Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2010, Article ID 258494, 12 pages
doi:10.1155/2010/258494

Research Article

Statistical Analysis of Variation in the Human Plasma Proteome

Todd H. Corzett,[1] Imola K. Fodor,[2] Megan W. Choi,[1] Vicki L. Walsworth,[1] Kenneth W. Turteltaub,[1] Sandra L. McCutchen-Maloney,[1] and Brett A. Chromy[1]

[1] *Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, 7000 East Avenue, Livermore, CA 94550, USA*
[2] *Department of Biostatistics, Genentech, Inc., 1 DNA Way, South San Francisco, CA 94080, USA*

Correspondence should be addressed to Brett A. Chromy, chromy1@llnl.gov

Received 12 July 2009; Accepted 19 October 2009

Academic Editor: Helen J. Cooper

Quantifying the variation in the human plasma proteome is an essential prerequisite for disease-specific biomarker detection. We report here on the longitudinal and individual variation in human plasma characterized by two-dimensional difference gel electrophoresis (2-D DIGE) using plasma samples from eleven healthy subjects collected three times over a two week period. Fixed-effects modeling was used to remove dye and gel variability. Mixed-effects modeling was then used to quantitate the sources of proteomic variation. The subject-to-subject variation represented the largest variance component, while the time-within-subject variation was comparable to the experimental variation found in a previous technical variability study where one human plasma sample was processed eight times in parallel and each was then analyzed by 2-D DIGE in triplicate. Here, 21 protein spots had larger than 50% CV, suggesting that these proteins may not be appropriate as biomarkers and should be carefully scrutinized in future studies. Seventy-eight protein spots showing differential protein levels between different individuals or individual collections were identified by mass spectrometry and further characterized using hierarchical clustering. The results present a first step toward understanding the complexity of longitudinal and individual variation in the human plasma proteome, and provide a baseline for improved biomarker discovery.

1. Introduction

Mapping the human proteome presents a significant scientific challenge, partly because of the complexity of the population and partly because of technological limitations [1]. However, potential rewards in the diagnosis and treatment of diseases make proteomic characterization of human plasma a very worthwhile endeavor. The Human Proteome Organization (HUPO) represents an international consortium of academic and industrial partners whose common goal is to foster collaboration and facilitate a better understanding of the human proteome. Recognizing the need for reproducibility, and following in the footsteps of the more mature field of gene expression analysis, proteomic standards are starting to emerge [2, 3]. The Human Plasma Proteome (HPP) project [4] of HUPO, which specifically targets plasma proteins, has made considerable progress while highlighting the complexity of plasma proteomics. For example, protein identification of the same specimen resulted in less than 50% agreement when repeated multiple times [5, 6], reflecting the challenges involved in biomarker discovery from human plasma [7, 8] and underlining the need for improvements in plasma proteomic characterization. Studies providing prefractionation and other sample preparation aspects are looking to improve this process [9–12].

A primary technological problem that needs to be addressed is the quantification of the experimental variation on a given proteomic platform. Next, the baseline variation within individuals over time and the variation between multiple individuals also need to be quantitated. Searching for disease-specific biomarkers makes sense only after these two steps are addressed. Our recent study, referred to as the Technical Variation Study (TVS) [13] throughout the manuscript, addressed the first question for two-dimensional difference gel electrophoresis (2-D DIGE) experiments by processing one human plasma sample eight

times and analyzing each of the resulting eight technical replicates in triplicate on twelve gels [13–16]. The present study is a follow-up to the TVS, whereby plasma samples from eleven healthy volunteer subjects, taken at three time points separated by two weeks, were analyzed in triplicate on 50 gels.

The goal of this study was to assess longitudinal and individual variation in human plasma and to compare results to the experimental variation detected in the previously reported TVS [13]. While differences were detected in the plasma proteome within individuals over time, our analyses indicate that individual variation contributes the largest observed proteomic variability. Further, our results demonstrate that gender-related proteomic differences can be detected by 2-D DIGE and should also be considered in biomarker discovery. Overall, this work represents a first step in quantitating the variability in human plasma by addressing the individual and longitudinal proteomic variation in human plasma.

2. Materials and Methods

2.1. Sample Collection. Blood samples were collected from eleven healthy volunteers (five males, six females) at three time points separated by two weeks, with informed consent under Institutional Review Board approval from Lawrence Livermore National Laboratory. To minimize the effect of daily variations within an individual, the samples from a given subject were taken at approximately the same time, within a thirty-minute window, in the morning for each time point. Other variables were not controlled for in order to better mimic the variability in typical human plasma samples (age, fasting, illness, medication, etc.). To better examine the longitudinal variation and minimize the chance of an individual providing samples while experiencing an underlying condition such as a cold, two weeks between sample collection were chosen. Each individual was assigned an identification number to blind the samples and ease the experimental design.

2.2. Top-6 High-Abundance Protein Depletion and Sample Preparation. To increase the resolution of the 2-D DIGE technique, the six most abundant plasma proteins were depleted using affinity chromatography, as previously reported [12]. The sample cleanup and protein assay was performed as described previously [12, 13, 17, 18].

2.3. 2-D DIGE and Gel Imaging. The 33 top-6-depleted plasma samples from the 11 individuals were analyzed in triplicate in a 50-gel 2-D DIGE experiment [13–16] (see Table 1 in Supplementary Material available online at doi: 10.1155/2010/258494). Each gel contained three samples, one internal pooled standard and two experimental samples. The internal pooled standard consists of an equal amount of each of the 33 samples and was labeled with the Cy2 dye (GE Healthcare). Each experimental sample was dye-swapped and labeled with both the Cy3 and the Cy5 dyes (GE Healthcare) in the experimental design to mitigate the effect of potential dye-specific variations [19]. Samples from individuals obtained at different times were compared on some gels, while samples from two different individuals were compared on other gels. Gels were run randomly in batches of twelve in order to minimize batch-to-batch variability. Supplementary Table 1 shows the complete experimental design. Labeling first dimension (*pI*) separation, second dimension (mw) separation, and gel imaging was performed as described previously [13]. Mass spectrometry was carried out as previously described [20].

