Title

Age-Related Changes in Hair Shaft Protein Profiling and Genetically Variant Peptides

Authors
Plott, Tempest J
Karim, Noreen
Durbin-Johnson, Blythe P
et al.

Publication Date
2020-07-01

DOI
10.1016/j.fsigen.2020.102309

Peer reviewed

Age-Related Changes in Hair Shaft Protein Profiling and Genetically Variant Peptides

Tempest J. Plott[ab1], Noreen Karim[b1], Blythe P. Durbin-Johnson[c], Dionne P. Swift[d],

R. Scott Youngquist[d], Michelle Salemi[e], Brett S. Phinney[e], David M. Rocke[c], Michael G. Davis[d],

Glendon J. Parker[ab2], Robert H. Rice[ab2]

[a]Forensic Science Graduate Program, University of California, Davis, CA, USA

[b]Department of Environmental Toxicology, University of California, Davis, CA, USA

[c]Division of Biostatistics, Department of Public Health Sciences, Clinical and Translational Science Center Biostatistics Core, University of California, Davis, CA, USA

[d]Procter & Gamble, Mason Business Center, Mason, OH, USA

[e]Proteomics Core Facility, University of California, Davis, CA, USA

[1]These authors contributed equally

[2]These authors contributed equally

Keywords

Proteomic profiling, genetically variant peptides, human hair, ageing, forensic investigation

Acknowledgments

This work was supported by NIH grant UL1 TR001860 from the National Center for Advancing Translational Sciences, NIJ grants 2011-DN-BX-K543 and 2015-DN-BX-K065, USDA (NIFA)/University of California Agricultural Experiment Station project CA-D-ETX-2152-H, and a research contract from Procter & Gamble. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests

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

Abstract

Recent reports highlight possible improvements in individual identification using proteomic information from human hair evidence. These reports have stimulated investigation of parameters that affect the utility of proteomic information. In addition to variables already studied relating to processing technique and anatomic origin of hair shafts, an important variable is hair ageing. Present work focuses on the effect of age on protein profiling and analysis of genetically variant peptides (GVPs). Hair protein profiles may be affected by developmental and physiological changes with age of the donor, exposure to different environmental conditions and intrinsic processes, including during storage. First, to explore whether general trends were evident in the population at different ages, hair samples were analyzed from groups of different subjects in their 20's, 40's and 60's. No significant differences were seen as a function of age, but consistent differences were evident between European American and African American hair profiles. Second, samples collected from single individuals at different ages were analyzed. Mostly, these showed few protein expression level differences over periods of 10 years or less, but samples from subjects at 44 and 65 year intervals were distinctly different in profile. The results indicate that use of protein profiling for personal identification, if practical, would be limited to decadal time intervals. Moreover, batch effects were clearly evident in samples processed by different staff. To investigate the contribution of storage (at room temperature) in affecting the outcomes, the same proteomic digests were analyzed for GVPs. In samples stored over 10 years, GVPs were reduced in number in parallel with the yield of identified proteins and unique peptides. However, a very different picture emerged with respect to personal identification. Numbers of GVPs sufficed to distinguish individuals despite the age differences of the samples. As a practical matter, three hair samples per person provided nearly the maximal number obtained from 5 or 6 samples. The random match probability (where the log increased in proportion to the number of GVPs) reached as high as 1 in 10^8. The data indicate that GVP results are dependent on the single nucleotide polymorphism profile of the donor genome, where environmental/processing factors affect only the yield, and thus are consistent despite the ages of the donors and samples and batchwise effects in processing. This conclusion is critical for application to casework where the samples may be in storage for long periods and used to match samples recently collected.

Introduction

Protein profiling (comparison of relative protein expression levels) and proteomic genotyping (inferring single nucleotide polymorphisms in the genome using the proteome) for human hair comparison and individual identification have shown promise as potential tools for forensic investigation. For example, large inter-individual differences in protein profile are evident in hair shafts (Laatsch et al, 2014). Studies using human twins (Wu et al, 2017) support the conclusion reached using inbred mouse strains (Rice et al, 2012) that differences in profile have primarily a genetic basis. Corneocyte proteins of the hair shaft (Wu et al, 2017), epidermis (Borja et al, 2019) and appendages provide an even more direct connection to genotype in their reflection of individual allelic differences in the genome. Thus, detection of genetically variant peptides (GVPs) containing single amino acid polymorphisms (SAPs) that could be matched to single nucleotide polymorphisms (SNPs) in the coding region of the genome provides a more discriminating way to infer the genotype and even ancestry of the donor (Parker et al, 2016).

From a forensic perspective, limitations on the use of samples for such identifications are important to know. For example, recent findings show that the hair shaft is equally useful for profiling or GVP analysis regardless of its state of pigmentation (Parker et al, 2019) or anatomic site of origin (Chu et al, 2019; Milan et al, 2019), although GVP analysis can offer much greater discrimination. A property that remains to be examined is the reproducibility of such samples with age of donor or period of storage. This issue is pertinent because the protein content of samples may change with the age of the donor at collection, and casework samples are often in storage for many years. Thus, investigators are likely to compare samples from individuals at different ages and originating many years apart.

First, to determine whether global changes in hair are evident with age, present work compares protein profiles in samples from groups of individuals of different age. Samples collected at roughly the same time are compared from American females in their 20’s, 40’s and 60’s from European and African backgrounds, also permitting investigation of the role of ethnic origin. Second, to examine changes in hair from individuals over time, samples were compared in protein profile and GVP content from 9 subjects at age intervals of 4 to 65 years. The results of both studies are presented and reconciled.

MATERIALS AND METHODS

Sample collection

For analysis of samples from different age groups, hair was collected by a commercial supplier from 30 African Americans (10 each of ages 20, 40, 60) and 40 European Americans (20 of age 20 and 10 each of ages 40 and 60), all female (Cohort 1). Samples are referred to as "African" or "European" for simplicity. One sample from each donor was analyzed. To find the effect of age on individuals, a second set of samples that had been collected at different times (stored at room temperature) from nine individuals (A – E (Cohort 2) and F-I (Cohort 3), total three females and six males), each analyzed in sets of 2-6 replicates (**Table S1)**. According to donors, the hair was not chemically treated (dyed, bleached, straightened). These samples were collected with informed consent approved by the University of California Davis Institutional Review Board (protocol 896494) and processed within a year.

Sample processing for protein isolation and mass spectrometry

In each case, aliquots of 4 mg were processed essentially as previously described (Laatsch et al, 2014) except for using 0.05 M ammonium bicarbonate instead of 0.1 M sodium phosphate buffer during reduction and alkylation. Each cohort of samples was processed at a different time by a different investigator. Hair protein digests from the age groups and from individuals were randomized and analyzed by LC-MS/MS on a Thermo Scientific Q Exactive Plus Orbitrap mass spectrometer essentially as previously described (Wu et al, 2017).

Database searching and proteomic profiling based on weighted spectral counts and statistical analysis

Data files generated for the samples of age groups (Cohort 1) and the individuals A-E (Cohort 2) were analyzed using X!Tandem (2016.10.15.2) to search a Uniprot human database with an appended database of common human contaminants and an appended identical but reversed (decoy) peptide database for estimating false discovery rates. The proteomics data are available in the MassIVE repository as #MSV000085030, Proteome Exchange #PXD017771 (https://massive.ucsd.edu/ProteoSAFe/dataset.jsp?task=4a43733eab0c45a0a78a7afc7ad4f685). Also, the data from Cohorts 2 and 3 have been deposited to the ProteomeXchange Consortium via the PRIDE (Perez-Riverol et al, 2019) partner repository with the dataset identifier PXD016169. Scaffold (version 4.8.2) was used to validate peptide and protein identifications. Accepted protein identifications contained at least 2 identified peptides. False discovery rates were estimated as 0.1% and 2.9% for peptides and proteins, respectively. The MS results were analyzed as weighted spectral counts (with clusters containing shared peptides) after removal of entries not genuinely present judging by their exclusive peptides. Differential protein abundance analyses were conducted using the limma-voom Bioconductor pipeline, originally developed for analysis of RNA-Seq data and applied here to weighted spectral counts (Ritchie et al, 2015). Standard errors of estimates were adjusted for correlation between replicates from the same sample; subject was included as a fixed effect in all models. The R code is provided in supplemental files.

Protein profiling using PEAKS

Label-free quantitation was performed on the LC-MS/MS datasets of individuals A-I (Cohorts 2 and 3) using PEAKS Studio 10.0 (Bioinformatics Solutions Inc., Waterloo, ON, Canada) to obtain their protein profiles (Zhang et al, 2012). From 2 - 6 samples for each age from all nine individuals amounting to a total of 67 datasets were analyzed against a validated UNIPROT human reference proteome (uniprot-proteome_UP000005640_Human). Default settings of the algorithm were employed except that the precursor mass error range and fragment ion were set to 10 ppm and 0.04 Da, respectively. Cysteine carbamidomethylation (+57 Da) was set as a fixed post translational modification, while deamidation on glutamines and asparagines (+0.98 Da), oxidation of histidines, tryptophan, and methionine (+15.99 Da), dioxidation of methionines (+29.99 Da), pyroglutamation at glutamines (-17.02 Da) and glutamates (-18.01 Da), and acetylation (+42.01) and formylation (+27.99) of N-termini and lysines were variable modifications. The resulting datasets, filtered with a 1% false discovery rate, were analyzed using the Q-module function of PEAKS Studio, and a heat map was generated by label free quantitation for proteins with at least 2 fold difference in the levels among the groups and a significance of 13 (p value = 0.05; -10log(0.05) = 13.01). Due to batch effects identified by comparing profiles of the most recent samples of Cohorts 2 and 3 **(Figure S1)** a collective comparison of the profiles of individuals A-I was not performed.

GVP analysis

The data files of the nine individuals (A-I) sampled at different ages were searched to generate GVP profiles to determine whether the individuals could be distinguished from each other by this criterion. For GVP analysis, raw data files were submitted to X!Tandem peptide spectra matching algorithm (Global Proteome Machine Fury, X!Tandem Alanine 149 (2016.10.15.2)) after conversion to MzML format by MSConvertGUI (Proteowizard 2.1 http://proteowizard.sourceforge.net). Default search parameters of the algorithm were used except that the virus and prokaryote reference libraries were excluded and point mutations were included in the search. Protein and peptide log(e) scores of -1, and fragment and parent mass error of 20 ppm and 100 ppm, respectively, were used. The files generated by X!Tandem (.XML, thegpm.org) were used to obtain the peptide data, which was then provided to/pasted into GVP Finder (Goecker et al, 2019). From the list of putative GVPs, unique tryptic peptides carrying log(e) scores of < -2 were used for GVP profiling if they displayed no other genetic or chemical modifications (except N/Q deamidation, methionine oxidation, cysteine carboxymethylation and N-terminal acetylation) and, if corresponding to a minor allele, with no major fragmentation masses corresponding to the reference alleles. The GVPs observed in the current study were not validated by DNA sequencing. However, the previously observed rate of false positive identifications of 1.5-2% (Borja et al, 2019; Parker et al, 2016) using the employed method provides high confidence in the GVP profiles. The mass spectrometry proteomics data from Cohorts 2 and 3 have been deposited to the ProteomeXchange Consortium via the PRIDE (Perez-Riverol et al, 2019) partner repository.

Random match probability calculation

Random match probabilities (RMPs) were calculated for the GVP profile of each sample using the genotype frequencies of the identified loci from the 1000 Genomes Project Consortium (2015). As all the studied subjects in Cohorts 2 and 3 were of European origin, only European genotype frequencies were used for estimation of RMP. For the calculation, each SNP was treated as independent except the multiple GVPs/alleles from one gene that were treated as one locus. The frequency for the allele combination was then used to estimate the RMPs. The product rule was applied to calculate the RMP for each specific GVP profile (Parker et al, 2016).

Hierarchical clustering

For statistical analysis, all the GVPs detected in the biological replicates were collated. GVPs detected in one or more replicates were given the same weight. All the detections were assigned the value “1”, and those that were not detected in the samples were assigned the value “0”. GVPs that were either detected or not detected throughout the samples (and thus were without probative value) were excluded from the analysis. Agglomerative hierarchical clustering with complete linkage was performed based on the Euclidean distance data for the samples, and a dendrogram for the clustering was plotted using the hclust function of R (Version 3.6.2) (Milan et al, 2019).

RESULTS

Hair proteome comparison among age groups

To study the effect of age and ethnicity on the hair proteome, hair samples from European-American and African Americans of three age groups (20s, 40s, 60s) were studied. The data were analyzed against the Uniprot human database using X!Tandem (2016.10.15.2) and peptide and protein identifications were validated using Scaffold (version 4.8.2). The weighted spectral counts of 241 proteins were used for analyzing pairwise differences in protein profile. As illustrated in **Table 1**, significant pair-wise differences were not detected in different age groups within each ethnic category or within the ethnic groups of combined ages. However, some significant differences between samples from African-American and European-American subjects were discernable (**Figure 1**). Proteins higher in the African samples included TYRP1 (Tyrosinase Related Protein 1) and GPNMB (Glycoprotein Nonmetastatic Melanoma Protein B), which participate in melanin biosynthesis (Kobayashi et al, 1998; Zhang et al, 2012), and are a reflection of the higher melanin content in samples from the African-American cohort. In addition, certain keratins (i.e., KRTs 1, 2, 5, 9, 10, 24) were among the proteins higher in level in the African samples. Two proteins involved in membrane lipid metabolism, PLD3 (Gonzalez et al, 2018) and LPCAT3 (Rong et al, 2015), were higher in the European hair samples. As the cuticle cells are bounded by a protein membrane surrounded by lipids (Dias, 2015), the higher number of cuticle layers in the European compared to African samples could contribute to the differences in level of these hair proteins in the two populations. Other proteins higher in the European samples are involved in autophagy (HSP90AA1, ATG9b), ribosomal function (RPS2, EEF1D), and calcium binding (CALML5). The overall data obtained from Cohort 1 identified no consistent proteomic differences in hair shafts as a function of age in the range of 20 to 60 years. Likewise, the lack of overall proteomic differences precludes the possibility of global changes in GVP profile as a function of age. Importantly, however, the data do not exclude the possibility that age-related changes in protein abundance are not detected due to compensating individual variation over time.

Proteomic profile comparisons at different ages in given individuals based on weighted spectral counts

Because a lack of differences in the hair proteome as a function of age in unrelated individuals could be attributed to compensating individual variation, a complementary analysis was also conducted on recent hair samples and those that had been stored over 4 to 65 years from 9 individuals (Supplementary Table S1). Two different groups of subjects (Individuals A-E in Cohort 2 and Individuals F to I in Cohort 3) were analyzed. For the first longitudinal study, proteomic datasets from hair shafts from 5 individuals were processed, and significant differences in pair-wise protein abundances among a total of 211 proteins were tabulated. As shown in **Table 2**, data from three subjects (A, D, and E) showed few protein differences (0-6) with age in two-way comparisons over periods of 4-11 years. Samples from one subject (C) showed few differences (5-7) over a span of 6 years, but a substantial number (27) over 11 years. One subject (B) showed a substantial number of differences (32) over a span of 65 years. As shown in **Figure 2**, the protein profiles from a single subject at different ages were much closer in distance than the profiles among different individuals. The data in **Table 2** indicated that subjects D and E could be readily distinguished from all the other subjects, but some subject combinations would be more difficult (e.g., A0 or A6 versus C6 or C11). Also the subjects B and C had high levels of internal differences, but these were consistent with longer time frames, a 65 year storage time for subject B and an 11 year difference for subject C. Storage time of the hair sample may have contributed to these differences in protein profiling, although physiological changes due to subject aging cannot be excluded.

