Published in final edited form as:

Proteomics. 2017 July ; 17(13-14): . doi:10.1002/pmic.201600462.

Proteomic analysis of hair shafts from monozygotic twins: Expression profiles and genetically variant peptides

Pei-Wen Wu¹, Katelyn E. Mason², Blythe P. Durbin-Johnson³, Michelle Salemi⁴, Brett S. Phinney⁴, David M. Rocke³, Glendon J. Parker^{2,5}, and Robert H. Rice^{*,1}

¹Forensic Science Graduate Program and Department of Environmental Toxicology, University of California, Davis

²Forensic Science Center, Lawrence Livermore National Laboratory

³Division of Biostatistics, Department of Public Health Sciences, Clinical and Translational Science Center Biostatistics Core, University of California, Davis

⁴Proteomics Core Facility, University of California, Davis

⁵Protein-Based Identification Technologies

Abstract

Forensic association of hair shaft evidence with individuals is currently assessed by comparing mitochondrial DNA haplotypes of reference and casework samples, primarily for exclusionary purposes. Present work tests and validates more recent proteomic approaches to extract quantitative transcriptional and genetic information from hair samples of monozygotic twin pairs, which would be predicted to partition away from unrelated individuals if the datasets contain identifying information. Protein expression profiles and polymorphic, genetically variant hair peptides were generated from 10 pairs of monozygotic twins. Profiling using the protein tryptic digests revealed that samples from identical twins had typically an order of magnitude fewer protein expression differences than unrelated individuals. The data did not indicate that the degree of difference within twin pairs increased with age. In parallel, data from the digests were used to detect genetically variant peptides that result from common non-synonymous single nucleotide polymorphisms in genes expressed in the hair follicle. Compilation of the variants permitted sorting of the samples by hierarchical clustering, permitting accurate matching of twin pairs. The results demonstrate that genetic differences are detectable by proteomic methods and provide a framework for developing quantitative statistical estimates of personal identification that increase the value of hair shaft evidence.

Keywords

Agglomerative hierarchical clustering; Forensic evidence; Genetically variant peptides; Monozygotic twin pairs; Protein profiling

*Correspondence: Department of Environmental Toxicology, University of California, One Shields Avenue, Davis, CA 95616-8588; Tel. 530-752-5176, rhrice@ucdavis.edu.

Conflict of interest statement

The authors have declared no conflict of interest.

1 Introduction

Hair evidence can be an important element in criminal investigations since it is commonly recovered at crime scenes, durable and easily transferred between individuals and inanimate objects [1]. However, microscopic comparison is not amenable to rigorous quantitative evaluation due to its subjective nature [2]. Mitochondrial DNA is commonly used in view of difficulties in obtaining sufficient nuclear DNA [3], but it provides exclusionary information of limited discrimination ability due to its maternal inheritance pattern and haplotype distribution in sample populations [4, 5]. Present work explores a more quantitative approach for distinguishing among individuals based on proteomic instead of DNA analysis.

A finding that inbred mouse strains are distinguishable by hair protein profiling [6] led to the realization that individual humans, constituting an outbred population, are also distinguishable in this way [7]. Using MZ (monozygotic or identical) twins, present work explores the degree to which protein expression levels are genetically determined in the human hair shaft, an important factor in application of such analysis to forensic science. While MZ twins have shared genotypes, they can also display phenotypic discordances. These are attributable to a number of factors, including epigenetic differences that could increase the discordance with age [8].

Beyond protein profiling, the hair proteome offers an opportunity to exploit genomic differences that are manifest as polymorphic or genetically variant peptides (GVPs) [9]. Common non-synonymous single nucleotide polymorphisms (SNPs) are transcribed as single amino acid polymorphisms that occur in GVPs. Identification of GVPs in proteomic datasets allows for imputation of SNP alleles. As demonstrated in the present work, the same mass spectrometric datasets permit both protein profiling and GVP analysis. Each depends on the genetics of the hair donor, but GVPs reflect coding and profiling mostly noncoding sequence. The combination has the potential to greatly increase the value of hair shaft evidence as a tool in human identification.