2.4. Data Analysis. The DeCyder Differential Analysis Software v5.01 (GE Healthcare) was used for quantitating differential abundance of proteins. The Differential In-gel Analysis (DIA) module was used to determine the optimal spot detection settings. Images were loaded into the Batch Processor module with the estimated number of spots set to 2,500. The master gel was assigned automatically to the gel with the most spots detected. Each sample was grouped for analysis in the Biological Variation Analysis (BVA) module. During batch processing, the Cy2 channel from each gel was used for normalization of the spot intensities and for automated matching between gels. For each spot on each gel, the software reported the standardized abundance (SA) as the ratio of the volume in the Cy3 (or Cy5) sample to the volume of the pooled standard sample labeled with Cy2, where the volumes were normalized across the gels. Standardized log abundance (SLA), defined as $\log_{10}(\mathrm{SA})$, was used in quantifying differential expression. Fold change between groups was calculated as the ratio of the average SA in the two groups. If R denotes that ratio, the fold change F was defined as $F = R$ if $R \geqq 1$ and $F = -1/R$ otherwise. A k-fold expression increase/decrease corresponded to a $+k/-k$ value of F.

Using DeCyder, all possible pairwise comparisons were made between the 33 groups defined by the eleven subjects at the three times. Within the BVA module each comparison was filtered to find the spots having (a) P-value $\leq .05$ and (b) greater than 1.5-fold change in expression between the groups. The analysis was converted into DeCyder 2D (v6.5), and the Extended Data Analysis (EDA) module (GE Healthcare) was used to perform expression pattern clustering [21–26]. Data from the TVS were integrated into the analysis, pooled standards were normalized and principal component analysis was conducted [27–30] on the spots that were successfully matched on >75% of the gels from the TVS and the present study [13].

Spot characteristics calculated by DeCyder were exported for further statistical processing into the R statistical computing environment [19] (http://www.r-project.org/). Summary statistics for the spot matching across the gels were calculated. The high-quality spots, defined as the spots matched on at least 75% of the gels, were subjected to further analysis. Spotwise standard deviations (SDs) of the SLA values and coefficients of variations (CVs) of the SA values were calculated, first by using all the data at a given spot to obtain one SD and one CV for that spot, then by performing the same calculations separately for the eleven subjects, thus

obtaining eleven SD and CV values for each spot. The former method estimated the protein expression variation among all subjects and time points, while the latter addressed the variation within individuals through time. The results were compared to the spotwise SD and CV values obtained from the TVS [13].

To quantitate the relative contribution of the components of variation, mixed-effects statistical modeling [31] was performed. Let y_{ijk} denote the SLA at spot i on gel j measured with dye k, with $i = 1,\ldots,I$, where I represents the number of spots matched on 75% of the gels, $j = 1,\ldots,50$, and $k = 1,2$. In addition, let $l = 1,\ldots,11$ and $m = 1,2,3$ indicate the subject and time indices, respectively. Let $g = 1,2$ indicate male and female genders, respectively. The assumed model was of the form

$$y_{ijklm} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \gamma_{ig} + a_{il} + b_{il(m)} + e_{ijklm}, \quad (1)$$

where μ is the overall mean, α_j denote the coefficients for the fixed gel effects, β_k the coefficients for the fixed dye effects, $(\alpha\beta)_{jk}$ the coefficients for the gel-dye interactions, γ_{ig} the coefficient for the fixed gender effect at spot i, and a_{il}, $b_{il(m)}$, and e_{ijklm} the random effects components for subject (individual), time (longitudinal), and error at spot i, independent and normally distributed [32] with mean zero and variance σ^2_{si}, σ^2_{ti}, and σ^2_{ei}, respectively. The gender, subject, and time subscripts in (1) are redundant, as for any given j and k, the identity of the sample, including the subject, gender and time, was known. However, we included them for the clarity of the model description. Since the gel and dye factors were balanced with respect to the spots (the first four terms in the model were common to all spots), (1) was fit in two stages, with results equivalent to, and computationally more efficient than, the full one-stage solution in (1) repeated at each spot. Similar methods have been established for microarrays [33]. In the first stage, the data from all I spots were used to estimate the global dye and gel effects; that is, only the first four terms in (1) plus error were included in the model. In the second stage, the last four terms in (1) were fit to the residuals from the first stage, one spot at a time. In essence, this first stage amounted to a normalization step, whereby the fixed dye and gel effects were estimated and removed by pooling the information across all the spots. In the second stage, a fixed gender effect and random variance components of subject, time, and error were estimated separately at each spot. Thus, at spot i, the total variance σ^2_{yi} was separated into its random components as $\sigma^2_{yi} = \sigma^2_{si} + \sigma^2_{ti} + \sigma^2_{ei}$.

The effect of additional statistical normalizations of the SLA on the variance component estimates and on the differentially expressed spots was investigated. The SLA values obtained from DeCyder were further normalized by statistical methods that corrected for potential dye biases within gels and range differences among the gels as previously described [34].

The spots that were determined to be of differential abundance (>1.5-fold difference with P-value <.05) were excised from the pick gel and identified by mass spectrometry as previously reported [12, 18, 20]. Identified spots were selected in DeCyder for additional expression pattern clustering.

3. Results and Discussion

3.1. Experimental Design. The experimental design is shown in Supplementary Table 1 and Supplementary Figure 1. Rather than randomly pairing the samples on the gels, the design was selected to minimize the experimental variation among the samples whose comparison was of most interest. By placing the samples from a subject across different time points on the same gel, gel-related variations for intrasubject comparisons were minimized. Essentially, our design was based on the requirements that (1) each of the 33 samples has three replicates and (2) comparisons of the samples from the same subject were of more interest than comparisons of different subjects across time points. This led to an experimental design that contained 22 gels used for comparing the same individual at different time points and 28 gels to compare two individuals at different time points. In addition, the use of dye swapping and triplicates also contributed to overall quality of the data. Our results suggest that gender variability may also be present. As individual and longitudinal variability was our main objective, we did not attempt to control for gender differences. We selected both males and females for our study to get a more appropriate human sample set. Future work looking at human proteome variability should account for gender-specific variability in the design of the experiment.