Proteomic profile comparisons at different ages among individuals based on heatmaps

An additional batch of hair samples (Cohort 3) was processed to expand the number of longitudinal samples. The resulting proteomic profiles were bioinformatically processed to obtain label free quantitation and subsequent heat maps using Q-module in the PEAKs™ software package (version 10.0) (Zhang et al, 2012). The samples were divided into two groups, new (recent samples) and old (collected 7 or more years before present) based on the time since collection. As can be seen in **Figure 3A**, when protein profiles were filtered based on a 2-fold change and p-value of 0.05, little difference was seen in the proteomes of older and recent samples when compared collectively. Only 3 protein differences were detected, one of which, KRTAP7-1, was a structural protein and one, SEC23B, is involved in endosomal transport and was significantly increased in pigmented hair (Parker et al, 2019). The low number of significant differences, again, could be attributed to the higher variation in proteomic profiles from individual to individual that could cancel statistically significant effects. Another analysis was therefore conducted on the most extreme case, individual I, with a 44 year gap in subject age. Samples from this individual showed 54 proteins that had a 2-fold change in abundance (p=0.05) (**Figure 3B**) with fifty proteins higher in level in the recent samples compared to the older ones. These included proteins reported to be concentrated in the cuticle (S100A3, KRT40, KRT82, KRTAP16-1, 24-1, and 3-2) among other hair KRTs and KRTAPs (http://www.proteinatlas.org; (Moll et al, 2008; Uhlén et al, 2015). The higher amounts of cuticle concentrated proteins in the recent samples could reflect the loss of cuticle in the older samples (Thibaut et al, 2010). Four of the proteins were higher in level in the older samples, SYNE2 (cytoskeletal protein), AKAP9 (scaffolding protein), and GFAP (an intermediate filament protein) (http://www.proteinatlas.org). A similar analysis from individuals F, G, and H showed considerably fewer proteomic changes over a period of 7 years with 2, 13, and 4 proteins respectively, differing among the stored and recent samples.

Genetically variant peptide analysis

To determine the effect of potential sample degradation with storage, GVPs in each sample were first identified and evaluated. The total number of unique peptides was also measured in each proteomic dataset. Sample storage/age was not seen to affect the average number of identified unique peptides in the samples over periods of <10 years (**Figure 4A**). However, decreases of ~38, 27, and 33% of the unique peptides, relative to their corresponding recent samples (stored <1 year), were observed in the samples B, C and I over storage periods of 65, 11 and 44 years, respectively **(Figure 4A and Table S1).** These results are consistent with the previous observations of a reduction in the complexity of proteomes over long periods of time, leading to a loss/degradation of certain proteins (Thibaut et al, 2010; Parker et al, 2016). By contrast, the samples from individual A did not show significant alterations in the amounts of detected proteins or unique peptides over a period of 11 years. The samples from individual E at both ages provided very low numbers of identified unique peptides (≈1200) and proteins (≈300) compared to the average numbers observed in the other samples (≈3000 and ≈600, respectively) (**Table S1**), an example of a substantial individual effect.

Genetically variant peptide profiles were identified for each individual (A-I) in the longitudinal study with 2 to 6 biological replicates. Overall, 237 different GVPs at 127 loci were identified with 67 ± 18 GVPs per sample (**Table S2)**. A straightforward relationship could not be made between the age of the sample and the number of GVPs observed except for the individuals B, C, and I (**Figure 4B**). The numbers of GVPs decreased 1.48 fold from 57.6 ± 8.5 to 36.6 ± 7 ($p=0.03$) in individual B, 1.5 fold from 63.3 ± 10.5 to 40.3 ± 14 ($p=0.015$) for individual C, and 2.1 fold from 63.6 ± 6 to 33 ± 3 ($p=0.007$), for individual I with storage over periods of 65, 11 and 44 years, respectively. However, the number of GVPs detected was seen to be proportional to the number of identified unique peptides in the samples (R=0.86**, Figure 5A)** as also observed by others (Catlin et al, 2019). GVP detections, when compared with the number of replicates used for each sample, showed that three biological replicates provide enough information to cover 97% of the GVPs, and adding more replicates is hardly more effective (**Figure S2**).

Random match probability

To calculate the random match probability (RMP) at each age, SNP profiles were inferred for each of the samples from their respective GVP profiles. The genotype frequencies from the 1000 Genomes Project for the inferred SNPs were used to calculate the RMPs. The calculation employed the product rule with complete independence between GVPs in different genes and complete dependence with GVPs from the same gene. The calculated random match probabilities ranged from 1 in 73 (for sample E1) to 1 in 185 million (for sample A3). The log of the RMP was found to be proportional to the number of GVPs detected **(Figure 5B)** with rare SNPs considerably increasing the RMPs.

Hierarchical clustering

Proteomic changes observed over 4-7 years were modest. However, more substantial changes over time were observed proteomically in the older samples from 44 and 65 year intervals. This was true for both total numbers of identified proteins (Table S1) and total unique peptide levels (Figure 4A, Table S1). Significant changes were also observed due to batch effects between the second and third cohort of longitudinal samples. A central question of this study was whether these changes also affected the profile of GVP-based inferred SNP genotypes. Therefore, GVP profiles of the individuals at different ages were also compared side by side. Samples from the same individuals were found to carry a large proportion of GVPs common at all ages with some unique GVPs (**Figure S3**). For the GVP profiles generated for individuals A-I, every GVP detection was assigned a value 1 and a non-detection a value 0 to create a binary data file for calculating Euclidean distances and from them to plot an agglomerative hierarchical clustering dendrogram. As seen in **Figure 6,** samples collected at different time points from the same individuals were clustered together, although distances among subjects varied. This includes the samples that had the longest storage periods and greatest level of changes, individuals B and I. It also includes samples from different cohorts of longitudinal samples, individuals A to E and F to I, despite recognizable batch effects (Figure S1). This indicates that the GVP-inferred profiles of SNP alleles were more dependent on individual genotypes than changes occurring as a result of storage with proteome degradation and batch effects.

DISCUSSION

Previous work has shown that inbred mouse strains can be distinguished by their hair protein profiles (Rice et al, 2012). Subsequently, human individuals were also shown to be distinguishable in this way (Laatsch et al, 2014). Studies of monozygotic twins indicate that the basis for such differences is largely genetic (Wu et al, 2017). That the twin profiles were not found to diverge with age would be consistent with a lack of effect of age or changes with age in the same direction within twin pairs. Present results support the latter alternative. Inasmuch as the different hair shaft layers (e.g., cuticle) have different protein profiles from the rest of the shaft (Laatsch et al, 2014), also reported for sheep wool (Koehn et al, 2010), changing proportions of the layers over time as diameters change could result in altered profiles. Hair shaft diameters reportedly change with age, decreasing in the elderly (Robbins et al, 2012; Kim et al, 2013). This finding is consistent with a report that the relative content of mRNAs encoding keratins and keratin associated proteins in hair follicles also changes with age (Giesen et al, 2011). The basis for chronological ageing is multifactorial, but includes accumulation of oxidative damage from ambient oxidants, ultraviolet radiation, copper content (Marsh et al, 2014) and air pollution (De Vecchi et al, 2019).

Present results indicate a lack of consistent population-wide changes, but some changes are evident for individuals. This finding supports possible usefulness of hair shaft protein profiling in distinguishing among individuals over short time periods, but it highlights a dependence on a short interval between sample collections, a clear limitation. Finding a substantially larger difference in subject C after 11 years compared to 5 or 6 years (27 versus 5 or 7) could be rationalized by a drift in profile. Comparing hair samples from individuals collected at greater than 40 year intervals, as for subjects B and I, reveals a large drift. Such changes could result from effects of normal ageing on hair follicle function/gene expression and profile modifications due to exposure to different physicochemical factors during storage. Therefore, proteomic profiling alone would not likely provide sufficient information to distinguish individuals from each other on a large scale. Moreover, batch effects from processing the samples at different times could confound use of a database of proteomic profiles for individual identification.

GVP analysis, on the other hand, was found to be a powerful tool to identify the source of the hair sample in each of the nine subjects studied despite the samples being stored even for periods >40 years. GVP analysis permits calculation of random match probabilities, providing a statistical basis for confidence in the results. The older samples of the individuals B and I, although deficient in proteins and peptides detected, provided GVP profiles with RMPs of 1 in nearly 1000 and 500, respectively. This capability is of particular interest for old and cold cases, where hair is present as evidence and nuclear DNA is not available. The relation between the number of unique peptides, GVPs, and the calculated RMPs testifies to the value of optimizing sample processing procedures and ongoing efforts to maximize their yields in problematic samples (e.g., from individual E).

The observation of lower unique peptide and protein yields with longer storage is consistent with loss of cuticle in older hair samples (Thibaut et al, 2010; Solazzo et al, 2013). This phenomenon could also rationalize the higher proportion in the recent samples of KRTAPs found in the present study. A factor of potential importance is the chemical modification of samples during long term storage. Deamidation, which has been linked with ageing of hairs (Robinson and Robinson, 2004; Adav et al, 2018), was higher in samples stored over a period of at least 10 years (R=0.97) (**Figure S4**). Other common chemical modifications were not consistent in their direction of change. Nevertheless, this observation raises the prospect in general of chemical modifications, some of which could depend on storage conditions. An important area for future investigation is the impact on protein profiles, and especially on GVP yield, of treatments individuals may use to reduce environmental damage, and common chemical treatments that are known to induce considerable damage and to reduce protein yields (Marsh et al, 2015).

Conclusion

The present study highlights that the hair, although very resilient in nature, could undergo developmental and environmental changes over decades, resulting in drift in profile and thus intra-individual variation. Therefore, proteomic profiling alone has limitations for human identification. GVP profiles, in contrast, were seen to be more robust over periods as long as 65 years. The stored hair samples, despite losing a fraction of unique peptides and proteins, were sufficient to provide high RMPs. These findings promise to be highly valuable in resolving routine and even old cases where hair samples are available for investigation.

REFERENCES

1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR (2015) A global reference for human genetic variation. Nature 526:68-74

Adav SS, Subbaiaih RS, Kerk SK, Lee AY, Lai HY, Ng KW, Sze SK, Schmidtchen A (2018) Studies on the proteome of human hair-Identification of histones and deamidated keratins. Scientific Reports 8(1):1599

Borja T, Karim N, Goecker Z, Salemi M, Phinney BS, Naeem M, Rice RH, Parker GJ (2019) Proteomic genotyping of fingermark donors with genetically variant peptides. Foren Sci Int: Genet 42:21-30

Catlin LA, Chou RM, Goecker ZC, Mullins LA, Silva DS, Spurbeck RR, Parker GJ, Bartling CM (2019) Demonstration of a mitochondrial DNA-compatible workflow for genetically variant peptide identification from human hair samples. Foren Sci Int: Genet 43:102148

Chu F, Mason KE, Anex DS, Jones AD, Hart BR (2019) Hair proteome variation at different body locations on genetically variant peptide detection for protein-based human identification. Scientific Reports 9(1):7641

De Vecchi R, da Silveira Carvalho Ripper J, Roy D, Breton L, Alexandre Germano Marciano AG, de Souza PMB, de Paula Corrêa M (2019) Using wearable devices for assessing the impacts of hair exposome in Brazil Scientific Reports 9(1):13357

Dias MF (2015) Hair cosmetics: an overview. Int J Trichology 7:2-15

Goecker ZC, Wills BM, Salemi SR, Phinney BS, Rice RH, Walsh S, Parker GJ (2019) Biogeographic classification of European and African hair using genetically variant peptides. 30th Annual International Symposium on Human Identification Poster #64

Gonzalez AC, Schweizer M, Jagdmann S, Bernreuther C, Reinheckel T, Saftig P, Damme M (2018) Unconventional trafficking of mammalian phospholipase D3 to lysosomes. Cell Reports 22:1040-1053

Kim SN, Lee SY, Choi MH, Joo KM, Kim SH, Koh JS, Park WS (2013) Characteristic features of ageing in Korean women's hair and scalp. Br J Dermatol 168:1215-1223

Kobayashi T, Imokawa G, Bennett DC, Hearing VJ (1998) Tyrosinase stabilization by Tyrp1 (the brown locus protein). J Biol Chem 273:31801-31805

Koehn H, Clerens S, Deb-Choudhury S, Morton J, Dyer JM, Plowman JE (2010) The proteome of the wool cuticle. J Proteome Res 9:2920-2928

Laatsch CN, Durbin-Johnson BP, Rocke DM, Mukwana S, Newland AB, Flagler MJ, Davis MG, Eigenheer RA, Phinney BS, Rice RH (2014) Human hair shaft proteomic profiling: individual differences, site specificity and cuticle analysis. PeerJ 2:e506

Marsh JM, Iveson R, Flagler MJ, Davis MG, Newland AB, Greis KD, Sun Y, Chaudhary T, Aistrup ER (2014) Role of copper in Photochemical damage to hair. Int J Cosmetic Sci 36:32-38

Marsh JM, Davis MG, Flagler MJ, Sun Y, Chaudhary T, M Mamak M, McComb DW, Williams REA, Greis KD, Rubio L, Coderch L (2015) Advanced hair damage model from ultra-violet radiation in the presence of copper Int J Cosmetic Sci 37:532-541

Milan J, Wu P-W, Salemi M, Durbin-Johnson B, Rocke DM, Phinney BS, Rice RH, Parker GJ (2019) Comparison of protein expression levels and proteomically-inferred genotypes using human hair from different body sites. Foren Sci Int: Genet 41:19-23

Moll R, Divo M, Langbein L (2008) The human keratins: biology and pathology. Histochem Cell Biol 129:705-733

Parker G, Goecker Z, Franklin R, Durbin-Johnson B, Milan J, Karim N, De Leon C, Matzoll A, Borja T, Rice B (2019) Proteomic genotyping: using mass specrometry to infer SNP genotypes in a forensic context. For Sci Intl: Genet Suppl Ser 7:664-666

Parker GJ, Leppert T, Anex DS, Hilmer JK, Matsunami N, Baird L, Stevens J, Parsawar K, Durbin-Johnson BP, Rocke DM, Nelson C, Fairbanks DJ, Wilson AS, Rice RH, Woodward SR, Bothner B, Hart H, Leppert M (2016) Demonstration of protein-based human identification using the hair shaft proteome. PLoS One 11(9):e0160653

Perez-Riverol Y, Csordas A, Bai J, Bernal-Llinares M, Hewapathirana S, Kundu D, Inuganti A, Griss J, Mayer G, Eisenacher M, Pérez E, Uszkoreit J, Pfeuffer J, Sachsenberg T, Yilmaz S, Tiwary S, Cox J, Audain E, Walzer M, Jarnuczak AF, Ternent T, Brazma A, Vizcaíno JA (2019)

The PRIDE database and related tools and resources in 2019: improving support for quantification data. Nucl Acids Res 47(D1):D442-D450

Rice RH, Bradshaw KM, Durbin-Johnson BP, Rocke DM, Eigenheer RA, Phinney BS, Sundberg JP (2012) Differentiating inbred mouse strains from each other and those with single gene mutations using hair proteomics. PLoS One 7:e51956

