11E+06	3.51E+06	6.08E+06	0.887
APOD	6.99E+06	9.04E+06	0.00E+00	0.00E+00	0.312
TGM3	3.94E+06	3.56E+06	2.68E+06	2.99E+06	0.665
GPRC5D	4.65E+06	5.86E+06	1.82E+06	1.66E+06	0.495

Table S-5.1 cont'd.

Gene	Exploded Hair Abundance		Control Hair Abundance		p-Value
	Mean	S.D.	Mean	S.D.	
GRN	3.45E+06	5.28E+06	8.66E+05	3.08E+05	0.486
SFN	6.37E+06	6.63E+06	5.07E+05	8.78E+05	0.264
SAMD1	5.90E+05	1.02E+06	3.56E+06	6.16E+06	0.493
PPL	6.50E+06	1.13E+07	7.19E+04	1.25E+05	0.427
HEPHL1	4.16E+06	3.65E+06	1.85E+06	8.48E+05	0.388
PMEL	0.00E+00	0.00E+00	6.00E+06	1.04E+07	0.423
CRYBG1	2.07E+06	2.52E+06	2.51E+06	1.68E+06	0.814
KRTAP13-1	4.18E+05	5.55E+05	4.64E+06	7.05E+06	0.409
NUP188	0.00E+00	0.00E+00	4.92E+06	8.52E+06	0.423
GDPD3	3.53E+06	3.14E+06	1.31E+06	1.04E+06	0.345
KRT80	1.08E+06	1.14E+06	2.45E+06	5.06E+05	0.163
GAPDH	4.29E+06	6.69E+06	0.00E+00	0.00E+00	0.383
FABP4	3.49E+06	3.97E+06	1.27E+05	2.20E+05	0.279
CCM2L	0.00E+00	0.00E+00	3.30E+06	5.72E+06	0.423
MIF	2.42E+06	2.37E+06	6.73E+05	1.16E+06	0.337
KRTAP5-2	2.38E+06	6.37E+05	5.60E+05	9.70E+05	0.062
KRTAP2-3	0.00E+00	0.00E+00	2.70E+06	4.18E+06	0.380
HSPB1	2.32E+06	2.91E+06	3.21E+05	5.56E+05	0.357
CPT1A	2.58E+06	3.79E+06	0.00E+00	0.00E+00	0.359
ATP5B	1.90E+06	2.13E+06	3.64E+05	6.30E+05	0.338
KRTAP10-7	1.30E+06	1.22E+06	9.12E+05	8.03E+05	0.673
PPIA	2.18E+06	3.78E+06	0.00E+00	0.00E+00	0.423
KRTAP12-3	0.00E+00	0.00E+00	9.56E+05	1.43E+06	0.367
TRIOBP	2.03E+06	3.52E+06	0.00E+00	0.00E+00	0.423
H1F0	2.02E+06	3.49E+06	0.00E+00	0.00E+00	0.423
TNIK	1.61E+06	1.44E+06	0.00E+00	0.00E+00	0.192
PKP3	4.70E+05	5.79E+05	5.99E+05	1.04E+06	0.862
KRT79	1.26E+05	2.18E+05	7.86E+05	1.36E+06	0.490
KRT10	0.00E+00	0.00E+00	8.10E+05	1.40E+06	0.423
GGCT	0.00E+00	0.00E+00	8.06E+05	8.22E+05	0.232
PRDX6	5.14E+05	8.91E+05	6.18E+05	5.35E+05	0.873
NEU2	5.33E+05	9.23E+05	6.36E+05	1.10E+06	0.907
KRTAP10-4	7.27E+03	1.26E+04	1.18E+05	2.04E+05	0.447
KRT38	8.92E+05	1.54E+06	4.04E+05	3.75E+05	0.643
DDX55	1.27E+06	2.20E+06	0.00E+00	0.00E+00	0.423
HSPA8	9.30E+05	8.72E+05	2.97E+05	5.15E+05	0.353
PLD3	6.03E+05	6.19E+05	6.09E+05	1.05E+06	0.994
SERPINB5	1.20E+06	1.14E+06	0.00E+00	0.00E+00	0.210
EEF2	5.38E+05	7.37E+05	6.05E+05	7.53E+05	0.917
KRTAP10-3	3.88E+05	6.72E+05	6.54E+05	5.72E+05	0.630
KRTAP7-1	0.00E+00	0.00E+00	5.33E+05	9.24E+05	0.423
CUX2	9.12E+05	1.58E+06	0.00E+00	0.00E+00	0.423

Table S-5.1 cont'd.