3.2. Spot Matching. Landmarks were placed manually on each gel to assist in the spot matching across the gels. Spots of interest identified through the analyses were verified to have the three-dimensional profile characteristics of a protein spot. Those spots with volumes close to background level and dust particles with very large slopes and small areas were eliminated. The total number of protein spots detected on the master gel was 2556. Three hundred and ninety seven (15%) spots were matched on at least 37 gels, and 1215 (46%) spots were matched on at least 25 gels. The following statistical analyses were restricted to the 397 high-quality spots matched on 75% of the gels. These high-quality spots were chosen to focus on spots that did not require warping or imputing of any missing data. Future data analysis may help determine if warping and data imputation can expand high-quality gel proteomic data. The latest version of DeCyder contains the ability to warp gels to potentially add missing data. These additions may provide additional spots that can be studied as high-quality, but this type of data manipulation may also create skewed expression values, as the results depend on the type of postrun imputation model that is utilized [35].

Our previous study with technical replicates of one human plasma sample [13] had 42% of the spots matched on 8 of 12 gels. The addition of biological samples from different subjects at varying time points added to the complexity of the current dataset and reduced the matching accuracy. While most of the decrease in the matching accuracy is expected to

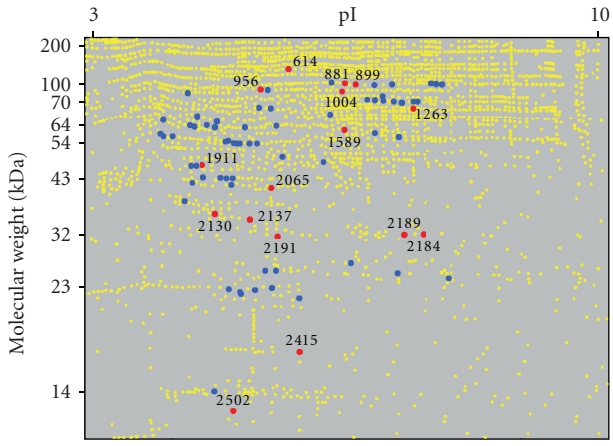

Figure 1: The spatial distribution of proteins spots (yellow dots) detected in human plasma by 2-D DIGE. Identified proteins (blue dots) and those showing differences between the theoretical and observed molecular weights (numbered red dots) are highlighted.

stem from the biological complexity of the experiment, part of it may be attributed to the larger number of gels, which inevitably increases the expected experimental variation. A study of five commercial software programs showed that an average of only 3% of the total analysis time was automated as opposed to manual, and as the number of gels increased, the percentage of automatically generated correct matches was dramatically reduced [36]. Taken together, these studies suggest that improved spot detection and matching software and algorithms are needed to increase the quality of spot matching. One such study was accomplished that created algorithms to improve spot matching with an integrated approach using hierarchical-based and optimization-based methods [37].

3.3. Differential Expression and Protein Identifications. The pairwise comparisons among the 33 samples identified over 1400 spots with P-value <.05 and fold-change > 1.5. Down selection using manual inspection eliminated most of these spots (due to three-dimensional profile characteristics not representative of protein or because of insufficient representation of the spot on enough gels) resulting in 427 spots of further interest. The majority of the spots that did not pass manual verification were similar to background levels, lacking visual characteristics of protein spots. The sensitive detection parameters used in this study, while allowing for the detection of low abundance proteins, results in increased detection of artifacts that require manual verification.

Of the 427 spots with differential abundance, those exhibiting the greatest differences in abundance levels were further characterized, and 78 proteins spots were identified by mass spectrometry. The identified proteins are listed in Table 2, along with the theoretical *pI* and molecular weight calculated from the full-length amino acid sequence of each protein. Figure 1 depicts the spatial distribution of the protein spots detected in human plasma by 2-D DIGE. Identified proteins are denoted by blue dots. Sixteen proteins (red dots) were found to have differences between the theoretical and observed molecular weights. The discrepancies are potentially due to posttranslational modifications or experimental processing; however, since all samples were treated identically, posttranslational modification is more likely. For example, spots 2189 and 2184, both identified as complement component C4A, were found to be statistically significant with at least a 1.5-fold difference between individuals. Complement component C4A has a theoretical molecular weight of 192.8 kDa; yet the protein spots identified indicate an approximate 32 kDa fragment. Since only C-terminal peptides were detected by mass spectrometry, the protein spot likely corresponds to the active Complement C4c fragment (mw = 33 kDa), which is a known cleavage product of Complement C4A [38, 39]. Variability in the amount of Complement C4c fragment between individuals could be a reflection of immune status, which may be a considerable variable when comparing human clinical subjects.

3.4. Spotwise Variation. The distribution of the SLA was consistent across the gels. The spot-wise SD values of the SLA for the 397 high-quality spots, when considering all samples, ranged from 0.04 to 0.53, with a median of 0.10. When broken down separately by subject, the range was 0.0002 to 0.50, with 0.06 as the median, reflecting the lower variation of time-within-subjects than variation between the different subjects. Both sets of values represented an increase from the spot-wise SDs observed among technical replicates of the same human plasma sample [13], where the maximum was 0.20 and the median 0.04. In the previous TVS work [13], the CV values of the SA had a median of 10% and a maximum of 42%. Here, the range of the spot-wise CVs was 10% to 93%, with a median CV of 23%. The higher CVs of the present study reflect the additional complexity due to the heterogeneity of the samples from multiple human subjects. These results are comparable to the recently reported 6% (min), 108% (max), and 19% (median) CVs found in a 2-D DIGE study of normal liver samples from ten human subjects [40]. Here, about 90% of the spots had less than 40% CV, and only 21 spots (5% of the 397 high-quality spots) had higher than 50% CV. The spots are likely not good biomarker candidates due to their high individual or longitudinal variability. These spots showing relatively high variation may correspond to a single isoform of individual proteins and do not represent all isoforms of any given protein. Notably, several of the proteins identified (Albumin, Transferrin, Haptoglobin, IgG, and IgA) are proteins removed by the Top-6 depletion process [12], which was subsequently found to result in variability when processing multiple samples in series. In future studies, column equilibration steps are recommended between samples to reduce this variability and ensure more complete depletion of high-abundant proteins.