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucl Acids Res 43(7):e47

Robbins C, Mirmirani P, Messenger AG, Birch MP, Youngquist RS, Tamura M, Filloon T, Luo F, Dawson TLJ (2012) What women want - quantifying the perception of hair amount: an analysis of hair diameter and density changes with age in caucasian women. Br J Dermatol 167:324-332

Robinson NE, Robinson AB (2004) Amide molecular clocks in drosophila proteins: potential regulators of aging and other processes. Mech Ageing Dev 125:259-267

Rong X, Wang B, Dunham MM, Hedde PN, Wong JS, Gratton E, Young SG, Ford DA, Tontonoz P (2015) Lpcat3-dependent production of arachidonoyl phospholipids is a key determinant of triglyceride secretion. Elife 4:e06557

Solazzo C, Dyer JM, Clerens S, Plowman J, Peacock EE, Collins MJ (2013) Proteomic evaluation of the biodegradation of wool fabrics in experimental burials. Int Biodeterior Biodegrad 80:48-59

Thibaut S, De Becker E, Bernard BA, Huart M, Fiat F, Baghdadli N, Luengo GS, Leroy F, Angevin P, Kermoal AM, Muller S (2010) Chronological ageing of human hair keratin fibres. Int J Cosmetic Sci 32:422-434

Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A, Olsson I, Edlund K, Lundberg E, Navani S, Al-Khalili Szigyarto C, Odeberg J, Djureinovic D, Takanen JO, Hober S, Alm T, Edqvist P-H, Berling H, Tegel H, Mulder J, Rockberg J, Nilsson P, Schwenk JM, Hamsten M, von Feilitzen K, Forsberg M, Persson L, Johansson F, Zwahlen M, von Heijne G, Nielsen J, Pontén F (2015) Tissue-based map of the human proteome. Science 347:394 (1260419)

Wu P-W, Mason KE, Durbin-Johnson BP, Salemi M, Phinney BS, Rocke DM, Parker GJ, Rice RH (2017) Proteomic analysis of hair shafts from monozygotic twins: Expression profiles and genetically variant peptides. Proteomics 17:13-14, 1600462

Zhang J, Xin L, Shan B, Chen W, Xie M, Yuen D, Zhang W, Zhang Z, Lajoie GA, Ma B (2012) PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. Mol Cell Proteomics 11(4):M111.010587

Zhang P, Liu W, Zhu C, Yuan X, Li D, Gu W, Ma H, Xie X, Gao T (2012) Silencing of GPNMB by siRNA inhibits the formation of melanosomes in melanocytes in a MITF-independent fashion. PLoS One 7(8):e42955

Table 1

Pairwise comparisons of differentially expressed proteins by age and ethnic origin.*

A	A20's	A40's
A40's	0	
A60's	0	0

B	E20's	E40's
E40's	0	
E60's	0	0

C	20's	40's
40's	0	
60's	0	0

D	A20's	A40's	A60's
E20's	8		
E40's		6	
E60's			2

E	All A
All E	19

*Ethnic groups are indicated by African (A) and European (E) and age groups by 20's, 40's and 60's. The numbers in table indicate the number of proteins with significant differences in expression level.

Table 2

Pairwise comparison of proteins significantly different in expression level (weighted spectral counts) in two-way comparisons.*

	A6	**A11**	**B0**	**B65**	**C0**	**C6**	**C11**	**D0**	**D5**	**E0**	**E4**
A0	***2***	***0***	34	4	64	7	7	23	22	206	132
A6		***6***	13	17	30	2	6	7	11	227	131
A11			30	15	56	6	11	26	23	168	103
B0				***32***	26	17	35	14	16	147	120
B65					88	23	9	24	26	196	132
C0						***5***	***27***	54	42	99	105
C6							***7***	10	9	35	28
C11								38	28	168	93
D0									***1***	135	118
D5										204	127
E0											***3***

*Subjects are identified by letter and years since the first collection (0). Comparisons within the same individual from different years are in bold italic. The numbers in the table indicate the number of differentially expressed proteins.

Figure Legends

Figure 1. Proteins differing in hair samples from African and European subjects. Shown are the ratios of relative amounts of proteins that differed significantly, judging by weighted spectral counts, between the samples collected from African and European subjects.

Figure 2. Distances in protein expression levels between samples from single individuals and between subjects. Box plots of Euclidean distances between samples, based on weighted spectral counts. The solid line on each box indicates the median, the lower and upper box edges indicate the 25th and 75th percentiles, respectively, and the lower and upper whiskers indicate the smallest and largest observations lying within 1.5 interquartile ranges of the box edges, respectively.

Figure 3. Heatmap showing differences in the proteomic composition of the newly and previously collected samples of (A) cohort 3 (individuals F-I), and (B) individual I at two times points with a difference of 44 years. The numbers after the hyphens in the sample names represent the storage time of the samples.

Figure 4: Unique peptides (A) and GVPs (B) in samples from individuals at different ages. The lines of different color show values (averages and standard deviations) for individuals at the

ages indicated. Significantly lower values in the unique peptides were observed in the stored samples of individuals B, C and I marked by asterisks. Periods of storage are indicated by the time span between points for given subjects.

Figure 5. The number of GVPs vs (A) the unique peptides identified in each sample and (B) calculated random match probabilities. The graph shows that the higher the number of unique peptides identified in a sample, the higher will be the number of GVPs observed (p value = 0.0001) and the higher the random match probabilities calculated (p value = 0.003).

Figure 6. Hierarchical clustering dendrogram of all the samples from individual subjects. Based on the Euclidean distances among the samples, the clustering shows that GVP profiles can distinguish individuals despite differences in hair collection and storage times.

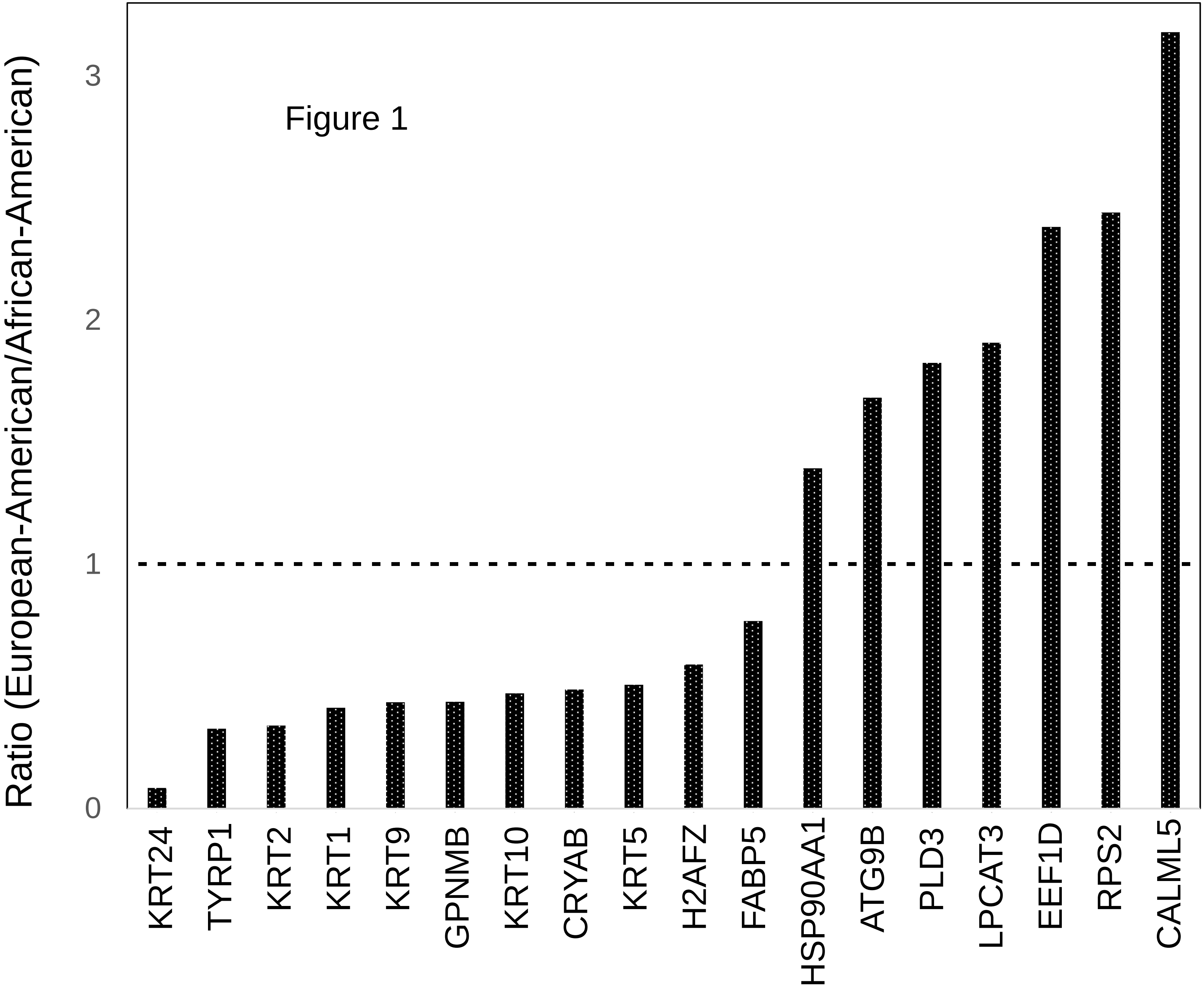
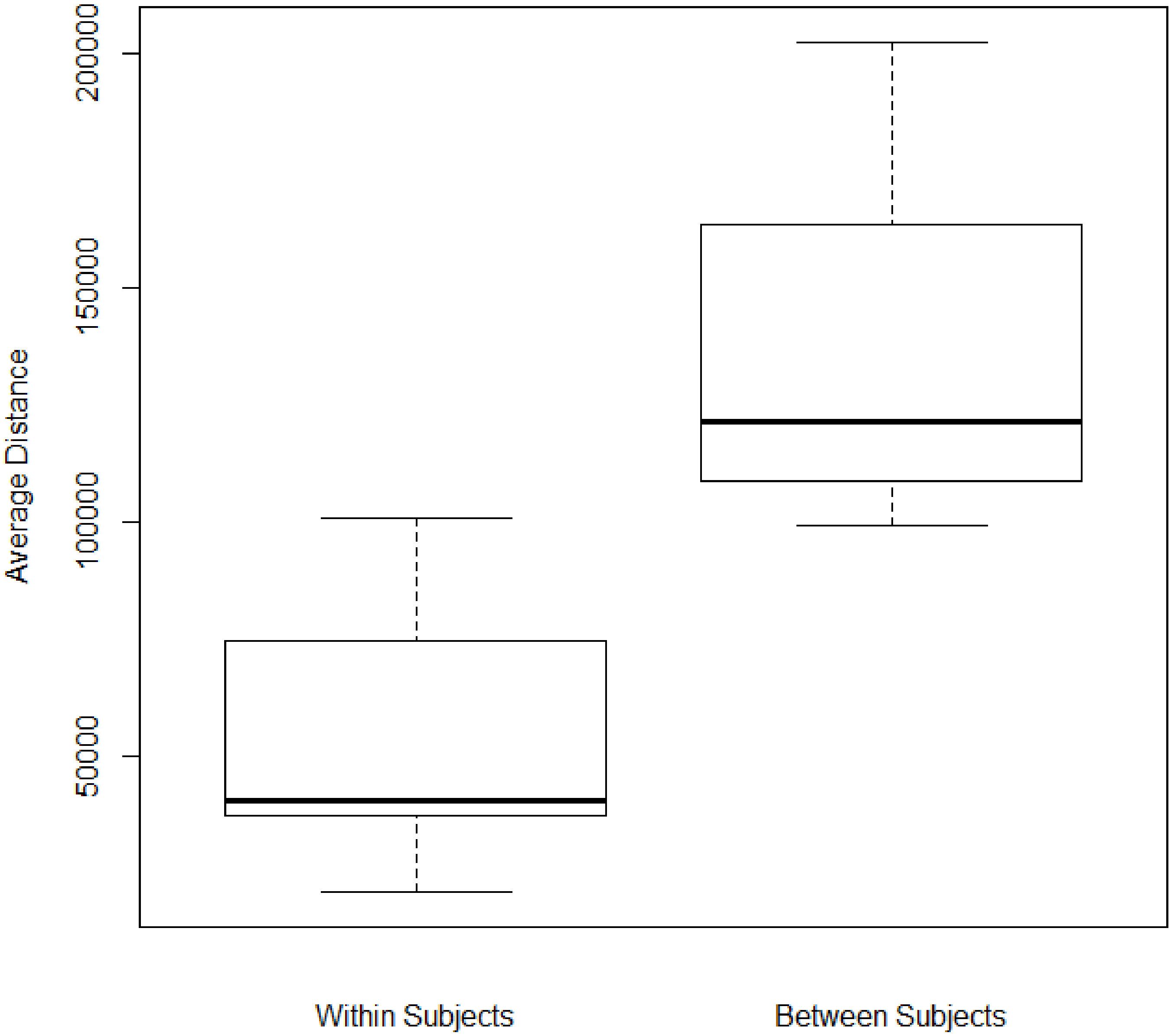


Figure 1
Ratio (European-American/African-American)
3
2
1
0
KRT24
TYRP1
KRT2
KRT1
KRT9
GPNMB
KRT10
CRYAB
KRT5
H2AFZ
FABP5
HSP90AA1
ATG9B
PLD3
LPCAT3
EEF1D
RPS2
CALML5