2 Materials and Methods

2.1 Human subjects

Primarily through SRI International's Twin Research Registry [10], 10 monozygotic (MZ) twin pairs (10 males and 10 females, ages 19 to 84) were recruited. Exclusion criteria included chemically treated hair (dyed, straightened, bleached) and balding syndromes. Qualified twins provided written consent, scalp hair samples and completed questionnaires, all collected by mail. To minimize a contribution of hair weathering to different protein profiles between hair distal to the scalp versus near its proximal end, most samples were collected from subjects with short hair (<10 inches). Three subjects with hair longer than 10 inches each provided a hair bundle with proximal end marked, the end sampled. Four samples from each subject were processed and analyzed at the same time. In compliance with ethical rules for research involving human subjects, the study was approved by the Institutional Review Board of the University of California, Davis (#491288).

2.2 Questionnaires

All subjects completed a questionnaire consisting of four categories of information: general (age, sex, hair length), similarity of lifestyles between the MZ twins (diet, sleep/work/exercise schedule, medical conditions, medications, work/home environment, level of stress), hair care habits (brushing, washing, hair care products) and whether the twin and his/her sibling shared the same placenta at birth.

2.3 Hair processing with detergent removal

Hair was processed as previously described [7, 11, 12] except that sodium dodecyl sulfate was replaced by sodium dodecanoate (SD). Hair samples were chopped, weighed (4 mg), rinsed twice in 2% SD (w/v) in 0.05 M ammonium bicarbonate to remove dust and debris, heated overnight at 70°C in 0.4 ml of the same buffer adjusted to 25 mM in dithioerythritol, and then alkylated with iodoacetamide for 45 min at room temperature with magnetic stirring in the dark. To avoid denaturation of trypsin, SD was removed by ethyl acetate extraction after acidification to pH 2–3 with trifluoroacetic acid (0.75%). After the upper organic phase was removed, the lower aqueous phase was readjusted to pH 8 with ammonium bicarbonate, and the protein was digested by daily addition of 40 µg of reductively methylated trypsin [13]. The digests were clarified by centrifugation and submitted for mass spectrometric analysis.

2.4 Mass spectrometry and protein identification

Digested peptides were analyzed by LC-MS/MS with a Thermo Scientific Q Exactive Orbitrap mass spectrometer fitted with a Proxeon Easy-nLC II HPLC (Thermo Scientific, Asheville, NC, USA) and a Proxeon nanospray source. The digested peptides were loaded on a 100 micron x 25 mm Magic C18 100Å 5U reverse phase trap where they were desalted online before being separated using a 75 micron x 150 mm Magic C18 200Å 3U reverse phase column. Peptides were eluted using a 90 minute gradient with a flow rate of 300 nL/min. An MS survey scan was obtained for the m/z range 300–1600, and MS/MS spectra were acquired where the top 15 ions in the MS spectra were subjected to high energy collisional dissociation. An isolation mass window of 2.0 m/z was used for the precursor ion selection, and normalized collision energy of 27% was used for fragmentation. A 5 sec duration was used for the dynamic exclusion.

Tandem mass spectra were extracted and charge states were deconvoluted and deisotoped. Sample data were analyzed using X!Tandem (thegpm.org; version CYCLONE 2013.02.01.1) to search a Uniprot human database appended to a database of common non-human contaminants (cRAP, common Repository of Adventitious Proteins database, <http://www.thegpm.org/crap/>). Both were appended to an identical but reversed database for calculating decoy false discovery rates (138,090 proteins total), assuming the digestion enzyme was trypsin. X!Tandem was searched with a fragment ion mass tolerance of 20 PPM and a parent ion tolerance of 20 PPM. Iodoacetamide derivative of cysteine was specified in X!Tandem as a fixed modification. Glu/Gln → pyro-Glu (N-term), deamidation of asparagine and glutamine, oxidation of methionine and tryptophan, dioxidation of methionine and tryptophan, N-terminal acetylation, and ammonia loss were specified in X!Tandem as variable modifications. Scaffold version 4.2.1 was used to validate MS/MS based