Gene	Exploded Hair Abundance		Control Hair Abundance		p-Value
	Mean	S.D.	Mean	S.D.	
LMNA	3.92E+05	6.78E+05	4.64E+05	4.95E+05	0.889
KRT13	0.00E+00	0.00E+00	8.41E+05	1.46E+06	0.423
YWHAE	5.52E+05	9.57E+05	2.85E+05	3.88E+05	0.688
RIDA	7.83E+05	9.16E+05	1.05E+04	1.82E+04	0.281
ALDH2	3.36E+05	5.83E+05	4.26E+05	7.37E+05	0.878
DYSF	7.62E+05	1.32E+06	0.00E+00	0.00E+00	0.423
LRP1	7.27E+05	1.26E+06	0.00E+00	0.00E+00	0.423
ACTBL2	2.35E+05	2.70E+05	4.65E+05	5.06E+05	0.537
KRTAP26-1	0.00E+00	0.00E+00	6.82E+05	9.84E+05	0.353
TPI1	0.00E+00	0.00E+00	6.17E+05	1.07E+06	0.423
HNRNPA1	0.00E+00	0.00E+00	5.99E+05	5.33E+05	0.191
TUBB2A	0.00E+00	0.00E+00	5.86E+05	2.47E+05	0.054
RALBP1	0.00E+00	0.00E+00	5.80E+05	1.01E+06	0.423
ATP5A1	5.57E+05	9.64E+05	0.00E+00	0.00E+00	0.423
CCDC157	0.00E+00	0.00E+00	4.84E+05	8.38E+05	0.423
TXN	4.80E+05	4.21E+05	0.00E+00	0.00E+00	0.187
VDAC2	3.71E+05	6.43E+05	8.26E+04	1.43E+05	0.521
NCCRP1	2.19E+05	1.92E+05	2.05E+05	1.85E+05	0.932
PLEC	3.10E+05	5.37E+05	6.28E+04	1.09E+05	0.511
LAP3	3.41E+05	5.90E+05	0.00E+00	0.00E+00	0.423
CRIP2	1.29E+05	2.23E+05	0.00E+00	0.00E+00	0.423
KRT5	0.00E+00	0.00E+00	2.68E+05	4.63E+05	0.423
FASN	0.00E+00	0.00E+00	2.41E+05	4.17E+05	0.423
HSPE1	0.00E+00	0.00E+00	2.20E+05	3.81E+05	0.423
SLC25A3	2.14E+05	3.71E+05	0.00E+00	0.00E+00	0.423
KRTAP10-5	0.00E+00	0.00E+00	1.91E+05	3.31E+05	0.423
IDH2	1.84E+05	3.19E+05	0.00E+00	0.00E+00	0.423
CXADR	1.73E+05	3.00E+05	0.00E+00	0.00E+00	0.423
KRT37	1.70E+05	2.94E+05	0.00E+00	0.00E+00	0.423
HERC4	0.00E+00	0.00E+00	1.65E+05	2.86E+05	0.423
SYNM	1.52E+05	2.63E+05	0.00E+00	0.00E+00	0.423
EEF1G	0.00E+00	0.00E+00	1.50E+05	2.59E+05	0.423
PHB2	1.36E+05	2.36E+05	0.00E+00	0.00E+00	0.423
SSBP1	0.00E+00	0.00E+00	1.24E+05	2.14E+05	0.423
TYRP1	0.00E+00	0.00E+00	1.07E+05	1.85E+05	0.423
HSPA5	0.00E+00	0.00E+00	9.72E+04	1.68E+05	0.423
TNC	0.00E+00	0.00E+00	8.66E+04	1.50E+05	0.423
TXNRD1	0.00E+00	0.00E+00	7.00E+04	1.21E+05	0.423
PHB	0.00E+00	0.00E+00	6.03E+04	1.04E+05	0.423
PADI3	4.41E+04	7.63E+04	0.00E+00	0.00E+00	0.423
ANXA2	3.18E+04	5.50E+04	0.00E+00	0.00E+00	0.423
SFPQ	1.77E+04	3.06E+04	0.00E+00	0.00E+00	0.423

Table S-5.1 cont'd.

Gene	Exploded Hair Abundance		Control Hair Abundance		p-Value
	Mean	S.D.	Mean	S.D.	
TGM1	0.00E+00	0.00E+00	2.69E+03	4.66E+03	0.423
MDH2	0.00E+00	0.00E+00	3.68E+02	6.37E+02	0.423
CHD5	1.26E+06	2.18E+06	0.00E+00	0.00E+00	0.423
EPPK1	0.00E+00	0.00E+00	1.74E+05	1.51E+05	0.184
FAM83H	4.55E+05	4.20E+05	3.12E+05	2.81E+05	0.655
KRTAP10-2	0.00E+00	0.00E+00	1.64E+04	2.84E+04	0.423
MYH6	1.35E+05	2.33E+05	0.00E+00	0.00E+00	0.423
NPC1	2.42E+05	4.19E+05	0.00E+00	0.00E+00	0.423
RPL18	1.51E+04	2.62E+04	0.00E+00	0.00E+00	0.423
TUBA4A	0.00E+00	0.00E+00	2.47E+05	4.27E+05	0.423
TUBB4B	0.00E+00	0.00E+00	4.17E+04	7.22E+04	0.423
GSDMA	1.20E+06	1.30E+06	0.00E+00	0.00E+00	0.250
KRT2	7.67E+05	1.33E+06	5.63E+05	9.76E+05	0.842
S100A6	1.47E+06	2.55E+06	0.00E+00	0.00E+00	0.423
RPS3A	4.52E+05	7.82E+05	0.00E+00	0.00E+00	0.423
GSTP1	6.49E+05	1.12E+06	8.90E+05	1.54E+06	0.839
HSPA6	7.24E+04	1.25E+05	0.00E+00	0.00E+00	0.423
RPL13	3.39E+05	5.87E+05	0.00E+00	0.00E+00	0.423
PABPC1	1.51E+05	2.62E+05	0.00E+00	0.00E+00	0.423
PCBP1	3.76E+04	6.52E+04	0.00E+00	0.00E+00	0.423
KRTAP5-1	6.11E+05	1.06E+06	0.00E+00	0.00E+00	0.423
CHUK	7.92E+05	1.37E+06	0.00E+00	0.00E+00	0.423
IL1F10	1.47E+06	1.38E+06	4.09E+06	3.83E+06	0.361
RTN3	3.07E+04	5.32E+04	0.00E+00	0.00E+00	0.423
RELL2	2.16E+05	3.74E+05	0.00E+00	0.00E+00	0.423
AHNAK	0.00E+00	0.00E+00	1.13E+06	1.95E+06	0.423
KRT87P	0.00E+00	0.00E+00	2.66E+05	4.61E+05	0.423
CRAT	0.00E+00	0.00E+00	5.51E+05	9.55E+05	0.423
EDRF1	0.00E+00	0.00E+00	6.56E+05	1.14E+06	0.423
FAHD1	0.00E+00	0.00E+00	3.71E+05	3.21E+05	0.184
ATG9B	0.00E+00	0.00E+00	6.23E+04	1.08E+05	0.423