In summary, the majority of the spots had small enough CV to indicate that the corresponding protein expressions were relatively constant across individuals, and thus could be potentially used as biomarkers. The minimum, maximum, and median CVs, when calculated separately for the subjects

Table 1: Frequency distribution of the variance component estimates.

% contribution to total variance	Subject		Time in subject		Error	
	(a)	(b)	(a)	(b)	(a)	(b)
0–10	6.31	6.31	63.89	63.89	6.57	6.57
10–20	9.34	15.66	17.68	81.57	17.17	23.74
20–30	12.89	28.50	7.83	89.39	16.67	40.40
30–40	14.14	42.68	4.29	93.69	9.34	49.75
40–50	13.63	56.31	2.27	95.96	10.86	60.61
50–60	11.36	67.68	3.03	98.99	11.87	72.47
60–70	11.36	79.04	0.51	99.49	10.37	82.83
70–80	11.61	90.66	0.50	100.00	8.08	90.91
80–90	7.83	98.48			4.80	95.71
90–100	1.51	100.00			4.29	100.00

The components of subject (σ_s), time-within-subject (σ_t), and random error (σ_e) are shown separately as (a) the percentage of spots and (b) the cumulative percentage of spots with contribution to the total variance indicated in the first column.

were 0.05%, 131%, and 14%, respectively. Over 95% of the CVs were below 35%, indicating that for most subjects, the variation over the three timepoints was comparable to the experimental variation in the previously published TVS data [13].

3.5. Statistical Normalization, Gel and Dye Effect Removal, and Variance Decomposition. The SLA values were further normalized as explained in the methods. The effect of the normalization on the results is addressed as appropriate in the following sections. The *F*-tests for the analysis of variance calculations corresponding to (1) indicated significant gel (*P*-value < 2.2e−16), dye (*P*-value < 3.2e−16), and gel-dye interaction (*P*-value < 2.2e−16) effects. The residual diagnostic plots did not reveal major departures from the assumptions, thus indicating the validity of the model. Similar analyses using the normalized SLA resulted in slightly higher *P*-values (gel effect *P*-value < 2.2e−16, dye effect *P*-value.003, gel-dye interaction *P*-value < 3.0e−09) but were consistent with the conclusions based on the calculations using the SLA.

The standard deviations corresponding to the random variance component estimates from the mixed-effects model (Figure 2) show the relative contribution of the three components at each of the spots matched on 75% of the gels. Overall, the time-within-subject component was found to have the smallest contribution to the total variance, while subject-related variation had the highest. The corresponding frequency distributions of the three variance components (Table 1) confirm that for most spots (89%) the contribution of the time component was less than 30% of the total variance. Only 5% of the spots had 70% to 80% of their variation explained by the time component, and no spot had the time component greater than 80% of the total variance. For 21% of the spots, the contribution of the subject component comprised over 70% of the total variance. For about 44% of the spots, the contribution of the subject component represented over 50% of the total variation.

To further elucidate the contributing factors involved in spot variance, we performed a meta-analysis on these data.

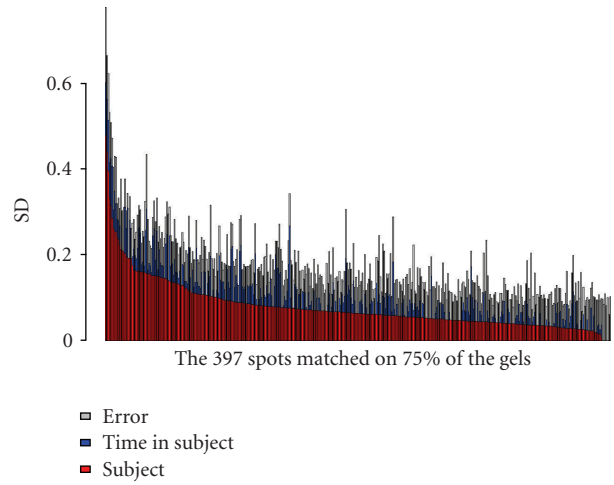

Figure 2: The subject (σ_s), time-within-subject (σ_t), and random error (σ_e) variance component estimates (on SD scale) for the 397 protein spots matched on at least 75% of the gels, ordered by the magnitude of the subject component.

Essentially, all the variances for the 397 high-quality spots were summed and the total variance that could be explained by the sum of the spot-wise subject, time within-subject, and error components was determined, respectively. A pooled estimate of the variance components was obtained by taking the average of the corresponding variance components over the spots. When aggregating the total variance over all the spots matched on at least 75% of the gels, the sum of the subject components explained 59% of the total variation and the sum of the time-within-subject components explained 12% of the total variation. The average subject variance component across the spots was 0.0097 (corresponding to σ_s = 0.098 on the Sd scale), and the average time-within-subject variance component was 0.0019 (σ_t = 0.044).

3.6. Multivariate Analysis of Expression Patterns. The EDA module of DeCyder 2D was used to visually display the

Table 2: Variable proteins identified from human plasma.