peptide and protein identifications. Peptide identifications were accepted if they could be established at greater than 94% probability to achieve an FDR less than 1% by the Scaffold Local FDR algorithm. This achieved a peptide decoy false discovery rate of 0.2%. Protein identifications were accepted to achieve an FDR less than 5% and to contain at least two identified peptides. This achieved an overall protein decoy false discovery rate of 0.9%. Since certain peptides can overlap, especially among the keratins, numbers of distributed (called weighted in Scaffold) spectral counts were tabulated using experiment-wide grouping as previously described [7]. Spectral counts provide a reliable measure of relative abundance over a wide dynamic range [14, 15]. While label-free quantitation methods generally are least able to detect protein variations of low amplitude [16], the present fold changes considered significantly different were all beyond the minimum considered reliable [15]. Scaffold's weighed spectral counts appropriate a percentage of each count divided among the protein groups that share that peptide. Thus, the percentage for peptide(i) assigned to protein(j) is $PPS(j)/\text{sum}(PPS(1\dots n))$, where $PPS(j)$ is the sum of the peptides(1...m) for protein(j) and (1...n) is the set of proteins that contain peptide(i). The weighted spectrum count is the sum of those percentages for each protein group. In addition, every protein identified was required to have at least one unique peptide that was not shared with any other protein. The Scaffold file containing all the peptide data used in the analysis is available on the MassIVE repository (ID: MSV000080295; <http://massive.ucsd.edu>).

2.5 Detection of GVPs

Acquired proteomic datasets from each individual were converted into mzML format using MSconvertGUI (Proteowizard 2.1, <http://proteowizard.sourceforge.net>) and submitted to the Global Proteome Machine webserver (www.thegpm.org; X! Tandem Piledriver (2015.04.01.1)). Default search settings were used with the exception of a 20 ppm error for the primary scan, inclusion of complete cysteine carbamidomethylation (C+57), and partial modifications of oxidized methionine (M+16), and deamidation (N+1, Q+1). Detected spectra were filtered for quality and specificity. Quality measures include the log(e) score of less than -2.0, and a mass error of fragment ions of less than 0.4 Da. Characterized GVPs were identified in each dataset [9] and given a weighted value by determining the inverse of the minor genotypic frequency of the corresponding non-synonymous SNP in a sample European population (1000 Genomes Project, phase 3). When combining datasets from two replicates of an individual twin, presence of a GVP was determined by detection in either or both datasets, with no additional weighting for the second identification.

2.6 Statistical analysis

For protein profiling, count data were transformed using a variance stabilizing transformation for negative binomial data, which takes the form $f_{\theta}(x) = \ln [x + (x^2 + x/\theta)^{0.5} + (2\theta)^{-1}]$. This transformation, when θ is selected to minimize the correlation between the variance and standard deviation of the transformed data, removes mean-variance dependency from the data so that they may be analyzed using methods that assume constant variance across the range of data. Data were then analyzed using the Bioconductor package for gene expression analysis limma [17], which fits linear models to each protein separately and then applies empirical Bayes shrinkage to the estimated variances to increase power. The application of the above transformation to RNA-Seq data is discussed in Rocke et al. (<http://>

[biorxiv.org/content/biorxiv/early/2015/06/11/020784.full.pdf](https://www.biorxiv.org/content/biorxiv/early/2015/06/11/020784.full.pdf)). The overall analysis approach is similar to that called “limma-trans” [18], which employs the variance stabilizing transformation from the DESeq RNA-Seq analysis package [19]. The correlation between twins was estimated by fitting a linear mixed effects model using the R package nlme, version 3.1–128 [20]. The “consensus” correlation (the hyperbolic tangent of the 15% trimmed mean of the inverse hyperbolic tangent of the per-protein correlations) was then input into limma to adjust for between-twin correlation. Distances between samples were estimated using Euclidean distances of variance stabilized data. Analyses were conducted using R, version 3.2.1 [21].

For evaluation of GVPs, profiles from two biological replicates were combined and each detected peptide was assigned a value of one. Agglomerative hierarchical cluster (AHC) analysis was performed using statistical software package XLSTAT (version 2013.5.0, Addinsoft; New York, NY) to group twin samples that displayed co-occurrence in their GVP profiles. Implementing the settings co-occurrence, flexible linkage beta = −0.5 in the AHC algorithm, similarity values were obtained across all samples. Resulting similarity coefficients indicate the proximity between two samples calculated by measuring at what point they are similar. GVP profiles were weighted by transforming a list of binary collated values (unweighted) that correspond to the presence/absence of potential GVPs detected in mass spectrometry data, (1) if peptide was observed in a single twin sample or both, and (0) if peptide was absent, to a list of genotypic frequencies. Values of (1) were multiplied by the genotypic frequency of the corresponding single-nucleotide polymorphism resulting in the detected GVP [22]. This statistical analysis was repeated for both weighted and unweighted GVP profiles from each individual replicate. The resulting data were cross-compared with other individuals in the study.