Table S-5.2. Comprehensive GVP profiles for each single one-inch hair sample, where 0 and 1 denote the presence of the major and minor GVP, respectively, and '--' represents non-detects for both major and minor GVP.

SNP	Exploded Hair			Control Hair				
	1-Ex.G	1-Ex.B.1	1-Ex.B.2	1-Tr.G	1-Tr.B	1-Ctrl.B	2-Tr.B	3-Ctrl.B
KRTAP26-1 rs147862769	--	--	--	0	0	0	--	--
KRT31 rs6503627	--	--	--	--	--	--	1	--
KRTAP16-1 rs72828116	--	0	0	--	0	0	1	--
S100A3 rs36022742	--	0,1	0,1	0,1	0	0,1	--	--
KRTAP10-6 rs62220887	--	1	--	--	1	--	--	--
KRTAP2-3 rs12937861	--	--	1	1	1	1	--	--
KRTAP3-2 rs9897046	0	0	0	0	0	0	0	0
KRTAP11-1 rs71321355	0,1	0,1	0,1	0,1	0,1	0,1	0	0
KRTAP10-12 rs34302939	--	0	0	0	--	0,1	0	0
KRT33A rs12937519	--	0	0	--	0	0	1	--
KRT84 rs2245203	--	1	--	--	--	--	--	--
KRT40 rs9908304	--	--	--	--	--	0	--	--
KRT81 rs2071588	1	1	1	1	1	1	--	--
KRT32 rs2071563	--	0	0	0	0	0	0	0

Table S-5.2 cont'd.

SNP	Exploded Hair			Control Hair				
	1-Ex.G	1-Ex.B.1	1-Ex.B.2	1-Tr.G	1-Tr.B	1-Ctrl.B	2-Tr.B	3-Ctrl.B
KRTAP4-5 rs1497383	0	0	0	0	0	0	0	0,1
GSDMA rs7212938	--	1	1	--	--	--	--	0
KRT36 rs2301354	--	0	0	0	0	0,1	0	--
KRT35 rs2071601	0,1	0,1	0,1	0,1	0,1	0,1	0,1	--
KRTAP4-7 rs11650484	0	0	0	0	0	0	1	0,1
KRTAP10-10 rs4818950	--	0	0	0	0	0	--	--
KRTAP4-11 rs9897031	--	1	1	1	1	1	1	0
TGM3 rs214830	--	1	1	--	--	--	--	--
KRTAP10-11 rs462007	--	--	--	0	--	0	--	1
KRTAP9-1 rs238824	--	--	--	--	--	--	1	--
KRTAP4-1 rs2320231	--	1	1	--	1	1	--	--

REFERENCES

REFERENCES

1. Chu, F.; Anex, D. S.; Jones, A. D.; Hart, B. R., Automated Analysis of Scanning Electron Microscopic Images for Assessment of Hair Surface Damage. *Royal Society Open Science* **2020,** *7* (1), 191438.

2. Szabo, S.; Jaeger, K.; Fischer, H.; Tschachler, E.; Parson, W.; Eckhart, L., In Situ Labeling of DNA Reveals Interindividual Variation in Nuclear DNA Breakdown in Hair and May Be Useful to Predict Success of Forensic Genotyping of Hair. *International Journal of Legal Medicine* **2012,** *126* (1), 63-70.

3. McNevin, D.; Wilson-Wilde, L.; Robertson, J.; Kyd, J.; Lennard, C., Short Tandem Repeat (STR) Genotyping of Keratinised Hair Part 2. An Optimised Genomic DNA Extraction Procedure Reveals Donor Dependence of STR Profiles. *Forensic Science International* **2005,** *153* (2), 247-259.

4. Rice, R. H.; Wong, V. J.; Price, V. H.; Hohl, D.; Pinkerton, K. E., Cuticle Cell Defects in Lamellar Ichthyosis Hair and Anomalous Hair Shaft Syndromes Visualized After Detergent Extraction. *The Anatomical Record* **1996,** *246* (4), 433-441.

5. Rice, R. H., Proteomic Analysis of Hair Shaft and Nail Plate. *Journal of Cosmetic Science* **2011,** *62* (2), 229-236.

6. Kim, Y.-D.; Jeon, S.-Y.; Ji, J. H.; Lee, W.-S., Development of a Classification System for Extrinsic Hair Damage: Standard Grading of Electron Microscopic Findings of Damaged Hairs. *The American Journal of Dermatopathology* **2010,** *32* (5), 432-438.

7. Lee, Y.; Kim, Y.-D.; Hyun, H.-J.; Pi, L.-q.; Jin, X.; Lee, W.-S., Hair Shaft Damage from Heat and Drying Time of Hair Dryer. *Annals of Dermatology* **2011,** *23* (4), 455-462.