Protein Number	Protein Identity	Accession Number	mw[a]	*pI*[a]	Gene
614a	clusterin	IPI00400826	57.8	6.25	CLU
614b	alpha-2-macroglobulin precursor	IPI00478003	163.3	6	A2M
835a	alpha-2-macroglobulin precursor	IPI00478003	163.3	6	A2M
835b	Complement C3 precursor	IPI00783987	187.1	6.02	C3
849	Plasminogen	IPI00019580	90.6	7.04	PLG
856	Plasminogen	IPI00019580	90.6	7.04	PLG
881a	complement factor B preproprotein	IPI00019591	85.5	6.67	CFB
881b	complement protein C7 precursor	IPI00296608	93.5	6.09	C7
881c	Complement C3 precursor	IPI00783987	187.1	6.02	C3
884	Plasminogen	IPI00019580	90.6	7.04	PLG
893	complement factor B preproprotein	IPI00019591	85.5	6.67	CFB
899a	complement factor B preproprotein	IPI00019591	85.5	6.67	CFB
899b	complement protein C7 precursor	IPI00296608	93.5	6.09	C7
899c	Complement C3 precursor	IPI00783987	187.1	6.02	C3
910a	fibrinogen gamma	IPI00219713	49.5	5.7	FGG
910b	complement factor B preproprotein	IPI00019591	85.5	6.67	CFB
956a	complement component 1, r subcomponent	IPI00296165	80.2	5.89	C1R
956b	complement component C4A	IPI00032258	192.8	6.66	C4A
963	complement component 1, r subcomponent	IPI00296165	80.2	5.89	C1R
1002	complement component 1,s subcomponent	IPI00017696	76.7	4.86	C1S
1004a	gelsolin	IPI00646773	80.6	5.58	GSN
1004b	complement component 2	IPI00303963	83.3	7.23	C2
1004c	complement factor B preproprotein	IPI00019591	85.5	6.67	CFB
1004d	complement protein C7 precursor	IPI00296608	93.5	6.09	C7
1004e	alpha-2-macroglobulin precursor	IPI00478003	163.3	6	A2M
1004f	Complement C3 precursor	IPI00783987	187.1	6.02	C3
1004g	complement component C4A	IPI00032258	192.8	6.66	C4A
1027	transferrin	IPI00022463	77	6.81	TF
1110	IGHM protein	IPI00828205	65.3	8.1	IGHM
1113a	IGHM protein	IPI00828205	65.3	8.1	IGHM
1113b	transferrin	IPI00022463	77	6.81	TF
1128a	IGHM protein	IPI00828205	65.3	8.1	IGHM
1128b	transferrin	IPI00022463	77	6.81	TF
1129a	histidine-rich glycoprotein precursor	IPI00022371	59.6	7.09	HRG
1129b	coagulation factor XII precursor	IPI00019581	67.5	7.94	F12
1129c	transferrin	IPI00022463	77	6.81	TF
1142a	histidine-rich glycoprotein precursor	IPI00022371	59.6	7.09	HRG
1142b	transferrin	IPI00022463	77	6.81	TF
1156	transferrin	IPI00022463	77	6.81	TF
1185	transferrin	IPI00022463	77	6.81	TF
1254	hemopexin precursor	IPI00022488	51.5	6.57	HPX
1263a	histidine-rich glycoprotein precursor	IPI00022371	59.6	7.09	HRG
1263b	Complement C3 precursor	IPI00783987	187.1	6.02	C3
1276a	hemopexin precursor	IPI00022488	51.5	6.57	HPX
1276b	Heparin cofactor II precursor	IPI00292950	57.1	6.41	HCF2
1276c	peptidoglycan recognition protein L precursor	IPI00163207	62.2	7.25	PGLYRP
1382	kininogen	IPI00215894	47.9	6.29	KNG
1394	albumin	IPI00745872	69.1	5.85	ALB
1456	alpha-1-antichymotrypsin precursor	IPI00550991	45.5	5.32	AACT
1471a	immunoglobulin alpha-1 heavy chain	IPI00166866	37.6	6.06	IGHA1
1471b	kininogen	IPI00215894	47.9	6.29	KNG

Table 2: Continued.