Figure 2



Average Distance
200000
150000
100000
50000
Within Subjects
Between Subjects

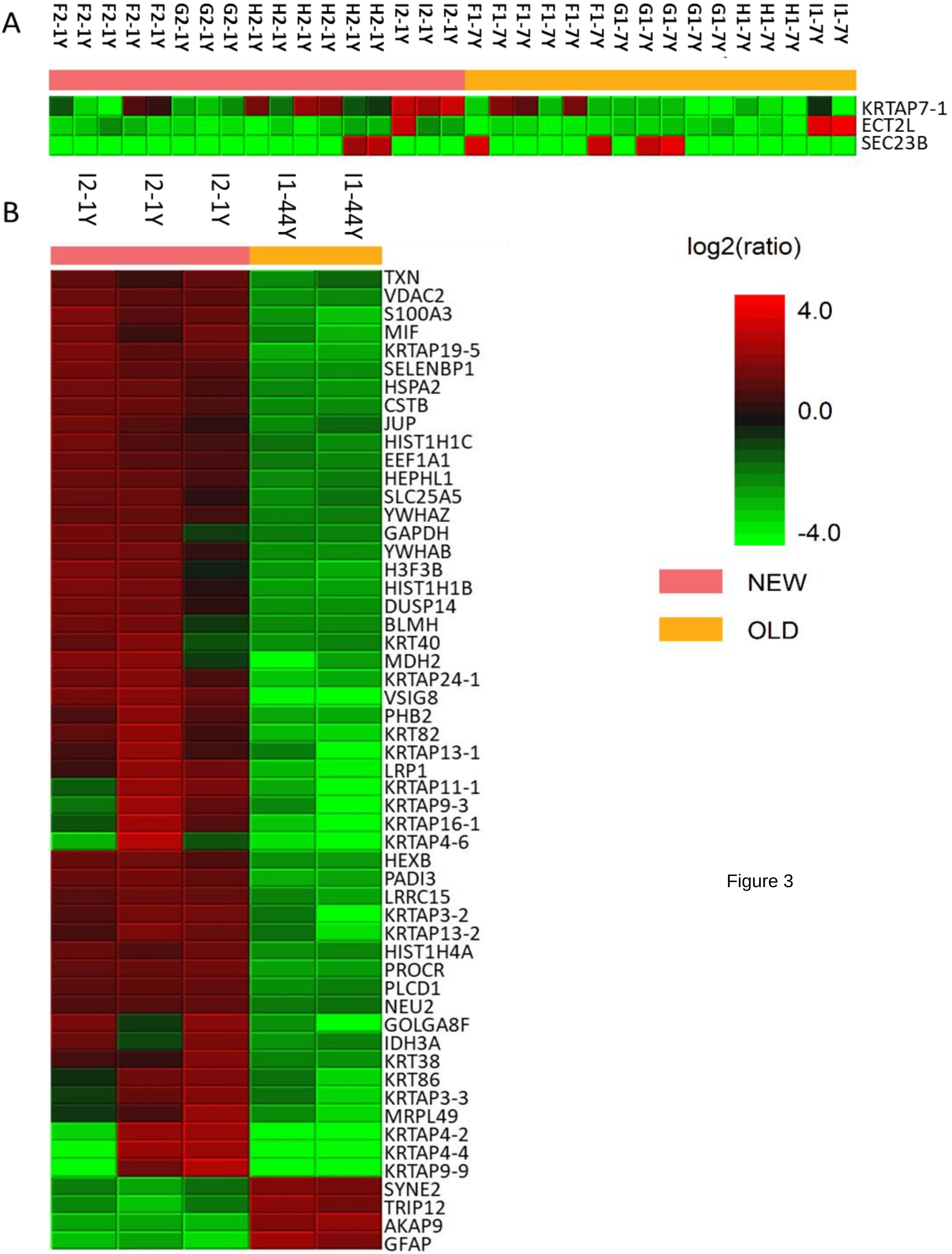

A
F2-1Y
F2-1Y
F2-1Y
F2-1Y
F2-1Y
G2-1Y
G2-1Y
G2-1Y
H2-1Y
H2-1Y
H2-1Y
H2-1Y
H2-1Y
H2-1Y
I2-1Y
I2-1Y
I2-1Y
F1-7Y
F1-7Y
F1-7Y
F1-7Y
F1-7Y
F1-7Y
G1-7Y
G1-7Y
G1-7Y
G1-7Y
G1-7Y
H1-7Y
H1-7Y
H1-7Y
I1-7Y
I1-7Y
KRTAP7-1
ECT2L
SEC23B
B
I2-1Y
I2-1Y
I2-1Y
I1-44Y
I1-44Y
TXN
VDAC2
S100A3
MIF
KRTAP19-5
SELENBP1
HSPA2
CSTB
JUP
HIST1H1C
EEF1A1
HEPHL1
SLC25A5
YWHAZ
GAPDH
YWHAB
H3F3B
HIST1H1B
DUSP14
BLMH
KRT40
MDH2
KRTAP24-1
VSIG8
PHB2
KRT82
KRTAP13-1
LRP1
KRTAP11-1
KRTAP9-3
KRTAP16-1
KRTAP4-6
HEXB
PADI3
LRRC15
KRTAP3-2
KRTAP13-2
HIST1H4A
PROCR
PLCD1
NEU2
GOLGA8F
IDH3A
KRT38
KRT86
KRTAP3-3
MRPL49
KRTAP4-2
KRTAP4-4
KRTAP9-9
SYNE2
TRIP12
AKAP9
GFAP
log2(ratio)
4.0
0.0
-4.0
NEW
OLD

Figure 3

Figure 4

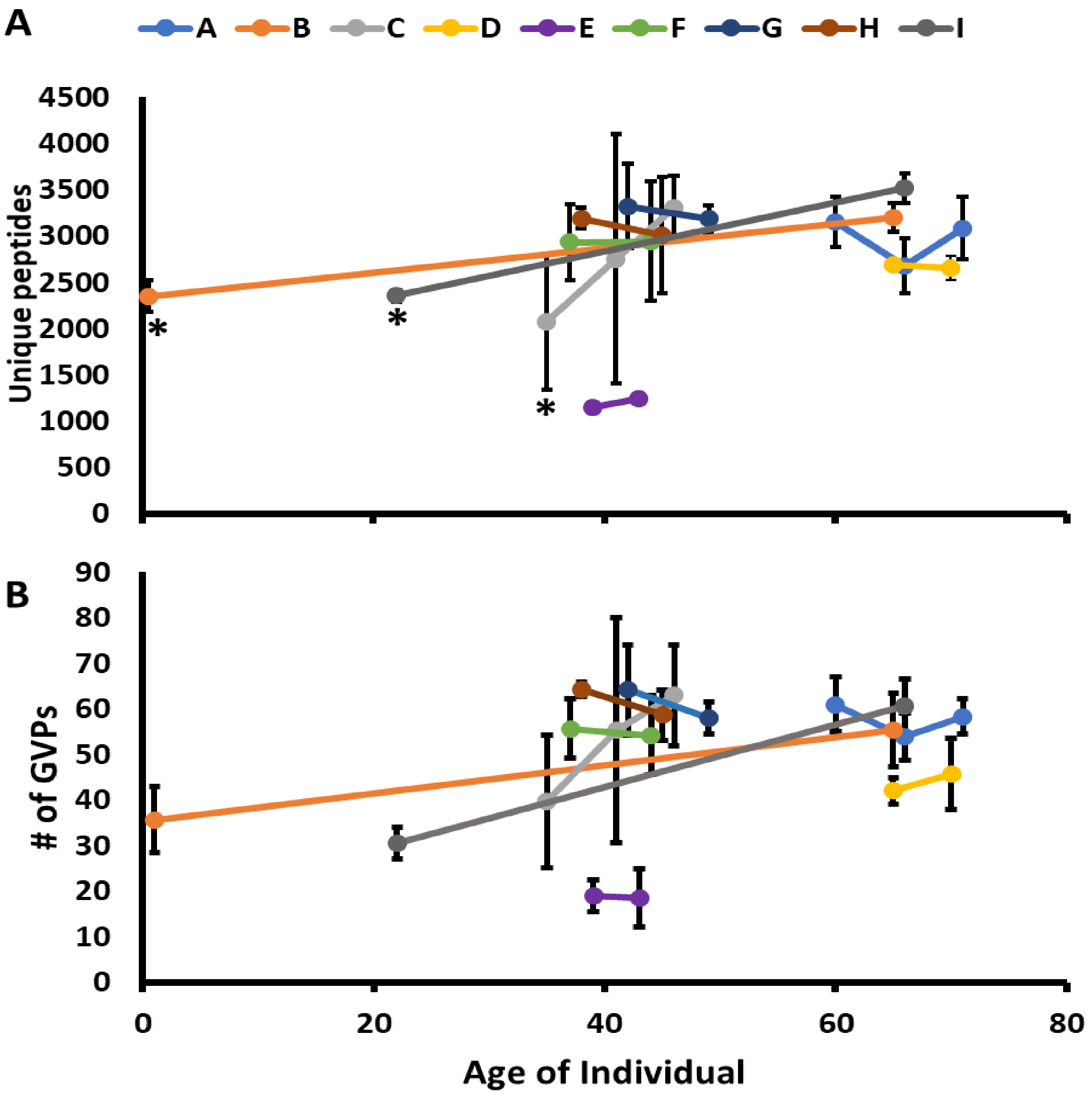

A
A
B
C
D
E
F
G
H
I
Unique peptides
4500
4000
3500
3000
2500
2000
1500
1000
500
0
*
*
*
B
# of GVPs
90
80
70
60
50
40
30
20
10
0
0
20
40
60
80
Age of Individual

Figure 5

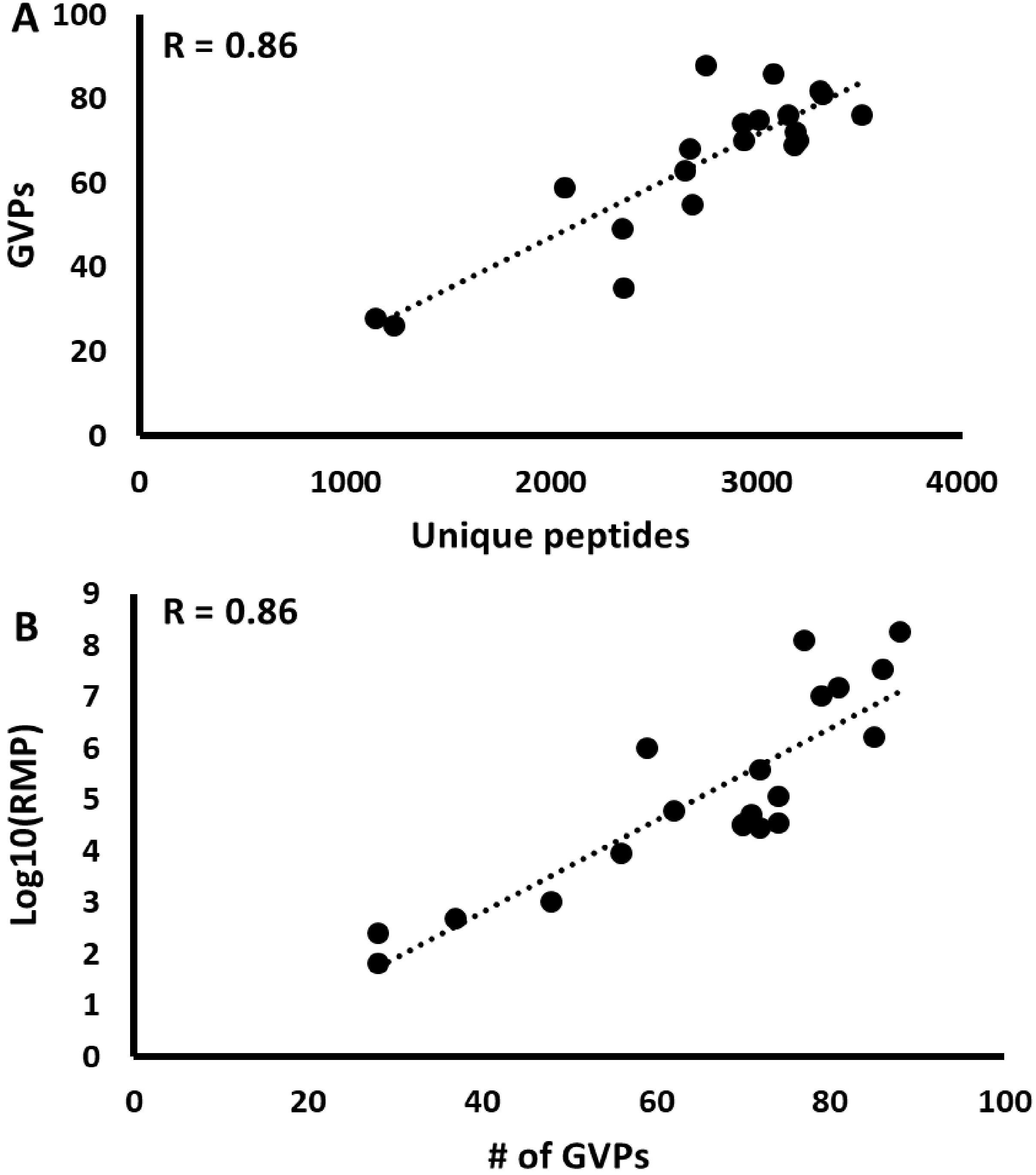

A
R = 0.86
GVPs
100
80
60
40
20
0
0
1000
2000
3000
4000
Unique peptides
B
R = 0.86
Log10(RMP)
9
8
7
6
5
4
3
2
1
0
0
20
40
60
80
100
# of GVPs

Figure 6

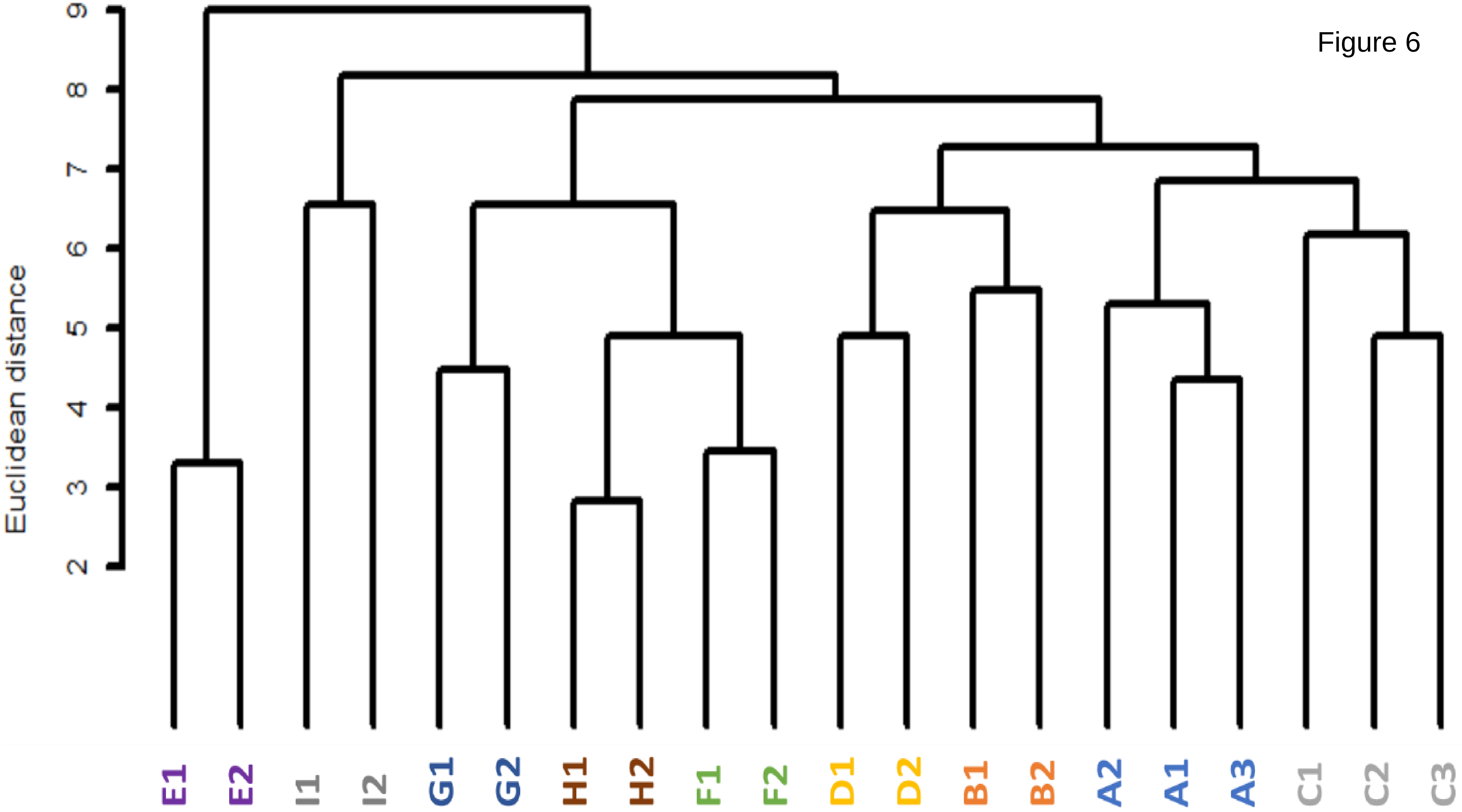

Euclidean distance
9
8
7
6
5
4
3
2
E1
E2
I1
I2
G1
G2
H1
H2
F1
F2
D1
D2
B1
B2
A2
A1
A3
C1
C2
C3

Table S1. Details of samples used in the study.