8. Milczarek, P.; Zielinski, M.; Garcia, M. L., The Mechanism and Stability of Thermal Transitions in Hair Keratin. *Colloid and Polymer Science* **1992,** *270* (11), 1106-1115.

9. Watt, I. C., Properties of Wool Fibers Heated to Temperatures Above 100°C. *Textile Research Journal* **1975,** *45* (10), 728-735.

10. Takada, K.; Nakamura, A.; Matsuo, N.; Inoue, A.; Someya, K.; Shimogaki, H., Influence of Oxidative and/or Reductive Treatment on Human Hair (I): Analysis of Hair-Damage after Oxidative and/or Reductive Treatment. *Journal of Oleo Science* **2003,** *52* (10), 541-548.

11. Sinclair, R.; Flagler, M. J.; Jones, L.; Rufaut, N.; Davis, M. G., The Proteomic Profile of Hair Damage. *British Journal of Dermatology* **2012,** *166* (s2), 27-32.

12. Ryu, S. R.; Jang, W.; Yu, S.-I.; Lee, B.-H.; Kwon, O.-S.; Shin, K., FT-IR Microspectroscopic Imaging of Cross-Sectioned Human Hair During a Bleaching Process. *Journal of Cosmetics, Dermatological Sciences and Applications* **2016,** *6* (5).

13. Dyer, J. M.; Bell, F.; Koehn, H.; Vernon, J. A.; Cornellison, C. D.; Clerens, S.; Harland, D. P., Redox Proteomic Evaluation of Bleaching and Alkali Damage in Human Hair. *International Journal of Cosmetic Science* **2013,** *35* (6), 555-561.

14. Richena, M.; Rezende, C. A., Effect of Photodamage on the Outermost Cuticle Layer of Human Hair. *Journal of Photochemistry and Photobiology B: Biology* **2015,** *153*, 296-304.

15. Nogueira, A. C. S.; Richena, M.; Dicelio, L. E.; Joekes, I., Photo Yellowing of Human Hair. *Journal of Photochemistry and Photobiology B: Biology* **2007,** *88* (2), 119-125.

16. Dyer, J. M.; Plowman, J. E.; Krsinic, G. L.; Deb-Choudhury, S.; Koehn, H.; Millington, K. R.; Clerens, S., Proteomic Evaluation and Location of UVB-Induced Photo-oxidation in Wool. *Journal of Photochemistry and Photobiology B: Biology* **2010,** *98* (2), 118-127.

17. McMullen, R.; Jachowicz, J., Thermal Degradation of Hair. I. Effect of Curling Irons. *Journal of the Society of Cosmetic Chemists* **1998,** *49* (4), 223-244.

18. Dyer, J. M.; Bringans, S. D.; Bryson, W. G., Characterisation of Photo-oxidation Products Within Photoyellowed Wool Proteins: Tryptophan and Tyrosine Derived Chromophores. *Photochemical & Photobiological Sciences* **2006,** *5* (7), 698-706.

19. Richena, M.; Silveira, M.; Rezende, C. A.; Joekes, I., Yellowing and Bleaching of Grey Hair Caused by Photo and Thermal Degradation. *Journal of Photochemistry and Photobiology B: Biology* **2014,** *138*, 172-181.

20. Chu, F.; Mason, K. E.; Anex, D. S.; Jones, A. D.; Hart, B. R., Hair Proteome Variation at Different Body Locations on Genetically Variant Peptide Detection for Protein-Based Human Identification. *Scientific Reports* **2019,** *9* (1), 7641.

21. Feltz, C. J.; Miller, G. E., An Asymptotic Test for the Equality of Coefficients of Variation from k Populations. *Statistics in Medicine* **1996,** *15* (6), 647-658.

22. Zhang, Y.; Alsop, R. J.; Soomro, A.; Yang, F.-C.; Rheinstädter, M. C., Effect of Shampoo, Conditioner and Permanent Waving on the Molecular Structure of Human Hair. *PeerJ* **2015,** *3*, e1296.

23. Kaliyadan, F.; Gosai, B. B.; Al Melhim, W. N.; Feroze, K.; Qureshi, H. A.; Ibrahim, S.; Kuruvilla, J., Scanning Electron Microscopy Study of Hair Shaft Damage Secondary to Cosmetic Treatments of the Hair. *International Journal of Trichology* **2016,** *8* (2), 94-98.

24. Lee, S. Y.; Choi, A. R.; Baek, J. H.; Kim, H. O.; Shin, M. K.; Koh, J. S., Twelve-Point Scale Grading System of Scanning Electron Microscopic Examination to Investigate Subtle Changes in Damaged Hair Surface. *Skin Research and Technology* **2016,** *22* (4), 406-411.

25. Verma, M. S.; Pratt, L.; Ganesh, C.; Medina, C., Hair-MAP: A Prototype Automated System for Forensic Hair Comparison and Analysis. *Forensic Science International* **2002,** *129* (3), 168-186.

26. Brooks, E.; Comber, B.; McNaught, I.; Robertson, J., Digital Imaging and Image Analysis Applied to Numerical Applications in Forensic Hair Examination. *Science & Justice* **2011,** *51* (1), 28-37.

27. McWhorter, A. S. Development and Evaluation of an Objective Method for Forensic Examination of Human Head Hairs Using Texture-Based Image Analysis. West Virginia University, 2015.

28. Mills, M.; Bonetti, J.; Brettell, T.; Quarino, L., Differentiation of Human Hair by Colour and Diameter Using Light Microscopy, Digital Imaging and Statistical Analysis. *Journal of Microscopy* **2018,** *270* (1), 27-40.