Protein Number	Protein Identity	Accession Number	mw[a]	*pI*[a]	Gene
1471c	antithrombin III	IPI00032179	52.6	6.32	AT3
1525	Vitronectin precursor	IPI00298971	54.3	5.55	VTN
1526a	kininogen	IPI00215894	47.9	6.29	KNG
1526b	Angiotensinogen	IPI00032220	53.2	5.78	AGT
1555	Vitronectin precursor	IPI00298971	54.3	5.55	VTN
1558	immunoglobulin alpha-2 heavy chain	IPI00641229	36.4	5.71	IGH
1568a	kininogen	IPI00215894	47.9	6.29	KNG
1568b	antithrombin III	IPI00032179	52.6	6.32	AT3
1568c	Angiotensinogen	IPI00032220	53.2	5.78	AGT
1577	Alpha-2-HS-glycoprotein	IPI00022431	39.3	5.43	AHSG
1589a	immunoglobulin alpha-1 heavy chain	IPI00166866	37.6	6.06	IGHA1
1589b	apolipoprotein H precursor	IPI00298828	38.3	8.34	APOH
1589c	prepro-plasma carboxypeptidase B	IPI00329775	48.4	7.61	pCPB
1589d	fibrinogen beta chain	IPI00298497	55.9	8.54	FGB
1589e	alpha-2-macroglobulin precursor	IPI00478003	163.3	6	A2M
1616a	apolipoprotein H precursor	IPI00298828	38.3	8.34	APOH
1616b	fibrinogen beta chain	IPI00298497	55.9	8.54	FGB
1626	Alpha-2-HS-glycoprotein	IPI00022431	39.3	5.43	AHSG
1648	fibrinogen beta chain	IPI00298497	55.9	8.54	FGB
1650a	apolipoprotein D	IPI00006662	28	5.14	APOD
1650b	Alpha-2-HS-glycoprotein	IPI00022431	39.3	5.43	AHSG
1650c	alpha-1-antichymotrypsin precursor	IPI00550991	45.5	5.32	AACT
1650d	fibrinogen gamma	IPI00219713	49.5	5.7	FGG
1652	Alpha-2-HS-glycoprotein	IPI00022431	39.3	5.43	AHSG
1725	vitamin D-binding protein precursor	IPI00742696	52.9	5.32	GC
1731	vitamin D-binding protein precursor	IPI00742696	52.9	5.32	GC
1740	vitamin D-binding protein precursor	IPI00742696	52.9	5.32	GC
1741	vitamin D-binding protein precursor	IPI00742696	52.9	5.32	GC
1744	vitamin D-binding protein precursor	IPI00742696	52.9	5.32	GC
1749	vitamin D-binding protein precursor	IPI00742696	52.9	5.32	GC
1752	vitamin D-binding protein precursor	IPI00742696	52.9	5.32	GC
1843a	pigment epithelial-differentiating factor	IPI00006114	46.3	5.84	PEDF
1843b	fibrinogen gamma	IPI00219713	49.5	5.7	FGG
1898	vitamin D-binding protein precursor	IPI00742696	52.9	5.32	GC
1911a	serum paraoxonase	IPI00218732	37.8	4.96	PON
1911b	fibrinogen gamma	IPI00219713	49.5	5.7	FGG
1911c	Complement C3 precursor	IPI00783987	187.1	6.02	C3
1918	serum paraoxonase	IPI00218732	37.8	4.96	PON
1925a	serum paraoxonase	IPI00218732	37.8	4.96	PON
1925b	fibrinogen gamma	IPI00219713	49.5	5.7	FGG
1985a	serum paraoxonase	IPI00218732	37.8	4.96	PON
1985b	haptoglobin	IPI00641737	45.2	6.13	HP
1986a	apolipoprotein A-IV precursor	IPI00304273	43.4	5.22	APOA4
1986b	haptoglobin	IPI00641737	45.2	6.13	HP
1986c	serum paraoxonase	IPI00218732	37.8	4.96	PON
1998	apolipoprotein A-IV precursor	IPI00304273	43.4	5.22	APOA4
2008	apolipoprotein A-IV precursor	IPI00304273	43.4	5.22	APOA4
2029	alpha-2-glycoprotein 1	IPI00166729	34.3	5.71	AZGP1
2030a	apolipoprotein A-IV precursor	IPI00304273	43.4	5.22	APOA4
2030b	haptoglobin	IPI00641737	45.2	6.13	HP
2065a	haptoglobin	IPI00641737	45.2	6.13	HP

Table 2: Continued.

Protein Number	Protein Identity	Accession Number	mw[a]	*pI*[a]	Gene
2065b	Complement factor I	IPI00291867	65.8	7.72	CFI
2095a	alpha-1-antichymotrypsin precursor	IPI00550991	45.5	5.32	AACT
2095b	clusterin	IPI00400826	57.8	6.25	CLU
2130a	Proapolipoprotein	IPI00021841	29	5.45	APOA1
2130b	Complement factor I	IPI00291867	65.8	7.72	CFI
2137	clusterin	IPI00400826	57.8	6.25	CLU
2184	complement component C4A	IPI00032258	192.8	6.66	C4A
2189	complement component C4A	IPI00032258	192.8	6.66	C4A
2191	transthyretin	IPI00022432	15.9	5.5	TTR
2236	immunoglobulin kappa light chain	IPI00784070	26	8.16	IGKC
2259	amyloid P component	IPI00022391	25.4	6.1	APCS
2260	amyloid P component	IPI00022391	25.4	6.1	APCS
2272a	lambda-chain precursor	IPI00154742	24.7	7.54	IGL
2272b	immunoglobulin kappa light chain	IPI00784070	26	8.16	IGKC
2284	immunoglobulin kappa light chain	IPI00784070	26	8.16	IGKC
2314	Proapolipoprotein	IPI00021841	29	5.45	APOA1
2325	Proapolipoprotein	IPI00021841	29	5.45	APOA1
2326	Proapolipoprotein	IPI00021841	29	5.45	APOA1
2338	Proapolipoprotein	IPI00021841	29	5.45	APOA1
2346	plasma glutathione peroxidase	IPI00026199	16.7	8.93	GPx-P
2415	haptoglobin	IPI00641737	45.2	6.13	HP
2468	transthyretin	IPI00022432	15.9	5.5	TTR
2520	haptoglobin	IPI00641737	45.2	6.13	HP

[a]Theoretical molecular weight (mw) in kDa and isoelectric point (*pI*) values.

results of the current study and the previous TVS study [13] (Figure 3). The multivariate expression profiles of the samples across 328 spots that were matched on >75% of the spot maps from both studies were transformed into the principal component basis and the projection of the samples onto the first two principal components displayed (Figure 3). The tight scatter of the samples from the TVS (encircled in black) indicates the small magnitude of the experimental variability when analyzing technical replicates of the same human sample. The magnitude of the longitudinal variation exceeded the technical variation, as evidenced by the larger scatter of the sample points of a given subject at the different time points. For example, the two red and green ellipses (Figure 3) highlight the longitudinal variation for subjects 1 and 11, respectively. The differential scattering of the samples, from the subjects into varying regions of the principal components plot, indicates that the subject-to-subject variation exceeded the longitudinal variation within subjects.