Sample Name	Collection Year	Age of Individual	Age of Sample	Sex (M/F)	Cohort	Identified Proteins	Unique Identified Peptides	Average Identified Protein ± SD	Average Unique Peptides ± SD
A1	2005	60	11	M	2	671	3337	652 ± 18	3156 ± 273
A1						651	3288		
A1						634	2842		
A2	2011	66	5	M	2	591	2757	571 ± 22	2677 ± 291
A2						573	2919		
A2						548	2354		
A3	2016	71	1	M	2	674	3241	672 ± 69	3082 ± 335
A3						603	2697		
A3						740	3309		
C1	2005	35	11	M	2	567	2547	471 ± 130	2069 ± 736
C1						324	1221		
C1						523	2439		
C2	2011	41	5	M	2	296	1223	565 ± 240	2752 ± 1346
C2						643	3273		
C2						756	3760		
C3	2016	46	1	M	2	672	3553	678 ± 29	3305 ± 345
C3						644	2911		
C3						702	3453		
B1	1951	< 1	65	M	2	575	2545	571 ± 7	2347 ± 174
B1						575	2216		
B1						563	2281		
B2	2016	65	1	M	2	687	3038	673 ± 30	3200 ± 157
B2						639	3212		
B2						693	3352		
D1	2011	65	5	F	2	603	2662	622 ± 27	2688 ± 36
D1						641	2714		
D2	2016	70	1	F	2	623	2585	609 ± 42	2651 ± 128

D2						642	2799		
D2						562	2570		
E1	2012	39	4	M	2	274	1157	282 ± 18	1148 ± 15
E1						269	1130		
E1						302	1156		
E2	2016	43	1	M	2	275	1226	304 ± 41	1236 ± 14
E2						333	1246		
G2	2017	49	1	M	3	633	3073	672 ± 42	3184 ± 147
G2						716	3350		
G2						667	3128		
G1	2010	42	7	M	3	683	3475	657 ± 107	3320 ± 456
G1						716	3589		
G1						467	2526		
G1						722	3639		
G1						695	3369		
F2	2017	44	1	F	3	723	3446	639 ± 107	2939 ± 646
F2						779	3805		
F2						568	2484		
F2						527	2349		
F2						598	2609		
F1	2010	37	7	F	3	577	2685	643 ± 70	2932 ± 408
F1						567	2457		
F1						597	2579		
F1						719	3442		
F1						688	3244		
F1						711	3187		
H1	2010	38	7	F	3	622	3064	664 ± 39	3190 ± 115
H1						698	3290		
H1						671	3217		
H2	2017	45	1	F	3	522	2452	622 ± 105	3011 ± 628
H2						544	2479		

H2						527	2442		
H2						661	3267		
H2						748	3681		
H2						732	3745		
I2	2017	66	1	M	3	801	3504	773 ± 39	3513 ± 161
I2						730	3357		
I2						788	3679		
I1	1973	22	44	M	3	681	2398	700 ± 28	2353 ± 64
I1						720	2308		

Table S2. GVPs identified at 127 loci.

Gene Nar	rs#_nuc	SAP	peptide sequence	A1	A2	A3	B1	B2	C1	C2	C3	D1	D2	E1	E2	F1	F2	G1	G2	H1	H2	I1	I2
ACAA1	rs222952	V294A	QVITLLNELK	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	1
ACAA1	rs222952	V294A	QaITLLNELK	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ALDH2	rs671_G	E504K/E45	ELGEYGLQAYTEVK	1	1	1	1	1	0	1	1	0	0	0	0	1	0	0	0	0	0	0	1
ALDH2	rs671_A	E504K	ELGEYGLQAYTk	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ANAX2	rs178452	V98L	*SALSGHLETlILGLLK*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
ATP5A1	rs790112	A32S	VLSIGDGIAR	1	1	0	1	0	0	1	1	1	1	0	0	1	1	1	1	1	1	1	1
ATP5A1	rs790112	A32S	*VLSIGDGIsR*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CSRP1	rs373828	K108I	HEEAPGHRPTTNPNA	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	0	1
CSRP1	rs373828	K108I	HEEAPGHRPTTNPNA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DSC3	rs352969	K180Q	GVDKEPLNLFYIER	1	1	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
DSC3	rs352969	K180Q	GVDqEPLNLFYIER	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DSP	rs287639	N1526K	ANSSATETINK	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
DSP	rs287639	N1526K	ANSSATETIk	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DSP	rs287639	R1537C	VQEQELTR	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1
DSP	rs287639	R1537C	VQEQELTcLR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DSP	rs692906	R1738Q	GqSEADSDKNATILE	1	0	1	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
DSP	rs692906	R1738Q	GRSEADSDKNATILE	1	0	1	1	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0
EFHD1	rs115506	K90R	LSEIDVALEGVK	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
EFHD1	rs115506	K90R / K1	LSEIDVALEGVr	1	0	1	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
EIF2S2	rs178560	E177D	DYTYEELLNR	1	0	0	0	0	0	1	0	0	0	0	0	1	1	1	1	1	1	0	1
EIF2S2	rs178560	E177D	DYTYdELLNR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GPNMB	rs353632	P324L	AAAPGPCPPPPPPPR	0	0	0	0	0	1	1	0	0	1	0	0	0	0	1	1	0	0	0	0
GPNMB	rs353632	P324L	AAAPGPClPPPPPPR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GSDMA	rs389419	R18Q	QLNPqGDLTPLDSLI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GSDMA	rs389419	R18Q	QLNPR/GDLTPLDSL	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1
GSDMA	rs721293	V128L	ALETVQER	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GSDMA	rs721293	V128L	ALETlQER	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GSTP1	rs113827	A114V	YISLIYTNYEAGKDDYV	1	1	1	0	1	0	1	1	1	1	0	0	0	1	1	0	1	1	0	1
GSTP1	rs113827	A114V	YISLIYTNYEvGKDDYV	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GSTP1	rs1695_A	I105V	YISLIYTNYEAGKDD	0	0	1	0	1	0	1	1	0	0	0	0	0	1	0	0	0	1	0	0
GSTP1	rs1695_G	I105V	YvSLIYTNYEAGKDD	1	1	1	0	0	0	1	1	1	1	0	0	0	0	1	0	1	1	0	1
HEXB	rs108058	I207V	GILIDTSR	0	0	0	1	1	1	1	1	0	1	0	0	1	1	1	1	1	1	1	1
HEXB	rs108058	I207V	GILvDTSR	1	1	1	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	1	1
HEXB	rs774999	I420V	K.LAPGTIVEVWKD	1	1	1	1	1	0	1	1	1	0	0	0	1	1	0	0	1	1	1	1
HEXB	rs774999	I420V	*K.LAPGTvVEVWKD*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
IL1F10	rs676127	T44I	ICTLPNR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
IL1F10	rs676127	T44I	ICiLPNR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
JUP	rs412834	R142H	SAIVHLINYQDDAEL	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	0	1
JUP	rs412834	R142H	SAIVHLINYQDDAEL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
JUP	rs143043	V648I	NEGTATYAAAVLFR	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1
JUP	rs143043	V648I	NEGTATYAAAiLFR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRT32	rs207156	S222Y	ADLEAQVEyLK	1	1	1	1	1	0	0	0	1	1	0	0	1	1	0	0	1	1	0	1
KRT32	rs728300	R280H	CQYEAMVEANRR	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1
KRT32	rs728300	R280H	CQYEAMVEANhR	1	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	0	1
KRT32	rs260495	P427T	SLLENEDCKLPCNPC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRT32	rs260495	P427T	SLLENEDCKLPCNPC	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1
KRT32	rs374478	Q72R	TYLSSSCQAASGISGS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
KRT32	rs374478	Q72R	TYLSSSCr	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRT32	rs207156	I171T	MVVNIDNAK	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1
KRT32	rs207156	I171T	MVVNtDNAK	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRT32	rs2604955	N402S	*LEGEINTYRSLLEsED*	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
KRT32	rs146792	A255T	LNIEVDAAPPVDLTR	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
KRT32	rs146792	A255T	LNIEVDtAPPVDLTR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRT34	rs223971	I238T/ I28	SQYEALVEINR / SQ	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	1	0	1	1
KRT34	rs207159	H348R	DSLENTLTESEAHYS	1	1	1	1	1	0	1	0	1	1	0	0	1	1	1	1	1	1	1	1
KRT35	rs743686	S36P	VSAMYSSSSCKLPSL	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	1	0	0	1	1
KRT35	rs743686	S36P	VSAMYSSSpCKLPSL	1	1	1	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
KRT35	rs138303	R163W	YETEVSLwQLVESDI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRT35	rs138303	R163W	YETEVSLRQLVESDI	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1
KRT35	rs124516	C441Y	TNCSPRPICVPCPGG	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
KRT35	rs124516	C441Y	TNySPRPICVPCPGG	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRT35	rs207160	P413A	TNCSaRPICVPCPGG	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRT35	rs207160	P413A	TNCSPRPICVPCPGG	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0

KRT36	rs757906	A202G / A	CQLGDRLNVEVDAA	1	1	1	1	1	0	1	1	1	1	0	0	1	1	1	1	1	1	1	1
KRT36	rs757906	A202G	CQLGDRLNVEVDgA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRT36	rs116573	N357T	YSSQLAQMQCLISN	1	0	1	0	0	0	0	0	1	0	0	1	1	1	0	1	1	1	0	1
KRT36	rs116573	N357T	YSSQLAQMQCLIStV	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRT36	rs990410	R277C	CQYEALVENNR	1	1	1	0	1	1	1	1	0	1	0	0	1	1	1	1	1	1	0	0
KRT36	rs990410	R277C	CQYEALVENNcR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRT37	rs991672	N39S	NVFVSPIDVGCQPV	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
KRT37	rs991672	N39S	NVFVSPIDVGCQPV	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
KRT37	rs991648	T72A	PSLCLPPTSHTACPLPG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
KRT37	rs991648	T72A	PSLCLPPaSHTACPLPG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRT37	rs991647	S73C	PSLCLPPTSHTACPLPG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
KRT37	rs991647	S73C	PSLCLPPTcHTACPLPG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRT37	rs169668	A217V	LLDDvTLAK	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
KRT38	rs897416	S423P	LPCNPCSTSPSCVTA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRT38	rs897416	S423P	LPCNPCSTpPSCVTA	1	1	1	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0	1
KRT39	rs178430	T341M	DSQECILTETEAR	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1
KRT39	rs178430	T341M	DSQECILmETEAR	1	0	1	0	0	1	1	1	0	0	0	0	0	0	0	0	1	1	0	0
KRT39	rs142154	S86N	FSLDDCSWYGEGIN	0	1	0	0	0	0	1	0	1	0	0	0	1	1	1	1	1	1	1	1
KRT39	rs142154	S86N	FSLDDCnWYGEGIN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRT39	rs721325	R456Q	SGAIESTAPACTSSS	1	1	0	0	1	1	1	1	1	1	0	0	1	1	1	1	1	1	0	1
KRT39	rs721325	R456Q	SGAIESTAPACTSSS	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0
KRT39	rs178430	L383M	QNQEYEILLDVK	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
KRT39	rs178430	L383M	QNQEYEILmDVK	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
KRT40	rs150812	C349R	TASALEIELQAQQSL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRT40	rs150812	C349R	TASALEIELQAQQSL	0	1	0	0	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1
KRT40	rs201002	R235H	NHEEEVNLLREQLG	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	0	1
KRT40	rs201002	R235H	NHEEEVNLLhEQLG	1	0	1	0	1	0	1	1	0	0	0	0	1	0	1	1	1	1	0	0
KRT40	rs140634	R108H	R.SLEETNAELESR	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1
KRT40	s1406344	R108H	*VhSLEETNAELESR*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRT40	rs721957	C265Y	CQCETVLANN RR	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
KRT40	rs721957	C265Y	CQyETVLANN RR	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	1
KRT75	rs223239	E242G	YEDEINKRTAAENEFV	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
KRT75	rs223239	E242G	YEDgINK	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRT81	rs658087	L248R	LYEEEILILQSHISDTS	1	0	1	1	1	1	0	1	1	0	0	1	0	0	0	0	0	0	1	1
KRT82	rs265865	T458M	GAFLYEPCGVSmPV	1	1	1	0	1	0	0	0	0	1	0	0	1	1	1	1	1	1	0	0
KRT82	rs265865	T458M	GAFLYEPCGVSTPVL	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	0	1	1
KRT82	rs1732263	E452D	GAFLYEPCGVSTPVL	1	0	1	0	1	0	1	1	0	0	1	1	1	1	1	0	1	1	0	1
KRT82	rs1732263	E452D	GAFLYdPCGVSTPVL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRT82	rs179163	E219Q	KYEEELSLRPCVENEFV	1	1	1	1	1	0	1	1	1	1	0	0	1	1	1	1	1	1	0	1
KRT82	rs179163	E219Q	KYEEELSLRPCVqNEFV	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRT83	rs285246	I279M	DLNMDCmVAEIK	0	0	0	0	0	1	1	1	1	1	0	0	1	1	1	1	1	1	0	0
KRT83	rs285246	I279M	DLNMDCIVAEIK	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	0	0	1	1
KRT83	rs285767	H493Y	GGVVCGDLCVSGSR	1	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
KRT83	rs285767	H493Y	GGVVCGDLCVSGSR	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
KRT83	rs285766	R149C	LQFYQNR.ECCQSN	1	1	1	1	1	0	0	1	1	1	0	0	1	1	1	1	1	1	1	1
KRT83	rs285766	R149C	LQFYQNCECCQSNL	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
KRT84	RS951773	C446R	CEYQELMNAKLGLD	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
KRT84	RS951773	C446R	QLrEYQELMNAKLG	0	0	1	0	0	0	0	0	1	0	0	0	1	1	1	0	1	1	1	1
KRT85	rs616300	R78H	IAVGGFRAGSCGR /	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRT85	rs616300	R78H	IAVGGFRAGSCGhS	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRT86	rs587172	Q139P	LpFYQNR	0	0	0	1	1	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0
KRTAP1-1	rs150218	P12R	ACCQTSFCGFPSCSTS	1	1	1	0	1	1	1	1	0	0	0	0	1	1	1	1	1	1	0	0
KRTAP1-1	rs150218	P12R	ACCQTSFCGFr	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRTAP1-1	rs138200	C14F	ACCQTSFCGFPSCST	1	1	1	0	1	1	1	1	0	0	0	0	1	1	1	1	1	1	0	0
KRTAP1-5	rs148449	T32S	TCCQTSFCGYPSFSI	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0
KRTAP1-5	rs148449	T32S	*TCCQTSFCGYPSFSI*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRTAP1-5	rs626233	C35Y	MTCCQTSFCG YPS	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0
KRTAP1-5	rs626233	C35Y	*MTCCQTSFCG YPS*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRTAP1-5	rs138758	T52A	SCQTSFCGFPSFSTS	1	1	1	0	1	1	1	1	0	0	1	1	1	1	1	1	1	1	0	0
KRTAP1-5	rs138758	T52A	*SCQaSFCGFPSFSTS*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRTAP3-2	rs989704	S8G	MDCCASRSCSVPTG	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
KRTAP3-2	rs381305	I46T	CGVCLPSTCPHTVWL	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
KRTAP3-2	rs382959	R27C	SCSVPTGPATTICSS	0	1	1	1	1	0	1	0	0	0	0	0	1	1	1	1	1	1	0	0
KRTAP3-2	rs382959	R27C	K.SCCCGVCLPSTCP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