29. Park, K. H.; Kim, H. J.; Oh, B.; Lee, E.; Ha, J., Assessment of Hair Surface Roughness Using Quantitative Image Analysis. *Skin Research and Technology* **2018,** *24* (1), 80-84.

30. Gurden, S. P.; Monteiro, V. F.; Longo, E.; Ferreira, M. M. C., Quantitative Analysis and Classification of AFM Images of Human Hair. *Journal of Microscopy* **2004,** *215* (1), 13-23.

31. Moyo, T.; Bangay, S.; Foster, G. In *The Identification of Mammalian Species Through the Classification of Hair Patterns Using Image Pattern Recognition*, AFRIGRAPH '06 Proceedings of the 4th international conference on computer graphics, virtual reality, visualisation and interaction in Africa, Cape Town, South Africa, Cape Town, South Africa, 2006; pp 177-181.

32. Seiler, H., Secondary Electron Emission in the Scanning Electron Microscope. *Journal of Applied Physics* **1983,** *54* (11), R1-R18.

33. Sutton, M. A.; Li, N.; Joy, D. C.; Reynolds, A. P.; Li, X., Scanning Electron Microscopy for Quantitative Small and Large Deformation Measurements Part I: SEM Imaging at Magnifications from 200 to 10,000. *Experimental Mechanics* **2007,** *47* (6), 775-787.

34. Foracchia, M.; Grisan, E.; Ruggeri, A., Luminosity and Contrast Normalization in Retinal Images. *Medical Image Analysis* **2005,** *9* (3), 179-190.

35. Shinohara, R. T.; Sweeney, E. M.; Goldsmith, J.; Shiee, N.; Mateen, F. J.; Calabresi, P. A.; Jarso, S.; Pham, D. L.; Reich, D. S.; Crainiceanu, C. M., Statistical Normalization Techniques for Magnetic Resonance Imaging. *NeuroImage: Clinical* **2014,** *6*, 9-19.

36. Shan, S.; Gao, W.; Cao, B.; Zhao, D. In *Illumination Normalization for Robust Face Recognition Against Varying Lighting Conditions*, 2003 IEEE International SOI Conference Proceedings, 17-17 Oct. 2003; 2003; pp 157-164.

37. Du, S.; Ward, R. In *Wavelet-Based Illumination Normalization for Face Recognition*, IEEE International Conference on Image Processing 2005, 14-14 Sept. 2005; 2005; pp II-954.

38. Schneider, C. A.; Rasband, W. S.; Eliceiri, K. W., NIH Image to ImageJ: 25 Years of Image Analysis. *Nature Methods* **2012,** *9*, 671.

39. Schindelin, J.; Rueden, C. T.; Hiner, M. C.; Eliceiri, K. W., The ImageJ Ecosystem: An Open Platform for Biomedical Image Analysis. *Molecular Reproduction and Development* **2015,** *82* (7-8), 518-529.

40. Jain, A.; Nandakumar, K.; Ross, A., Score Normalization in Multimodal Biometric Systems. *Pattern Recognition* **2005,** *38* (12), 2270-2285.

41. Davies, P. J.; Horrocks, A. R.; Miraftab, M., Scanning Electron Microscopic Studies of Wool/Intumescent Char Formation. *Polymer International* **2000,** *49* (10), 1125-1132.

42. Song, D.; Wang, J., Modified Resolution Factor for Asymmetrical Peaks in Chromatographic Separation. *Journal of Pharmaceutical and Biomedical Analysis* **2003,** *32* (6), 1105-1112.

43. Parker, G. J.; Leppert, T.; Anex, D. S.; Hilmer, J. K.; Matsunami, N.; Baird, L.; Stevens, J.; Parsawar, K.; Durbin-Johnson, B. P.; Rocke, D. M.; Nelson, C.; Fairbanks, D. J.; Wilson, A. S.; Rice, R. H.; Woodward, S. R.; Bothner, B.; Hart, B. R.; Leppert, M., Demonstration of Protein-Based Human Identification Using the Hair Shaft Proteome. *PLoS ONE* **2016,** *11* (9), e0160653.

44. Langbein, L.; Rogers, M. A.; Winter, H.; Praetzel, S.; Beckhaus, U.; Rackwitz, H.-R.; Schweizer, J., The Catalog of Human Hair Keratins: I. Expression of the Nine Type I Members in the Hair Follicle. *Journal of Biological Chemistry* **1999,** *274* (28), 19874-19884.

45. Rogers, M. A.; Langbein, L.; Winter, H.; Ehmann, C.; Praetzel, S.; Korn, B.; Schweizer, J., Characterization of a Cluster of Human High/Ultrahigh Sulfur Keratin-Associated Protein Genes Embedded in the Type I Keratin Gene Domain on Chromosome 17q12-21. *Journal of Biological Chemistry* **2001,** *276* (22), 19440-19451.

46. Rogers, M. A.; Winter, H.; Beckmann, I.; Schweizer, J.; Langbein, L.; Praetzel, S., Hair Keratin Associated Proteins: Characterization of a Second High Sulfur KAP Gene Domain on Human Chromosome 21. *Journal of Investigative Dermatology* **2004,** *122* (1), 147-158.

47. Schweizer, J.; Langbein, L.; Rogers, M. A.; Winter, H., Hair Follicle-Specific Keratins and Their Diseases. *Experimental Cell Research* **2007,** *313* (10), 2010-2020.