Hierarchical clustering was used to group the 33 samples based on the similarity of their protein expression profiles along the 397 high-quality spots that were matched on >75% of the gels (Figure 4). Clustering was performed in on the proteins and experimental samples, using Euclidean distance and average linkage to define similarity. For all subjects, the first clustering step placed the three samples of the given subject into one cluster. Samples of the same subject collected at the three time points were most similar to each

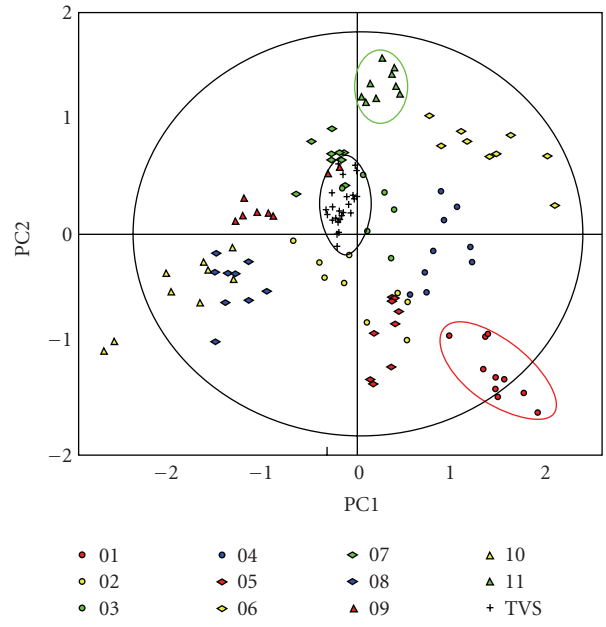

Figure 3: Principal component analysis of the 33 samples from the present study and the 8 replicates from the previous Technical Variation Study (TVS) [9], color-coded according to the legend, projected onto the first two principal components. Ellipses highlighting subjects 1 (red), 11 (green), and the TVS (black) are added for illustrative purposes only.

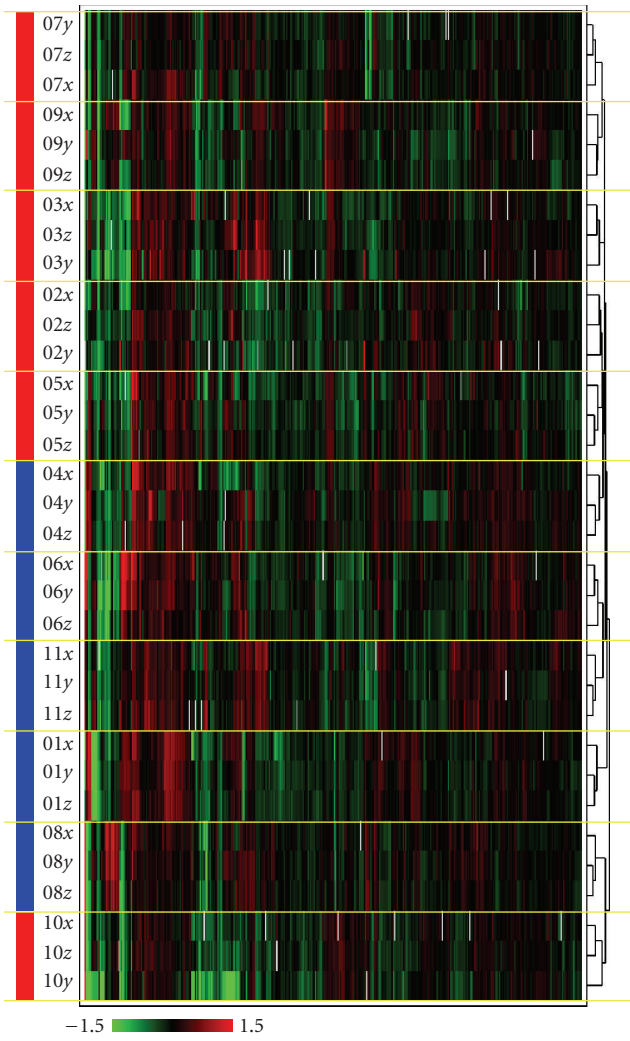

Figure 4: Hierarchical clustering of the 33 samples (y-axis) based on the abundance of the 397 high-quality protein spots on the x-axis, using Euclidean distance and average linkage. The samples are in SubjectNumberTime format, where SubjectNumber ranges from 01 to 11, and the Time values {x, y, and z } correspond to {T_1, T_2, and T_3}. The intensities range from −1.5-fold change (bright green) to 1.5-fold change (bright red). The dendrogram on the right indicates the order of the sample grouping, with more similar samples being grouped together first. The color band on the left shows the genders of the samples, with red for females, and blue for males.

other, as evidenced by the succession of self-similar bands of three rows (highlighted by the yellow lines in Figure 4). The clustering also shows a general trend of clustering the samples based on gender appeared (highlighted by the blue and red bars for males and females, resp., in Figure 4).

3.7. Gender Effects. In addition to the results seen in the hierarchical clustering (Figure 4), after fitting the mixed-effects model to the residuals from the SLA at the spots

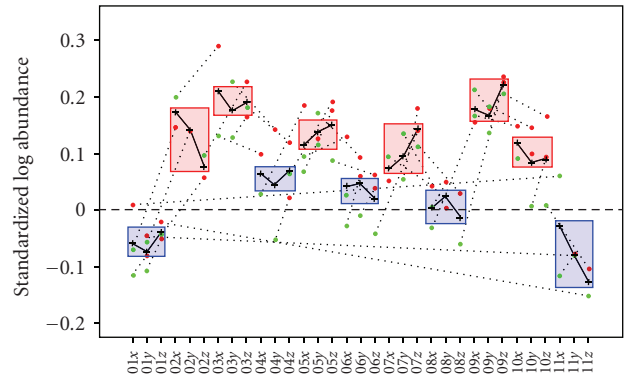

Figure 5: Expression data for alpha-2-HS-glycoprotein, with an average increase of 1.49-fold between the female and male groups, and FDR-adjusted gender-effect P-value =.055. The samples are in SubjectNumberTime format, where SubjectNumber ranges from 01 to 11, and the Time values {x, y, and z } correspond to {T_1, T_2, and T_3}. The annotations indicate the gels (numbers) and the dyes (red for Cy5, green for Cy3) corresponding to the samples. Dotted lines connect samples multiplexed on the same gel. Crosses indicate sample averages over the technical replicates. The solid line connects all sample averages. Boxes around the three Time values for each SubjectNumber highlight male and female genders (blue and red respectively) added for illustrative purposes only.