KRTAP4-2	rs620672	T59S	TTCCRPSCCVSSCCR	1	1	1	0	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0
KRTAP4-2	rs620672	T59S	*TTCCRPSCCVSSCCR*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRTAP4-2	rs389784	Y95C	*TTCCRPSCCVSSCFR*	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRTAP4-2	rs389784	Y95C	TTCCRPSCCVSSCFR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRTAP4-3	rs428371	P152S	PACCISSCCHPSCCVSS	1	1	1	1	1	1	1	1	0	1	0	0	1	1	1	1	1	1	0	1
KRTAP4-3	rs428371	P152S	sACCISSCCHPSCCVSS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRTAP4-4	rs366700	R154S	TTCCRPSCCVSRCYR	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1
KRTAP4-4	rs366700	R154S	*TTCCRPSCCVSsCYR*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRTAP4-4	rs385055	Y25C	VNSCCGSVCSDQGC	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	0	1
KRTAP4-4	rs444509	C35S	R.TTCCRPSCCVSSC	0	0	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
KRTAP4-4	rs444509	C35S	*R.TTsCRPSCCVSSCC*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRTAP4-4	rs750304	Q109R	TTCCRPSCCRPQCC	1	1	1	1	0	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0
KRTAP4-4	rs750304	Q109R	TTCCRPSCCRPr	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRTAP4-5	rs149738	R22C	*VSSCCGSVSSEQSCG*	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
KRTAP4-6	rs739831	P63S	*R.TTCCRPSCCVSSC*	1	1	1	0	1	1	1	1	0	1	0	0	0	0	1	0	0	0	0	0
KRTAP4-8	rs201814	T183S	VSCHTTCYRPACVIST	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
KRTAP4-8	rs201814	T183S	PACVISsCPR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRTAP4-8	rs138296	G7S	VNSCCGSVCSDQGC	0	0	0	0	0	0	1	0	0	0	1	0	0	1	1	0	1	1	0	0
KRTAP4-8	rs138296	G7S	VNSCCsSVCSDQGC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRTAP4-9	rs113059	D18V	VSSCCGSVCSDQGC	1	1	1	0	0	1	1	1	0	0	0	0	1	0	1	0	0	0	0	0
KRTAP4-9	rs113059	D18V	VSSCCGSVCSDQGC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRTAP4-1	rs989703	R17Q	VNSCCGSVCSHQGC	0	0	0	1	1	1	1	1	1	1	1	1	0	0	0	0	1	1	0	1
KRTAP4-1	rs989703	R17Q	VNSCCGSVCSHQGC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRTAP4-1	rs113376	R26H	LCQETCCRPSCCETT	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
KRTAP9-2	rs990223	S56C	CRPTSCQNTCCR	1	1	1	0	0	1	1	1	1	1	0	0	1	1	1	1	1	1	0	1
KRTAP9-2	rs990223	S56C	CRPTcCQNTCCR	1	1	1	1	1	1	1	1	0	1	0	0	0	0	1	1	0	0	0	0
KRTAP9-4	rs219137	S146Y	R.TCYYPTTVCLPGCI	0	0	1	0	0	0	1	0	0	0	1	1	0	0	0	0	1	0	0	0
KRTAP9-4	rs219137	S146Y	*RTCYYPTTVCLPGCL*	0	1	1	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	1
KRTAP9-6	rs129386	Y86C	TTCCQPTCVTSCCQ	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0
KRTAP9-6	rs129386	Y86C	*TTCCQPTCVTSCCQ*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRTAP9-6	rs576405	Y145C	R.RTCYHPTTVCLPG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
KRTAP9-6	rs576405	Y145C	*R.RTCYHPTTVCLPG*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRTAP9-6	rs537301	C146R	R.RTCYHPTTVCLPG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
KRTAP9-6	rs537301	C146R	*R.RTCYHPTTVCLPG*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRTAP9-6	rs129383	C51Y	TTCWQPTIVTTCSSTP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRTAP9-6	rs129383	C51Y	TTCWQPTIVTTCSSTP	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
KRTAP10	rs233252	C170Y	*STCCVPIPSCCAPAS*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
KRTAP10	rs233252	C170Y	*STyCVPIPSCCAPAS*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRTAP10	rs464391	R268P	PASCVSLLCRPACSRL	0	0	0	0	0	0	1	1	0	0	1	1	0	0	1	0	0	1	0	0
KRTAP10	rs464391	R268P	PASCVSLLCRPACSpL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRTAP10	rs465279	S300P	SSSSVSLLCHPVCK	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0
KRTAP10	rs111668	V24M	MADACCTRTYVIAAS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
KRTAP10	rs111668	V24M	MADACCTRTYVIAAS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRTAP10	rs411254	H26R	TYVIAASTMSVCSS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
KRTAP10	rs411254	H26R	TYVIAASTMSVCSS	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
KRTAP10	rs998012	C257R	PACCVPVSSCCAPTSS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRTAP10	rs998012	C257R	PACCVPVSSCCAPTSS	1	1	1	0	0	1	1	1	0	0	1	1	0	0	0	0	0	1	0	0
KRTAP10	rs481895	V158M	SVCYVPVCSGASTS	0	0	0	0	0	1	1	0	0	0	1	1	0	0	0	0	0	0	0	0
KRTAP10	rs481895	V158M	SVCYmPVCSGASTS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRTAP10	rs617459	C236Y	LASCGSLLCR	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	1
KRTAP10	rs617459	C236Y	LASCGSLLyR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRTAP10	rs343029	G226S	RVPVPSCCVPTSSCC	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	1
KRTAP10	rs343029	G226S	*RVPVPSCCVPTSSCC*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRTAP11	rs713213	R72Q	CIVPVAQVTTTSTTD	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
KRTAP11	rs713213	R72Q	CIVPVAQVTTTSTTD	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRTAP11	rs963684	C111S	QTTCISNPCSTTYSR	1	1	1	1	0	0	0	0	1	1	1	0	1	1	1	1	1	1	1	1
KRTAP11	rs963684	C111S	QTTCISNPCSTTYSR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRTAP13	rs380401	S74R	R.GCQEICWEPTSC	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
KRTAP13	rs380401	S74R	*R.GCQEICWEPTSC*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRTAP16	rs207428	P340R	RCPSVCPEPVSCPSTS	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1
KRTAP16	rs207428	P340R	RCrSVCPEPVSCPSTSC	1	1	1	1	1	1	1	1	0	1	0	0	1	1	1	1	0	0	0	1
LAMP1	rs957723	I309T	FFLQGIQLNTILPDAR	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
LAMP1	rs957723	I309T	FFLQGIQLNTtLPDAR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LRRC15	rs130606	V270L	LYLSNNHISQLPPSV	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	0	0

LRRC15	rs130606	V270L	LYLSNNHISQLPPSIF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
LRRC15	rs130705	P286L	ELSlGIFGPMPNLR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LRRC15	rs130705	P286L	ELSPGIFGPMPNLR	0	0	1	0	1	0	0	1	0	0	0	0	1	1	1	1	1	1	0	1
LGALS3	rs101483	R183K	LDNNWGR	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1
LGALS3	rs101483	R183K	LDNNWGk	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LGALS3	rs11125_	Q201H	IQVLVEPDHFK	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1
LGALS3	rs11125_	Q201H	IhVLVEPDHFK	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
NEU2	rs223338	S11R	ESVFQSGAHAYR	0	0	0	0	1	0	1	1	0	0	0	0	1	1	1	1	1	1	0	1
NEU2	rs223338	S11R	ASLPVLQKEr	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NEU2	rs223338	R41Q	IPALLYLPGQQSLLAFA	0	0	0	0	1	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
NEU2	rs223338	R41Q	IPALLYLPGQQSLLAFA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NEU2	rs223339	A145T	DLTDAAIGPAYR	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	1
NEU2	rs223339	A145T	DLTDtAIGPAYR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NEU2	rs223339	H168N	EWSTFAVGPGHCLC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
NEU2	rs223339	H168N	EWSTFAVGPGHCLC	0	0	0	0	1	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0
PCM1	rs412750	S159N	DASTSPPNR	0	1	1	0	0	1	0	0	0	1	0	0	0	0	1	0	1	0	0	0
PCM1	rs412750	S159N	DASTnPPNR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PKP1	rs618182	R684W	AAEAARLLLSDMWS	1	1	1	1	1	0	1	1	1	1	0	0	1	1	1	1	1	1	1	1
PKP1	rs618182	R684W	AAEAAwLLLSDMW	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PKP1	rs109201	A442V	NYSGLIDSLMAYVQN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
PKP1	rs109201	A442V	NYSGLIDSLMAYVQN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PLB1	rs675392	V167L	AFVNLVDLSEVAEVSR	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
PLB1	rs675392	V167L	AFlNLVDLSEVAEVSR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PLCD1	rs933135	R257H	EEAAGPALALSLIER	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
PLCD1	rs933135	R257H	EEAAGPALALSLIEhYE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PPL	rs2037912	Q1573E	QNLQLETR	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
PPL	rs2037912	Q1573E	eNLQLETR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PPL	rs143676	R1457Q	VVLQQDPQQAREH	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PPL	rs143676	R1457Q	VVLQQDPQQAqEH	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S100A3	rs116208	L62V	FMSVLDTNKDCEVD	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
S100A3	rs360227	R3K	ARPLEQAVAAIVCTF	1	0	1	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	1
S100A3	rs360227	R3K	AkPLEQAVAAIVCTF	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S100A3	rs412651	H87Q	SLACLCLYCHEYFKD	1	0	0	0	0	0	1	1	0	0	0	0	1	1	1	1	1	1	0	1
SERPINB5	rs145555	I319V	GVALSNVIHK	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
SERPINB5	rs145555	I319V	GVALSNVvHK	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SYNGR2	rs142608	A28S	FLTQPQVVAR	1	0	1	0	0	0	1	0	1	1	0	0	1	0	0	0	0	0	1	0
SYNGR2	rs142608	A28S	FLTQPQVVsR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TCHH	rs251566	L63R	TVDLILELLDLDSNGF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TCHH	rs251566	L63R	TVDLILELLDr	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
TGM3	rs214803	T13K	AALGVQSINWQkAF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TGM3	rs214803	T13K	AALGVQSINWQTAF	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	0	0
TGM3	rs214814	S249N	SWNGSVEILK	1	1	1	0	1	0	1	1	0	1	0	0	0	0	1	1	1	1	0	1
TGM3	rs214814	S249N	nWNGSVEILK	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
VSIG8	rs626244	V47I	R.LGCPYVLDPEDYG	1	0	0	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	1
VSIG8	rs626244	V47I	*R.LGCPYiLDPEDYG*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

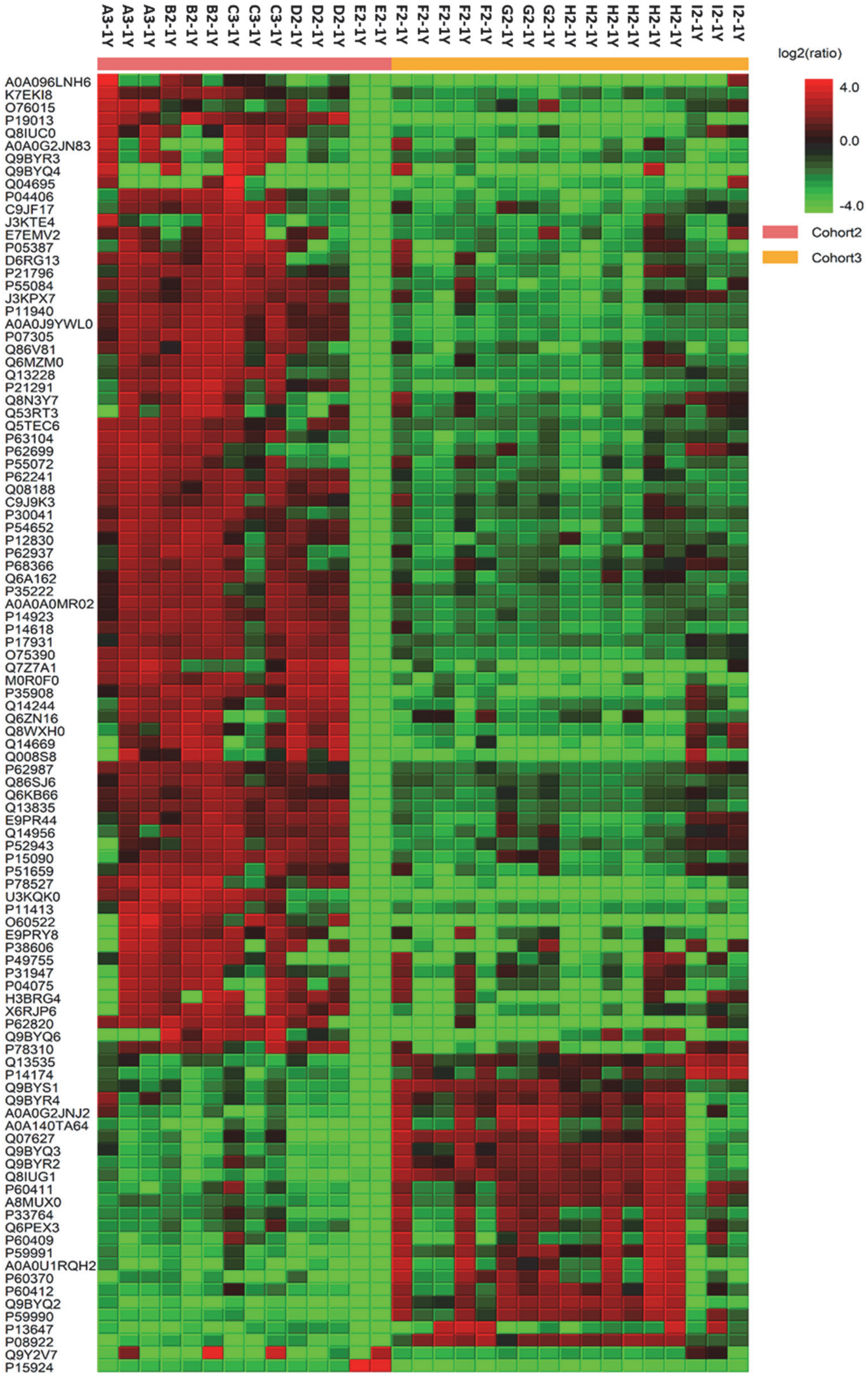


Figure S1. Heatmap showing at least 2 fold (p = 0.05) difference in the levels of proteins between the most recent samples of cohort 2 and cohort 3 emphasizing the batch effect on the proteomic profiling. The entries on the y axis denote the Uniprot IDs of the proteins while each column is a different sample. The numbers after the hyphens in the sample names represent the time of sample storage (1Y = 1 year).