48. Moll, R.; Divo, M.; Langbein, L., The Human Keratins: Biology and Pathology. *Histochemistry and Cell Biology* **2008,** *129* (6), 705.

49. Morioka, K., A Guide to Hair Follicle Analysis by Transmission Electron Microscopy: Technique and Practice. *Experimental Dermatology* **2009,** *18* (7), 577-582.

50. Mesler, A. L.; Veniaminova, N. A.; Lull, M. V.; Wong, S. Y., Hair Follicle Terminal Differentiation Is Orchestrated by Distinct Early and Late Matrix Progenitors. *Cell Reports* **2017,** *19* (4), 809-821.

51. Stadtman, E. R.; Van Remmen, H.; Richardson, A.; Wehr, N. B.; Levine, R. L., Methionine Oxidation and Aging. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **2005,** *1703* (2), 135-140.

52. Zhang, X.; Monroe, M. E.; Chen, B.; Chin, M. H.; Heibeck, T. H.; Schepmoes, A. A.; Yang, F.; Petritis, B. O.; Camp, D. G.; Pounds, J. G.; Jacobs, J. M.; Smith, D. J.; Bigelow, D. J.; Smith, R. D.; Qian, W.-J., Endogenous 3,4-Dihydroxyphenylalanine and Dopaquinone Modifications on Protein Tyrosine. *Molecular & Cellular Proteomics* **2010,** *9* (6), 1199-1208.

53. Jeong, J.; Jung, Y.; Na, S.; Jeong, J.; Lee, E.; Kim, M.-S.; Choi, S.; Shin, D.-H.; Paek, E.; Lee, H.-Y.; Lee, K.-J., Novel Oxidative Modifications in Redox-Active Cysteine Residues. *Molecular & Cellular Proteomics* **2011,** *10* (3), M110.000513.

54. Langbein, L.; Eckhart, L.; Rogers, M. A.; Praetzel-Wunder, S.; Schweizer, J., Against the Rules: Human Keratin K80: Two Functional Alternative Splice Variants, K80 and K80.1, with Special Cellular Localization in a Wide Range of Epithelia. *Journal of Biological Chemistry* **2010,** *285* (47), 36909-36921.

55. Kitts, A.; Sherry, S., The Single Nucleotide Polymorphism Database (dbSNP) of Nucleotide Sequence Variation. In *The NCBI Handbook*, McEntyre, J.; Ostell, J., Eds. National Center for Biotechnology Information (US): Bethesda, MD, 2002 [Updated 2011].

CHAPTER 6: Conclusions and Broader Impacts

6.1 Conclusions and Broader Impacts

The findings described in this dissertation have concurrently demonstrated success detecting genetically variant peptides in hair and addressed knowledge gaps in hair protein chemistry across a spectrum of conditions in which hair evidence can be obtained for forensic analysis. Leveraging an integrated exome sequence and mass spectrometry-based workflow, GVP detection in single one-inch hairs from different body locations attains similar power to differentiate individuals, aged hairs up to a year of hair growth have shown minimal differences in GVP abundance levels, and damaged hairs recovered after an explosive blast exhibit similar GVP detection rates as their undamaged counterparts. Furthermore, the advancements in GVP analysis from this work include extended capabilities to co-detect GVPs from hair proteins and SNPs from mtDNA. In addition, this research has proposed an objective framework for quantifying morphological hair damage using scanning electron microscopic images. These capabilities benefit both the forensic and medical sciences, providing increased information content and specificity in discriminative potential from co-detection of variant markers.

But what challenges remain before implementing this protein-based approach in forensic casework? As the technologies for GVP analysis mature, the need for data-independent mass spectrometry approaches to support protein-based human identification has become increasingly apparent because false negatives result from data-dependent proteomic protocols. Targeted data-independent methods are expected to improve reproducibility in GVP detection, thereby facilitating validation of GVP candidates for inclusion into a human identification marker panel. However, during this period of discovery and method development, data-dependent and data-independent approaches should both be utilized in tandem to identify and evaluate potential GVP

candidates. This should include use of exclusion lists of previously identified candidates for data-dependent analysis discoveries with subsequent analyses of the same hair samples, which enables deeper interrogation of the hair proteome for identification of novel GVP markers. Analysis of hair samples from a larger group of individuals, from different ancestries, is expected to expand the breadth of novel GVPs. Novel markers can then be validated using peptide standards, synthesized peptide sequences that include any associated chemical modifications, to confirm the identity of candidate GVP markers detected via mass spectrometry and selected for inclusion in a GVP panel.

The path forward for GVP analysis of forensic evidence does not rely on developments in analytical tools alone. As a multi-disciplinary endeavor interpretation of GVP profiling as applicable to the end-user, i.e., forensic analysts, depends on a more complete understanding of population genetics, advances in tools to examine genetic variation, and use of appropriate statistical methods to support biological observations.

Mutation frequencies within a population comprise an important component for quantification of discriminative power from GVP analysis, but true SNP frequencies may differ from what has been documented; the current sampled population, 141,456 individuals from the Genome Aggregation Database[1], is likely not representative of the global population. In such case, assessment of reliability, e.g., using confidence intervals, to quantify discriminative potential may be poorly estimated. Public databases that curate mutation frequencies continue to be updated as the population of individuals included in clinical studies of disease states grows. As greater efforts are expended towards expanding profiling of individuals' genotypes, a more complete measure of discriminative potential using GVPs is anticipated.