3 Results

3.1 Distinguishing MZ sibling protein profiles

Profiles were analyzed focusing on 193 consistently identified proteins. Global testing within the 10 twin pairs revealed 16 instances of differential protein expression involving 14 different proteins. The specific proteins and their spectral counts are illustrated in Figure 1. Numbers of significant differences within each pair, as shown in Figure 2, ranged from 0 to 6 proteins (mean 1.6). As is evident in the age distribution of the subjects, the number of differences was not an obvious function of age. Information obtained from questionnaires soliciting information about lifestyles did not reveal factors that correlated with the observed differences (Supporting Table S1).

Previous work has shown that hair samples from unrelated individuals usually display significantly different levels of numerous proteins [7]. The present samples permitted comparison of the differences within MZ twin pairs to those among unrelated twins. A resampling-based comparison of differences between twins and non-twins was conducted. One twin was selected from each of the pairs, where these individuals were randomly assigned to artificial “twin pairs”, and then the result of differential expression was compared to analysis from the real MZ twin pairs. Of 50 random re-pairings of unrelated individuals, an average of 51 proteins was differentially expressed among artificial “twins”

and an average of 1 protein was differentially expressed between the same number of actual MZ twins.

To illustrate this observation more generally, the relative degree of difference is illustrated as a measure of distance between samples (Figure 3). As shown, the distance between different samples from the same person's hair was the same on average as that between samples from two MZ twins, whereas the distance between samples from unrelated twins was distinctly larger.

3.2 Analysis of GVPs among MZ twin pairs

Polymorphic peptides were identified in proteomic datasets from digests of hair samples from the twin subjects [9]. A tally was compiled with values weighted as the inverse of allelic frequency from the European populations (www.internationalgenome.org; Supporting Tables S2 and S3). From the data a proximity matrix was calculated in Euclidian space (Supporting Table S4) from which a dendrogram was constructed using agglomerative hierarchical clustering of individual samples from the 10 twin pairs. As illustrated in Figure 4, all twins clustered according to their MZ status. When the values were not weighted, the clustering was less efficient, with 8 of the 10 twin pairs partitioning in this way and individuals from two twin pairs associating within the range of values for other matched twin pairs (MZ pair D and I; Supporting Figure S1 and Supporting Table S5).

For visualization of relative similarities of GVP values in MZ twins versus unrelated individuals, Figure 5 presents a measure of distance based on GVP differences in the various hair samples. Distances between MZ twins were relatively small, as close on average as samples from a given individual were to each other. By contrast, distances between unrelated individuals calculated in this way were much larger (3 fold), validating the hypothesis that MZ twins would partition away from other individuals.

4 Discussion

As previously reported [7, 9], the proteomic approaches described here provide quantitative criteria for judging whether hair samples are derived from different individuals. As now demonstrated, the peptide data from mass spectrometric analysis of hair digests can be analyzed in parallel for levels of relative protein expression and for identification of GVPs. Further advances in expanding the range of proteins detected is also anticipated to increase individual discrimination by expression level profiling, including among MZ twins and may be enhanced by expanding the pool of characterized GVPs. Identification of GVPs is less sensitive to digestion yields than protein profiling; a single good scoring peptide assignment is sufficient. Expanding the analysis of GVP detection beyond biological duplicates, for instance to analyze all 4 replicates, would also increase the sensitivity and discrimination of the study. Since the level of protein digestion influences peptide yields, it may also influence the discriminatory power of GVP identification. When proteins show differential levels of expression, GVPs present in these proteins are more likely to fluctuate above and below the limits of detection. We observe that detection of some common GVPs, at the DNA level present in a majority of the population, occurred in only a single twin pair. In this manner GVP assignment data is influenced by protein expression levels, and potentially by other

biological processes, such as post-translational modification, protein turnover and protein-protein interaction.

Other statistical approaches can be employed to evaluate biodistance. Determination of genetic relationships at the DNA level, as used with paternity tests, requires a direct comparison of specific shared allele profiles and a calculation of the most likely genetic relationship [23]. The proximity, or similarity, measurements in this analysis instead compare the entire GVP profile. A comparison of the two approaches, an interesting investigative statistical question, is an ambitious project requiring a strict analysis of dependency patterns and subsequent down selection of available markers. High levels of correlation between different GVPs in this proximity analysis would result in some regions of the genome having greater weight than areas with less representation in the hair shaft proteome.