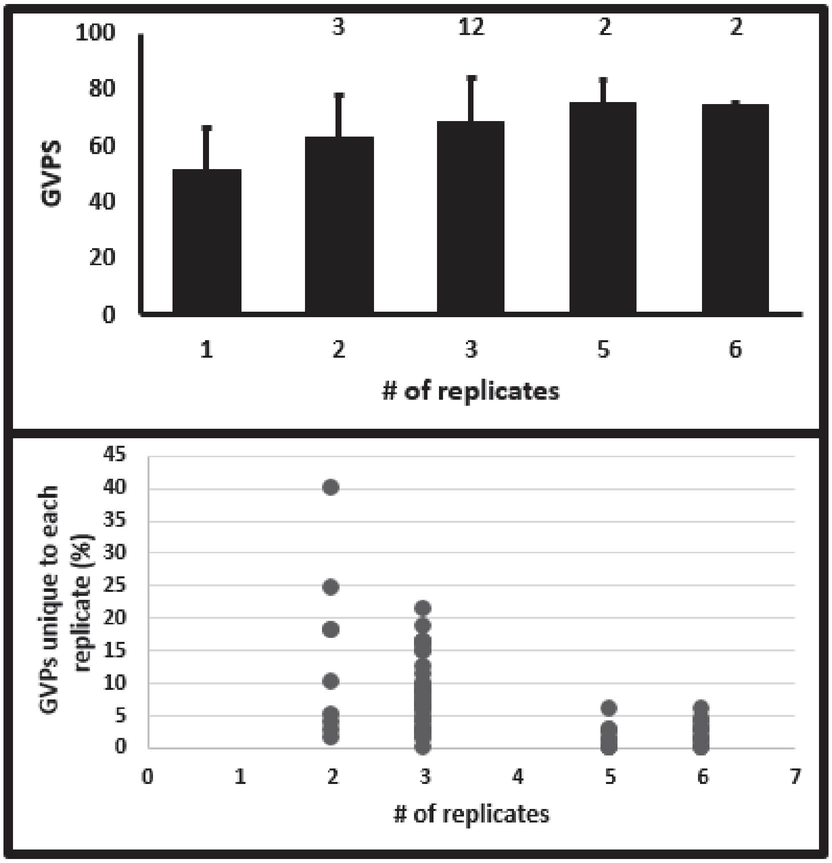


Figure S2. GVPs vs the number of replicates employed. The top panel presents the average number of GVPs identified vs the number of replicates used, while the bottom panel shows the percent GVPs unique to a replicate when 2, 3, 5 and 6 replicates were used. The number on the top of each bar indicates the number of different sample files analyzed for each scenario.

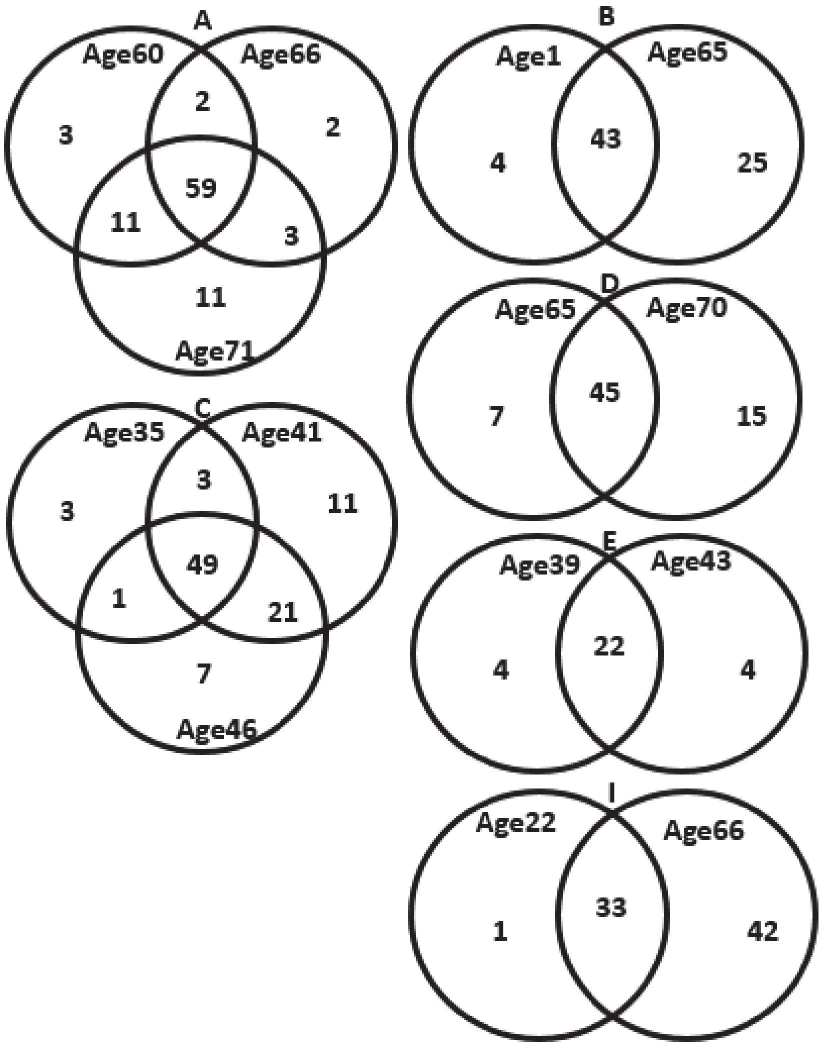


Figure S3. Number of GVPs common to samples at different ages or unique to a sample. Venn diagrams for each of the individuals are labeled on top of each diagram. The ages written at the tops of the circles represent ages of the individuals at the time of collection of samples.

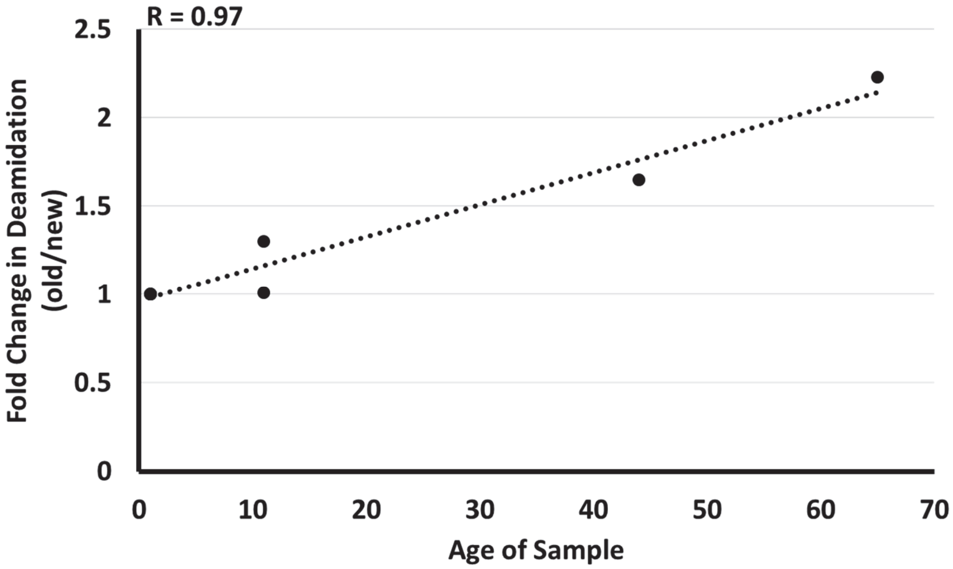


Figure S4. Deamidation of Q and N residues in proteins of hair samples stored for at least 10 years. Samples collected at different age points from individuals A, B, C and I were compared.

R code - Statistical Analysis - Age, Race

```
library(gdata)

library(edgeR)

library(dplyr)

library(RColorBrewer)


dat <- read.xls("WeightNotNorm-Ages.xlsx", stringsAsFactors = F, nrow = 261)

drop <- which(unlist(lapply(dat, function(x) all(is.na(x)))))

dat <- dat[,-drop]

anno <- dat[,1:4]


counts <- dat[,5:ncol(dat)]

rownames(counts) <- dat$Accession.Number


d <- DGEList(counts)

d <- calcNormFactors(d)


group <- unlist(lapply(strsplit(colnames(counts), split = ".", fixed = T),

                function(x)x[1]))


mm <- model.matrix(~0 + group)

y <- voom(d, mm, plot = T)


fit <- lmFit(y, mm)


# A1 vs A2

contr <- makeContrasts("groupA2 - groupA1", levels = colnames(coef(fit)))

tmp <- contrasts.fit(fit, contr)

tmp <- eBayes(tmp)
```

```
tmp2 <- topTable(tmp, n = Inf, sort.by = "P")
tmp2$Accession.Number <- rownames(tmp2)
tmp2 <- left_join(tmp2, anno)
tmp2 <- select(tmp2, Accession.Number, logFC, P.Value, adj.P.Val,
            Identified.Proteins)
write.csv(tmp2,file = "A2_v_A1.csv", row.names = F)

# B1 vs B2
contr <- makeContrasts("groupB2 - groupB1", levels = colnames(coef(fit)))
tmp <- contrasts.fit(fit, contr)
tmp <- eBayes(tmp)
tmp2 <- topTable(tmp, n = Inf, sort.by = "P")
tmp2$Accession.Number <- rownames(tmp2)
tmp2 <- left_join(tmp2, anno)
tmp2 <- select(tmp2, Accession.Number, logFC, P.Value, adj.P.Val,
            Identified.Proteins)
write.csv(tmp2,file = "B2_v_B1.csv", row.names = F)

# C1 vs C2
contr <- makeContrasts("groupC2 - groupC1", levels = colnames(coef(fit)))
tmp <- contrasts.fit(fit, contr)
tmp <- eBayes(tmp)
tmp2 <- topTable(tmp, n = Inf, sort.by = "P")
tmp2$Accession.Number <- rownames(tmp2)
tmp2 <- left_join(tmp2, anno)
tmp2 <- select(tmp2, Accession.Number, logFC, P.Value, adj.P.Val,
            Identified.Proteins)
write.csv(tmp2,file = "C2_v_C1.csv", row.names = F)
```

```
# A1 vs B1
contr <- makeContrasts("groupB1 - groupA1", levels = colnames(coef(fit)))
tmp <- contrasts.fit(fit, contr)
tmp <- eBayes(tmp)
tmp2 <- topTable(tmp, n = Inf, sort.by = "P")
tmp2$Accession.Number <- rownames(tmp2)
tmp2 <- left_join(tmp2, anno)
tmp2 <- select(tmp2, Accession.Number, logFC, P.Value, adj.P.Val,
         Identified.Proteins)
write.csv(tmp2,file = "B1_v_A1.csv", row.names = F)

# B1 vs C1
contr <- makeContrasts("groupC1 - groupB1", levels = colnames(coef(fit)))
tmp <- contrasts.fit(fit, contr)
tmp <- eBayes(tmp)
tmp2 <- topTable(tmp, n = Inf, sort.by = "P")
tmp2$Accession.Number <- rownames(tmp2)
tmp2 <- left_join(tmp2, anno)
tmp2 <- select(tmp2, Accession.Number, logFC, P.Value, adj.P.Val,
         Identified.Proteins)
write.csv(tmp2,file = "C1_v_B1.csv", row.names = F)

# A1 vs C1
contr <- makeContrasts("groupC1 - groupA1", levels = colnames(coef(fit)))
tmp <- contrasts.fit(fit, contr)
tmp <- eBayes(tmp)
tmp2 <- topTable(tmp, n = Inf, sort.by = "P")
tmp2$Accession.Number <- rownames(tmp2)
tmp2 <- left_join(tmp2, anno)
```

```
tmp2 <- select(tmp2, Accession.Number, logFC, P.Value, adj.P.Val,
            Identified.Proteins)
write.csv(tmp2,file = "C1_v_A1.csv", row.names = F)

# A2 vs B2
contr <- makeContrasts("groupB2 - groupA2", levels = colnames(coef(fit)))
tmp <- contrasts.fit(fit, contr)
tmp <- eBayes(tmp)
tmp2 <- topTable(tmp, n = Inf, sort.by = "P")
tmp2$Accession.Number <- rownames(tmp2)
tmp2 <- left_join(tmp2, anno)
tmp2 <- select(tmp2, Accession.Number, logFC, P.Value, adj.P.Val,
            Identified.Proteins)
write.csv(tmp2,file = "B2_v_A2.csv", row.names = F)

# B2 vs C2
contr <- makeContrasts("groupC2 - groupB2", levels = colnames(coef(fit)))
tmp <- contrasts.fit(fit, contr)
tmp <- eBayes(tmp)
tmp2 <- topTable(tmp, n = Inf, sort.by = "P")
tmp2$Accession.Number <- rownames(tmp2)
tmp2 <- left_join(tmp2, anno)
tmp2 <- select(tmp2, Accession.Number, logFC, P.Value, adj.P.Val,
            Identified.Proteins)
write.csv(tmp2,file = "C2_v_B2.csv", row.names = F)

# A2 vs C2
contr <- makeContrasts("groupC2 - groupA2", levels = colnames(coef(fit)))
tmp <- contrasts.fit(fit, contr)
tmp <- eBayes(tmp)

tmp2 <- topTable(tmp, n = Inf, sort.by = "P")

tmp2$Accession.Number <- rownames(tmp2)

tmp2 <- left_join(tmp2, anno)

tmp2 <- select(tmp2, Accession.Number, logFC, P.Value, adj.P.Val,

           Identified.Proteins)

write.csv(tmp2,file = "C2_v_A2.csv", row.names = F)


######

age <- substr(group, 1, 1)


mm <- model.matrix(~0 + age)

y <- voom(d, mm, plot = T)


fit <- lmFit(y, mm)


# A vs all B

contr <- makeContrasts("ageB - ageA", levels = colnames(coef(fit)))

tmp <- contrasts.fit(fit, contr)

tmp <- eBayes(tmp)

tmp2 <- topTable(tmp, n = Inf, sort.by = "P")

tmp2$Accession.Number <- rownames(tmp2)

tmp2 <- left_join(tmp2, anno)

tmp2 <- select(tmp2, Accession.Number, logFC, P.Value, adj.P.Val,

           Identified.Proteins)

write.csv(tmp2,file = "B_v_A.csv", row.names = F)


# B vs all C

contr <- makeContrasts("ageC - ageB", levels = colnames(coef(fit)))
```

```
tmp <- contrasts.fit(fit, contr)
tmp <- eBayes(tmp)
tmp2 <- topTable(tmp, n = Inf, sort.by = "P")
tmp2$Accession.Number <- rownames(tmp2)
tmp2 <- left_join(tmp2, anno)
tmp2 <- select(tmp2, Accession.Number, logFC, P.Value, adj.P.Val,
        Identified.Proteins)
write.csv(tmp2,file = "C_v_B.csv", row.names = F)

# A vs all C
contr <- makeContrasts("ageC - ageA", levels = colnames(coef(fit)))
tmp <- contrasts.fit(fit, contr)
tmp <- eBayes(tmp)
tmp2 <- topTable(tmp, n = Inf, sort.by = "P")
tmp2$Accession.Number <- rownames(tmp2)
tmp2 <- left_join(tmp2, anno)
tmp2 <- select(tmp2, Accession.Number, logFC, P.Value, adj.P.Val,
        Identified.Proteins)
write.csv(tmp2,file = "C_v_A.csv", row.names = F)

# MDS plot
cols <- brewer.pal(6, "Dark2")
tiff("MDS_age_race.tiff")
plotMDS(d, labels = group, col = cols[as.numeric(factor(group))])
dev.off()

# all 1 vs. all 2
race <- substr(group, 2, 2)
```

```
mm <- model.matrix(~race)
y <- voom(d, mm, plot = T)

fit <- lmFit(y, mm)

# A vs all B
tmp <- contrasts.fit(fit, coef = 2)
tmp <- eBayes(tmp)
tmp2 <- topTable(tmp, n = Inf, sort.by = "P")
tmp2$Accession.Number <- rownames(tmp2)
tmp2 <- left_join(tmp2, anno)
tmp2 <- select(tmp2, Accession.Number, logFC, P.Value, adj.P.Val,
             Identified.Proteins)
write.csv(tmp2,file = "2_v_1.csv", row.names = F)
```