Assessment of GVP marker independence also plays a large role in development of a GVP panel and affects quantification of discriminative potential, but the process is somewhat nebulous. The quantitative approach used in this work, random match probability, requires independent GVP markers or that the inferred SNP markers are not in linkage disequilibrium, or co-inherited from a single ancestral chromosome, as linked SNPs occur more frequently than expected in a population[2]. Elucidation of linked SNP pairs has been limited owing to availability of only a small population of phased genomes, namely from the 1000 Genomes Project, which has curated genetic variants from 2,504 individuals[3], and few accessible statistical tools to examine linkage disequilibrium. Phased genomes, which contain allele assignments to the maternal or paternal chromosomes, provide the basis for assessing whether SNPs are linked; only SNP pairs lying on the same chromosome have the potential to be linked. Quantitative metrics that correlate the presence or absence of SNP pairs within a population have been used conventionally to evaluate the extent to which SNPs are linked[4, 5], and the Linkage Disequilibrium Calculator hosted on ENSEMBL offers one publicly available tool to calculate these metrics[6], but there exists little guidance on metric thresholds for classifying linked SNPs from unrelated markers. GVP profile interpretations described herein assumed a one-SNP-per-gene rule to minimize disequilibria and outlined a general approach for forensic application. However, it is expected that with a better understanding of which SNPs are linked, through availability of phased genomes from a larger population, and further development of publicly available statistical tools and rigorous methods to evaluate linkage, clarification on the extent of marker independence and how to evaluate it can be obtained, which will guide selection of analytically-validated GVP markers.

A third consideration to GVP profile interpretation lies in the need for statistical validity, but no consensus has been reached in the forensic community regarding an approach for assessment of discriminative potential, despite heavy reliance upon statistics to provide interpretations regarding discriminative power of DNA evidence. Random match probability was selected for use in this work as a simple and more established concept in forensic investigations, though a different method, likelihood ratio, is available, and may be more suited for GVP analysis, as implemented in routine forensic investigations, to quantify discriminative potential. As an alternative to random match probability, likelihood ratios evaluate evidentiary strength by quantifying two contrasting hypotheses: the odds that a match did not occur at random over the probability that the same match occurred by random chance[7, 8]. Likelihood ratios can be converted to random match probabilities, but the reverse would be statistically meaningless since random match probability only quantifies the latter hypothesis[7]. Applicability of likelihood ratio to other evidence types, including genetic profiles and trace evidence[8], and their inferential property to combine evidentiary strength as the product of likelihood ratios for each evidence type[9], provide a more versatile and rigorous statistical approach than random match probability. However, some challenges that use of likelihood ratios encounters include development of methods and metrics to quantify the odds for both hypotheses, which may vary depending on the evidence type, and facilitating presentation of this statistical approach in a courtroom setting in a comprehensible manner; both are topics of ongoing research and discussion in the forensic community. Applied to co-detection of SNPs from mtDNA and GVPs from hair proteins, this statistical approach represents a more appropriate method to produce a combined discriminative potential from both types of evidence. Given these advantages and growing research in this area,

profile interpretation from GVP analysis would benefit from use of likelihood ratios and integrate well with current forensic analyses of other evidence types.

With the advances outlined above, how should GVP analysis be implemented in practice? The similarities between GVP and conventional STR analyses facilitate application of GVP markers for routine forensic operation, and GVP analysis fills the gap left by limitations of DNA recovery in hair evidence. Paralleling STR markers in DNA evidence test kits, the culmination of GVP marker validation and selection should include a list of GVP markers to be detected in GVP analyses of evidence and reference samples, an optimized mass spectrometry workflow for marker detection, guidelines and reagents for hair sample preparation, and peptide standard equivalents of GVP markers for quality control and assurance, perhaps with standard reference peptides certified by the National Institute of Standards and Technology. Subsequent processes for assembly of GVP profiles and downstream interpretation should also follow a similar path to STR analysis. However, unique to GVP analysis, the inferential properties from GVP to SNP and *vice versa*, via the genetic code from DNA to protein, offer versatility in profile comparison to match GVPs from recovered forensic evidence to either SNPs or GVPs from suspects or DNA databases. One route entails comparison of GVP profiles from recovered hair evidence to GVP profiles obtained from a suspect's hair or curated in a database, to evaluate the extent of a match. An alternative approach utilizes intact DNA from a suspect's blood sample or from exome sequences stored in a DNA database to compare with GVP profiles from recovered evidence. Advantages of this approach include GVP analysis success even when hair reference samples are not readily available and greater matching success by comparing to a larger pool of individuals' profiles within a database. Given the growing clinical research on SNPs using exome sequences, creation of DNA databases containing exome sequences is expected to receive broader interest,

which will facilitate GVP analysis and profile comparison of recovered forensic evidence when there are no suspects. Identification of SNPs carried by the suspect or the individual contributing the DNA in a database entry via exome sequencing and subsequent conversion of SNPs to GVPs allow comparison to GVP profiles from recovered evidence to determine the extent of a match. The work described herein primarily used exome sequencing as a research and development tool for discovery of a broad range of GVPs, but exome sequencing also finds utility in GVP analysis in this manner.