Inasmuch as MZ twins are much more difficult to distinguish than unrelated individuals by hair protein expression level, these data are consistent with those from inbred mouse strains indicating strong genetic control of expression [6]. Since a major factor in gene expression level is the transcription rate [24], this conclusion likely reflects the finding in the human genome of considerable variation in sequences of transcription factors, their binding sites in gene promoters and the chromatin states that result [25–28]. The influence of adult environment, including diet, on protein expression level appeared small or inconsistent in this study. A possibility that these levels might diverge with age of the subjects is not supported by the data. If the protein profiles change with age, the results indicate they change largely in the same direction. Limited divergence in profile due to environmental or stochastic effects during development or childhood are not ruled out, however.

5 Conclusion

The present study provides a foundation for distinguishing among individuals by proteomic analysis of their hair shafts. Further study of the influence of age and environment (including weathering and chemical treatment) on protein profile and yield of GVPs are in prospect. Application to forensic investigation will involve optimization of hair processing, preferably to the level of single hair shafts, and compilation of GVPs that can be identified in practice.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by National Institute of Justice grants 2011-DN-BX-K543 and 2015-DN-BX-K065 and NIH grant UL1 TR000002. We gratefully acknowledge the twins for having participated in this study and Lisa Jack of the SRI International Twin Research Registry for invaluable assistance. The mass spectrometry proteomics data have been deposited with the Center for Computational Mass Spectrometry (MassIVE ID: MSV000080295); reviewers may use the dataset password hair123.

Abbreviations

AHC	agglomerative hierarchical cluster
GVPs	genetically variant peptides
MZ	monozygotic
SD	sodium dodecanoate
SNPs	single nucleotide polymorphisms

References

- Robertson, J. Forensic Examination of Hair. Robertson, J., editor. Taylor & Francis; London: 1999. p. 79-154.
- Taupin JM. Forensic hair morphology comparison - a dying art or junk science? Science & Justice. 2004; 44:95–100. [PubMed: 15112598]
- 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. Foren Sci Int. 2005; 153:247–259.
- Pakendorf B, Stoneking M. Mitochondrial DNA and human evolution. Ann Rev Genomics Hum Genet. 2005; 6:165–183. [PubMed: 16124858]
- Mishra P, Chan DC. Mitochondrial dynamics and inheritance during cell division, development and disease. Nat Rev Molec Cell Biol. 2014; 15:634–646. [PubMed: 25237825]
- Rice RH, Bradshaw KM, Durbin-Johnson BP, Rocke DM, et al. Differentiating inbred mouse strains from each other and those with single gene mutations using hair proteomics. PLoS One. 2012; 7:e51956. [PubMed: 23251662]
- Laatsch CN, Durbin-Johnson BP, Rocke DM, Mukwana S, et al. Human hair shaft proteomic profiling: individual differences, site specificity and cuticle analysis. PeerJ. 2014; 2:e506. [PubMed: 25165623]
- Czyz W, Morahan JM, Ebers G, Ramagopalan SV. Genetic, environmental and stochastic factors in monozygotic twin discordance with a focus on epigenetic differences. BMC Med. 2012; 10:93. [PubMed: 22898292]
- Parker GJ, Leppert T, Anex DS, Hilmer JK, et al. Demonstration of protein-based human identification using the hair shaft proteome. PLoS One. 2016; 11(9):e0160653. [PubMed: 27603779]
- Krasnow RE, Jack LM, Lessov-Schlaggar CN, Bergen AW, Swan GE. The Twin Research Registry at SRI International. Twin Res Hum Genet. 2013; 16:463–470. [PubMed: 23084148]
- Lee YJ, Rice RH, Lee YM. Proteome analysis of human hair shaft: From protein identification to posttranslational modification. Molec Cell Proteom. 2006; 5:789–800.
- Rice RH. Proteomic analysis of hair shaft and nail plate. Int J Cosmetic Sci. 2011; 62:229–236.
- Rice RH, Means GE, Brown WD. Stabilization of bovine trypsin by reductive methylation. Biochim Biophys Acta. 1977; 492:316–321. [PubMed: 560214]
- Liu H, Sadygov RG, Yates JRI. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Analyt Chem. 2004; 76:4193–4201. [PubMed: 15253663]
- Dowle AA, Wilson J, Thomas JR. Comparing the diagnostic classification accuracy of iTRAQ, peak-area, spectral-counting, and emPAI methods for relative quantification in expression proteomics. J Proteome Res. 2016; 15:3550–3562. [PubMed: 27546623]
- Ramus C, Hovasse A, Marcellin M, Hesse AM, et al. Benchmarking quantitative label-free LC-MS data processing workflows using a complex spiked proteomic standard dataset. J Proteomics. 2016; 132:51–62. [PubMed: 26585461]
- Ritchie ME, Phipson B, Wu D, Hu Y, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucl Acids Res. 2015; 43doi: 10.1093/nar/gkv1007

18. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC bioinformatics*. 2013; 14(1):91. [PubMed: 23497356]
19. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol*. 2010; 11:R106. [PubMed: 20979621]
20. Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. Team, R. C. 2015. <http://CRAN.R-project.org/package=nlme>
21. R Core Team. R Foundation for Statistical Computing; Vienna, Austria: 2015. <http://www.R-project.org/>
22. Consortium GP, Auton A, Brooks LD, Durbin RM, et al. A global reference for human genetic variation. *Nature*. 2015; 526:68–74. [PubMed: 26432245]
23. Butler, JM. Fundamentals of Forensic DNA Typing. Academic Press; San Diego, CA: 2010.
24. Jovanovic M, Rooney MS, Mertins P, Przybylski D, et al. Immunogenetics. Dynamic profiling of the protein life cycle in response to pathogens. *Science*. 2015; 347:1112. (1259038).
25. Barrera LA, Vedenko A, Kurland JV, Rogers JM, et al. Survey of variation in human transcription factors reveals prevalent DNA binding changes. *Science*. 2016; 351:1450–1454. [PubMed: 27013732]
26. Kasowski M, Kyriazopoulou-Panagiotopoulou S, Grubert F, Zaugg JB, et al. Extensive variation in chromatin states across humans. *Science*. 2013; 342:750–752. [PubMed: 24136358]
27. McVicker G, van de Geijn B, Degner JF, Cain CE, et al. Identification of genetic variants that affect histone modifications in human cells. *Science*. 2013; 342:747–749. [PubMed: 24136359]
28. Waszak SM, Delaneau O, Gschwind AR, Kilpinen H, et al. Population variation and genetic control of modular chromatin architecture in humans. *Cell*. 2015; 162:1039–1050. [PubMed: 26300124]

Significance of the study

Together with our previous work, this twin study indicates that proteins of the human hair shaft permit distinguishing objectively among individual hair donors. Judging by the present evidence from monozygotic (MZ) twins, differences among individuals in expression levels of the various hair proteins appears to have primarily a genetic basis, where two unrelated individuals are much more readily differentiated than a pair of MZ twins. Moreover, proteins in the hair shaft contain genetically variant peptides traceable to non-synonymous single nucleotide polymorphisms in nuclear DNA distributed among the human population. Analysis of these peptides permitted sorting of the 20 individual subjects into 10 twin pairs. Application of such proteomic analysis has the potential to increase substantially the value of the human hair shaft as forensic evidence.

Figure 1. Detection of significant differences in protein expression in MZ twins. Among the 10 twin pairs (A1/A2 through J1/J2), 6 displayed differences in weighted spectral counts.

Figure 2. Protein differences as a function of age among the twins. The vertical axis shows the number of proteins that were significantly different in expression level for each MZ twin pair, arranged along the horizontal axis by twin age.

Figure 3.

Distances in protein expression levels between samples from single individuals, between MZ twins and between unrelated twin subjects. Box plots of Euclidean distances between samples, based on 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 (IQR) of the box edges, respectively. Observations lying more than 1.5 IQR from the box edges are indicated by circles.

Figure 4.

Agglomerative hierarchical clustering dendrogram of twin sample weighted genetically variant peptide profiles. Detected genetically variant peptides were weighted as the reciprocal of the minor allelic frequency so that co-detection of rare variants would have a greater weight than detection of variants that are more widely distributed in the population. Duplicate profiles were combined, weighted and then subject to proximity measurements in Euclidian space (see Methods) and analyzed for similarity using agglomerative hierarchical clustering. Similarity coefficients (y-axis) indicate the proximity between two samples calculated by measuring at what point they are similar.

Figure 5.

Differences in genetically variable peptide (GVP) content expressed as distances. Box plots of Euclidean distances between samples from single individuals, between MZ twins and between unrelated twin subjects, based on GVP content. 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 (IQR) of the box edges, respectively. Observations lying more than 1.5 IQR from the box edges are indicated by circles.