R code - Statistical Analysis - Individuals

```
library(gdata)
dat <- read.xls("ProfilesVsAge.xlsx", stringsAsFactors = F, skip = 1, nrow = 242, check.names = F)

dat2 <- dat
drop <- which(names(dat2) == "")
dat2 <- dat2[,-drop]
dat2[,5:73] <- lapply(dat2[,5:73], function(x)gsub(",", "", x, fixed = T))
counts <- data.matrix(dat2[,5:73])

library(edgeR)
d <- DGEList(counts)
d <- calcNormFactors(d)
rownames(d) <- dat$`#`

pdata <- read.xls("hair_aging_sample_info.xlsx", stringsAsFactors = F)
identical(pdata$sample, colnames(d))

# boxplot(d$sample$norm.factors ~ pdata$processed_by)

# Calculate batch-adjusted MDS plot
library(RColorBrewer)
cpms <- cpm(d, log = T)
resids <- t(apply(cpms, 1, function(x)resid(lm(x ~ processing_batch, data = pdata))))
cols <- c("black", brewer.pal(8, "Set2"))
tiff("./figures/MDS_batch_adjusted_by_subject_and_year.tiff", width = 8, height = 8, res = 400, units = "in")
plotMDS(resids, col = cols[as.numeric(factor(pdata$subject))], labels = pdata$collection_year)
legend("right", text.col = cols, legend = levels(factor(pdata$subject)), title = "Subject")
```

```
dev.off()

tiff("./figures/MDS_batch_adjusted_by_subject_and_sample.tiff", width = 8, height = 8, res = 400, units
= "in")

plotMDS(resids, col = cols[as.numeric(factor(pdata$subject))], labels = colnames(cpms))

legend("right", text.col = cols, legend = levels(factor(pdata$subject)), title = "Subject")

dev.off()


# derive time since sample collection as 2017 - year, or 2018 - year if second batch

pdata$sampage <- ifelse(pdata$processed_by == "TJP", 2017 - pdata$collection_year,

                     2018 - pdata$collection_year)


# Derive hair sample

pdata$hair <- substr(pdata$sample, 1, nchar(pdata$sample) - 1)


# Set age to 1 if lt 1

pdata$collection_age <- ifelse(pdata$collection_age == "< 1", 1, as.numeric(pdata$collection_age))


##########################################################################

######### Analysis by time since sample was collected


mm <- model.matrix(~sampage + subject, data = pdata)

y <- voom(d, mm, plot = T)


#####

write.csv(cbind(rownames(y), dat$Accession.Number, y$E), file = "normalized_counts.csv", row.names =
F)


#####
```

```
# Calculate within-hair correlations
cor <- duplicateCorrelation(y, mm, block = pdata$hair)$consensus

fit <- lmFit(y, mm, block = pdata$hair, correlation = cor)

# Estimate contrasts
#year
tmp <- contrasts.fit(fit, coef = 2)
tmp <- eBayes(tmp)
tmp2 <- topTable(tmp, sort.by = "P", n = Inf)
length(which(tmp2$adj.P.Val < 0.05))
anno <- dat[,1:4]
names(anno)[2] <- "Identified Proteins"
out <- merge(anno, tmp2, by.y = "row.names", by.x = "#")
out <- out[order(out$P.Value),c("Accession Number", "Identified Proteins", "MW", "logFC", "P.Value", "adj.P.Val")]
write.csv(out, "Protein_Expression_by_Years_Since_Collection_Results_ALL_SAMPLES.csv", row.names = F)

# Plot significant proteins by year
sigs <- rownames(tmp2)[which(tmp2$adj.P.Val < 0.05)]

f <- function(X){
	protein <- unlist(strsplit(dat$`Accession Number`[which(dat$`#` == X)], split = "|", fixed = T)[[1]])[3]
	x <- as.numeric(y$E[X,])
	plotname <- paste0("./figures/", protein, "_ALL_SAMPLES.tiff")
	tiff(plotname, width = 8, height = 8, res = 400, units = "in" )
	plot(x ~ collection_year, main = protein, xlab = "Year", ylab = "Normalized Expression", data = pdata)
```

```
            abline(lsfit(pdata$collection_year, x), col = 2)
  dev.off()
  drop <- which(pdata$hair == "R")
            plotname <- gsub("_ALL_SAMPLES", "_NO_SAMPLE_R", plotname)
            tiff(plotname, width = 8, height = 8, res = 400, units = "in" )
            plot(x[-drop] ~ pdata$collection_year[-drop],
                xlab = "Year", ylab = "Normalized Expression", main = protein)
            abline(lsfit(pdata$collection_year[-drop], x[-drop]), col = 2)
            dev.off()
}
sapply(sigs, f)


# Refit model without hair R
drop <- which(pdata$hair == "R")
mm <- model.matrix(~sampage + subject, data = pdata[-drop,])
y.no1951 <- voom(d[,-drop], mm, plot = T)
cor <- duplicateCorrelation(y.no1951, mm, block = pdata$hair[-drop])$consensus
fit <- lmFit(y.no1951, mm, block = pdata$hair[-drop], correlation = cor)
tmp <- contrasts.fit(fit, coef = 2)
tmp <- eBayes(tmp)
tmp2 <- topTable(tmp, sort.by = "P", n = Inf)
length(which(tmp2$adj.P.Val < 0.05))
anno <- dat[,1:4]
names(anno)[2] <- "Identified Proteins"
out <- merge(anno, tmp2, by.y = "row.names", by.x = "#")
out <- out[order(out$P.Value),c("Accession Number", "Identified Proteins", "MW", "logFC", "P.Value", "adj.P.Val")]
write.csv(out, "Protein_Expression_by_Years_Since_Collection_Results_NO_SAMPLE_R.csv", row.names = F)
```

```
#######################################################################
#######################################################################
# Analysis by subject age at collection
mm <- model.matrix(~collection_age + subject, data = pdata)
y <- voom(d, mm, plot = T)

# Calculate within-hair correlations
cor <- duplicateCorrelation(y, mm, block = pdata$hair)$consensus

fit <- lmFit(y, mm, block = pdata$hair, correlation = cor)

# Estimate contrasts
#year
tmp <- contrasts.fit(fit, coef = 2)
tmp <- eBayes(tmp)
tmp2 <- topTable(tmp, sort.by = "P", n = Inf)
length(which(tmp2$adj.P.Val < 0.05))
anno <- dat[,1:4]
names(anno)[2] <- "Identified Proteins"
out <- merge(anno, tmp2, by.y = "row.names", by.x = "#")
out <- out[order(out$P.Value),c("Accession Number", "Identified Proteins", "MW", "logFC", "P.Value", "adj.P.Val")]
write.csv(out, "Protein_Expression_by_Subject_Age_at_Collection_Results_ALL_SAMPLES.csv", row.names = F)

# Refit model without hair R
drop <- which(pdata$hair == "R")
mm <- model.matrix(~collection_age + subject, data = pdata[-drop,])
```

```
y.no1951 <- voom(d[,-drop], mm, plot = T)
cor <- duplicateCorrelation(y.no1951, mm, block = pdata$hair[-drop])$consensus
fit <- lmFit(y.no1951, mm, block = pdata$hair[-drop], correlation = cor)
tmp <- contrasts.fit(fit, coef = 2)
tmp <- eBayes(tmp)
tmp2 <- topTable(tmp, sort.by = "P", n = Inf)
length(which(tmp2$adj.P.Val < 0.05))
anno <- dat[,1:4]
names(anno)[2] <- "Identified Proteins"
out <- merge(anno, tmp2, by.y = "row.names", by.x = "#")
out <- out[order(out$P.Value),c("Accession Number", "Identified Proteins", "MW", "logFC", "P.Value", "adj.P.Val")]
write.csv(out, "Protein_Expression_by_Subject_Age_at_Collection_Results_NO_SAMPLE_R.csv", row.names = F)



cor(pdata$collection_age, pdata$collection_year)



################# Pairwise contrasts between hairs, within each batch
mm <- model.matrix(~0 + hair, data = pdata)
y <- voom(d, mm, plot = T)

fit <- lmFit(y, mm)

# Estimate contrasts--pairwise comparisons of all hairs
samps <- unique(pdata$hair[pdata$processed_by == "TJP"])
nsamp <- length(samps)
out <- dat[,c("Accession Number", "Identified Proteins (467)", "MW")]
```

```
names(out)[2] <- "Identified Proteins"

nsig <- matrix(nrow = nsamp, ncol = nsamp)

for (i in 1:(nsamp - 1)){

  for (j in (i + 1):nsamp){

    cont <- paste("hair", samps[i], " - hair", samps[j], sep = "")

    contr <- makeContrasts(cont, levels = colnames(coef(fit)))

    tmp <- contrasts.fit(fit, contr)

    tmp <- eBayes(tmp)

    tmp2 <- topTable(tmp, sort.by = "none", n = Inf)

    nsig[i, j] <- nsig[j, i] <- length(which(tmp2$adj.P.Val < 0.05))

    names(tmp2) <- paste(names(tmp2), samps[i], "v", samps[j], sep = ".")

    out <- cbind(out, tmp2[,c(1,4,5)])

  }

}

samps <- unique(pdata$hair[pdata$processed_by == "RHR"])

nsamp <- length(samps)

out <- dat[,c("Accession Number", "Identified Proteins (467)", "MW")]

names(out)[2] <- "Identified Proteins"

nsig <- matrix(nrow = nsamp, ncol = nsamp)

for (i in 1:(nsamp - 1)){

  for (j in (i + 1):nsamp){

    cont <- paste("hair", samps[i], " - hair", samps[j], sep = "")

    contr <- makeContrasts(cont, levels = colnames(coef(fit)))

    tmp <- contrasts.fit(fit, contr)

    tmp <- eBayes(tmp)

    tmp2 <- topTable(tmp, sort.by = "none", n = Inf)

    nsig[i, j] <- nsig[j, i] <- length(which(tmp2$adj.P.Val < 0.05))

    names(tmp2) <- paste(names(tmp2), samps[i], "v", samps[j], sep = ".")

    out <- cbind(out, tmp2[,c(1,4,5)])
  }
}
rownames(nsig) <- colnames(nsig) <- samps


library(openxlsx)
wb <- createWorkbook()
addWorksheet(wb, "Results of Pairwise Comparisons")
writeData(wb, "Results of Pairwise Comparisons", out)
posStyle <- createStyle(fontColour = "#006100", bgFill = "#C6EFCE")
pvalcols <- grep("adj", names(out))
sapply(pvalcols,function(x) conditionalFormatting(wb, "Results of Pairwise Comparisons", cols = x, rows
= 1:nrow(out),
                                          rule = "<0.05", style = posStyle))
addWorksheet(wb, "Num Sig Comparisons")
writeData(wb, "Num Sig Comparisons", nsig, rowNames = T)
Sys.setenv(R_ZIPCMD= "C:/Rtools/bin/zip")
saveWorkbook(wb, "Pairwise Comparisons Between Samples.xlsx", overwrite = TRUE)


#######################################################################################
###
#######################################################################################
###
# subject-time interaction
mm <- model.matrix(~sampage*subject, data = pdata)
y <- voom(d, mm, plot = T)


# Calculate within-sample correlations
cor <- duplicateCorrelation(y, mm, block = pdata$hair)$consensus


fit <- lmFit(y, mm, block = pdata$hair, correlation = cor)
```

```
# Estimate contrasts
f <- function(subject){
  if (subject == "A"){
    con <- "sampage"
  }else{
    con <- paste0("sampage + sampage.subject", subject)
  }
  contr <- do.call(makeContrasts, list(contrasts = con, levels = make.names(colnames(coef(fit)))))
  rownames(contr) <- colnames(coef(fit))
  tmp <- contrasts.fit(fit, contr)
  tmp <- eBayes(tmp)
  results <- topTable(tmp, sort.by = "none", n = Inf)[,c("logFC","P.Value","adj.P.Val")]
  names(results) <- paste(names(results), subject, sep = ".")
  return(results)
}
subs <- unique(pdata$subject)
out <- lapply(subs, f)

# Merge files
results <- do.call(cbind, out)
anno <- dat[,1:4]
out <- merge(anno, results, by.y = "row.names", by.x = "#")
library(openxlsx)
wb <- createWorkbook()
addWorksheet(wb, "Results")
writeData(wb, "Results", out)
posStyle <- createStyle(fontColour = "#006100", bgFill = "#C6EFCE")
pvalcols <- grep("adj", names(out))
```

```
sapply(pvalcols,function(x) conditionalFormatting(wb, "Results", cols = x, rows = 1:nrow(out),
                            rule = "<0.05", style = posStyle))
Sys.setenv(R_ZIPCMD= "C:/Rtools/bin/zip")
saveWorkbook(wb, "Subject by Time Since Sample Collection Interaction Model.xlsx", overwrite = TRUE)


############# Plots of distances
cpms <- cpm(d, log = T)
resids <- t(apply(cpms, 1, function(x)resid(lm(x ~ processing_batch, data = pdata))))

d <- dist(t(resids), diag = T)
d2 <- as.matrix(d)

subs <- unique(pdata$subject)
nsub <- length(subs)
between.subject.dists <- NULL
between.subject.names <- NULL
within.subject.dists <- NULL
within.subject.names <- NULL

for (i in 1:nsub){
  for (j in 1:i){
    subject1 <- subs[i]
    subject2 <- subs[j]
    if (i == j){
      t1 <- which(pdata$subject == subject1)
      #
      tmp <- d2[t1, t1]
      tmp0 <- as.numeric(tmp[lower.tri(tmp)])
      within.subject.dists <- c(within.subject.dists, tmp0)
      pairname <- paste(subject1, subject1, sep = ".")
      within.subject.names <- c(within.subject.names, rep(pairname, length(tmp0)))
    } else{
      t1 <- which(pdata$subject == subject1)
      t2 <- which(pdata$subject == subject2)
      tmp <- d2[t1, t2]
      tmp0 <- as.numeric(tmp)
      between.subject.dists <- c(between.subject.dists, tmp0)
      pairname <- paste(subject1, subject1, sep = ".")
      between.subject.names <- c(between.subject.names, rep(pairname, length(tmp0)))
    }
  }
}
names(within.subject.dists) <- within.subject.names
names(between.subject.dists) <- between.subject.names

avg.within.subject <- tapply(within.subject.dists, names(within.subject.dists),
                            function(x)sqrt(mean(x^2)))
avg.between.subject <- tapply(between.subject.dists, names(between.subject.dists),
                             function(x)sqrt(mean(x^2)))
tiff("./figures/Distance Boxplots.tiff", width = 8, height = 8, res = 400, units = "in")
boxplot(list(avg.within.subject, avg.between.subject), beside = T,
        ylab = "Average Distance", xaxt = "n")
axis(1, at = 1:2, labels = c("Within Subjects", "Between Subjects"),line = 1, tick = F)
dev.off()
```