While the need for advances in bioanalytical tools, population genetics, and statistics addresses the technical readiness aspects of GVP technologies, current capabilities in forensic laboratories must also be adapted to integrate GVP identification into the analytical scheme. Chiefly, analysis and detection of GVPs requires, at a minimum, liquid chromatography-tandem mass spectrometry capabilities. Forensic laboratories that perform toxicological analyses are more likely to meet this requirement as LC-MS/MS techniques have been more widely adopted in recent years. For example, LC-MS/MS approaches provide versatility and more confidence in confirming the presence of non-volatile or thermally labile drugs and their metabolites in blood and urine over traditional gas chromatography-mass spectrometry methods[10]. Existing LC-MS/MS instrumentation can be configured to include GVP analysis using test kit components. Although the work presented in this dissertation utilized nano-LC-MS/MS with an Orbitrap mass analyzer, the high mass resolution, while beneficial, is not a critical component of GVP analysis in practice, given a list of well-characterized and distinct GVP targets. Therefore, lower mass resolution mass analyzers such as quadrupole-time-of-flights and triple quadrupoles, which are more prevalent in forensic laboratories with LC-MS/MS instrumentation, suffice to accommodate GVP analysis. As forensic laboratories continue to expand their analytical

capabilities, GVP analysis is expected to be more accessible, which will drive more widespread adoption of this approach in forensic science. Collectively, these aspects outline the translation of GVP analysis from the laboratory to be implemented, strengthened, and gain acceptance in the forensic community.

This research not only represents an advancement in forensic proteomics, a rapidly growing discipline, but may also form the basis for development in other fields. Not limited to forensics or proteomics, investigation of GVPs including GVPs from other biological matrices, may offer an avenue to elucidate mechanisms of diseases associated with nonsynonymous SNPs where protein functions may be affected, with potential application in structural biology and the clinical sciences.

REFERENCES

REFERENCES

1. Karczewski, K. J.; Francioli, L. C.; Tiao, G.; Cummings, B. B.; Alföldi, J.; Wang, Q.; Collins, R. L.; Laricchia, K. M.; Ganna, A.; Birnbaum, D. P.; Gauthier, L. D.; Brand, H.; Solomonson, M.; Watts, N. A.; Rhodes, D.; Singer-Berk, M.; England, E. M.; Seaby, E. G.; Kosmicki, J. A.; Walters, R. K.; Tashman, K.; Farjoun, Y.; Banks, E.; Poterba, T.; Wang, A.; Seed, C.; Whiffin, N.; Chong, J. X.; Samocha, K. E.; Pierce-Hoffman, E.; Zappala, Z.; O'Donnell-Luria, A. H.; Minikel, E. V.; Weisburd, B.; Lek, M.; Ware, J. S.; Vittal, C.; Armean, I. M.; Bergelson, L.; Cibulskis, K.; Connolly, K. M.; Covarrubias, M.; Donnelly, S.; Ferriera, S.; Gabriel, S.; Gentry, J.; Gupta, N.; Jeandet, T.; Kaplan, D.; Llanwarne, C.; Munshi, R.; Novod, S.; Petrillo, N.; Roazen, D.; Ruano-Rubio, V.; Saltzman, A.; Schleicher, M.; Soto, J.; Tibbetts, K.; Tolonen, C.; Wade, G.; Talkowski, M. E.; Neale, B. M.; Daly, M. J.; MacArthur, D. G., Variation Across 141,456 Human Exomes and Genomes Reveals the Spectrum of Loss-of-Function Intolerance Across Human Protein-Coding Genes. *bioRxiv* **2019**, 531210.

2. Ardlie, K. G.; Kruglyak, L.; Seielstad, M., Patterns of Linkage Disequilibrium in the Human Genome. *Nature Reviews Genetics* **2002,** *3* (4), 299-309.

3. The 1000 Genomes Project Consortium, A Global Reference for Human Genetic Variation. *Nature* **2015,** *526*, 68-74.

4. Slatkin, M., Linkage Disequilibrium - Understanding the Evolutionary Past and Mapping the Medical Future. *Nature Reviews Genetics* **2008,** *9*, 477-485.

5. Stephens, J. C.; Schneider, J. A.; Tanguay, D. A.; Choi, J.; Acharya, T.; Stanley, S. E.; Jiang, R.; Messer, C. J.; Chew, A.; Han, J.-H.; Duan, J.; Carr, J. L.; Lee, M. S.; Koshy, B.; Kumar, A. M.; Zhang, G.; Newell, W. R.; Windemuth, A.; Xu, C.; Kalbfleisch, T. S.; Shaner, S. L.; Arnold, K.; Schulz, V.; Drysdale, C. M.; Nandabalan, K.; Judson, R. S.; Ruaño, G.; Vovis, G. F., Haplotype Variation and Linkage Disequilibrium in 313 Human Genes. *Science* **2001,** *293* (5529), 489.

6. Hunt, S. E.; McLaren, W.; Gil, L.; Thormann, A.; Schuilenburg, H.; Sheppard, D.; Parton, A.; Armean, I. M.; Trevanion, S. J.; Flicek, P.; Cunningham, F., Ensembl Variation Resources. *Database* **2018,** *2018*.

7. National Research Council, U. S. A., *The Evaluation of Forensic DNA Evidence*. National Academies Press: 1996.

8. Meuwly, D.; Ramos, D.; Haraksim, R., A Guideline for the Validation of Likelihood Ratio Methods Used for Forensic Evidence Evaluation. *Forensic Science International* **2017,** *276*, 142-153.

9. Nordgaard, A.; Rasmusson, B., The Likelihood Ratio as Value of Evidence—More than a Question of Numbers. *Law, Probability and Risk* **2012,** *11* (4), 303-315.

10. Mbughuni, M. M.; Jannetto, P. J.; Langman, L. J., Mass Spectrometry Applications for Toxicology. *EJIFCC* **2016,** *27* (4), 272-287.