matched on 75% of the gels, 17 spots showed gender-effect P-value <.01. None of these spots were found to be significant (P-value <.01) after the False Discovery Rate (FDR) method [41] for multiple comparisons was applied suggesting that larger numbers of samples are needed to validate gender differences in the human plasma proteome. Despite the lack of statistically significant data on gender differences, trends in this dataset suggest that future, larger datasets might enable the differentiation of protein expression levels due to gender. One spot, 1659, had FDR-adjusted gender-effect P-value equal to.055 with a 1.49-fold-change between the male and female groups (Figure 5). Five additional spots (466, 1626 alpha-2-HS-glycoprotein, 1650 alpha-2-HS-glycoprotein, 1652 alpha-2-HS-glycoprotein, 1678) had adjusted P-values of.11. The results were similar when fitting the same model to the residuals from the statistically normalized SLA, albeit with P-values that slightly exceeded their corresponding values based on the SLA. Three of the spots exhibiting gender effects were identified as alpha-2-HS-glycoprotein, which has been shown to vary between males and females. The concentration of alpha-2-HS-glycoprotein has been found to undergo a progressive age-related decrease in women, while men show no noticeable change [36].

The removal of the dye and gel effects using the model in (1) proved to be a beneficial preprocessing step. Without this step, when the mixed-effect model was fit to the original SLA, the smallest FDR-adjusted gender-effect P-value was.19 (spot 1659). When the same model was fit to the statistically normalized SLA, the smallest FDR-adjusted P-value was also.19 (spot 1659). The statistical normalization improved the quality of the data slightly, but it did not reduce dramatically the observed P-values. On the other

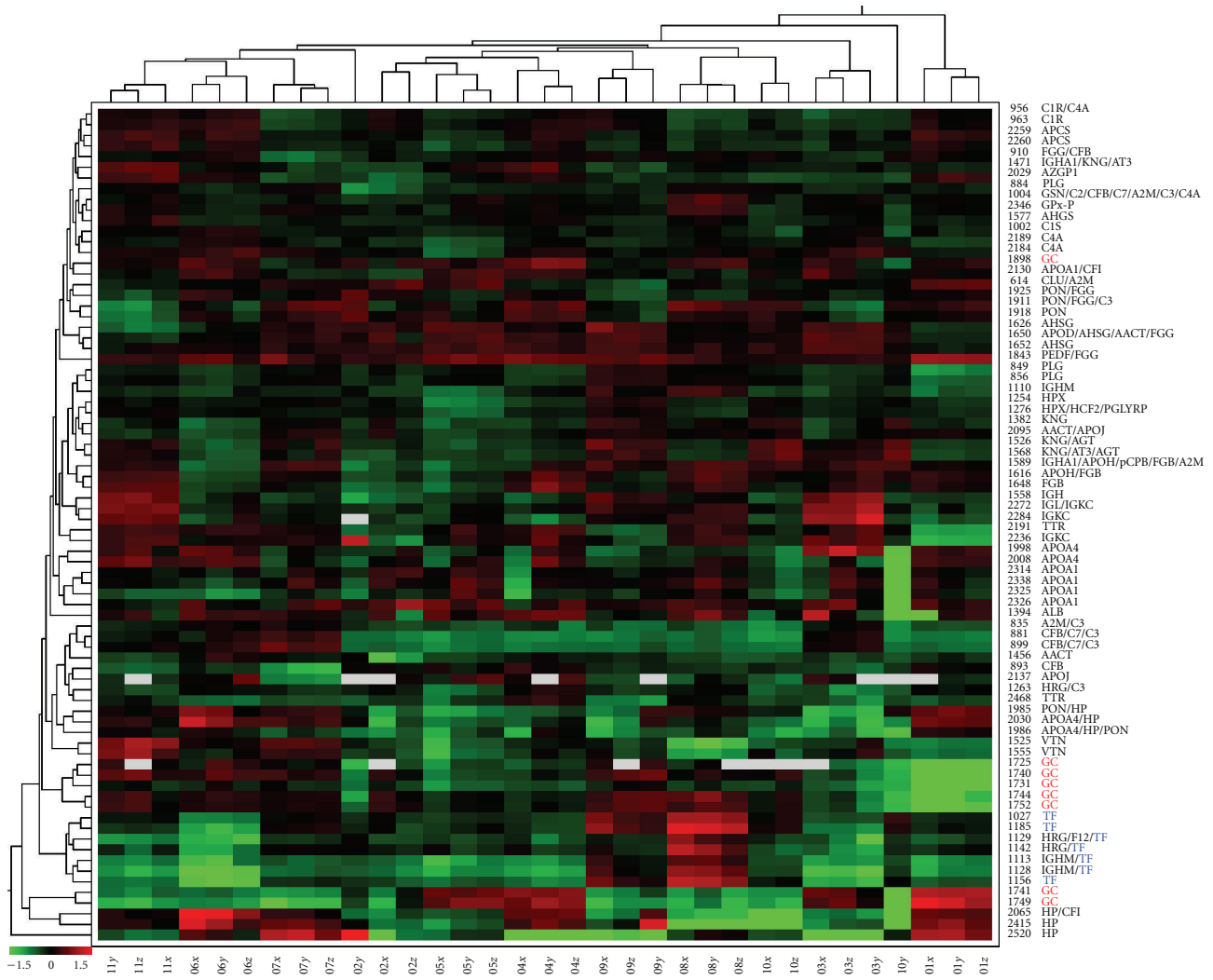

FIGURE 6: Hierarchical clustering of the 33 samples (x-axis) based on the abundance of the 78-identified protein spots on the y-axis, using Euclidean distance metrics and average linkage methods. The samples are in SubjectNumberTime format, where SubjectNumber ranges from 01 to 11, and the Time values (x, y, and z) correspond to (T_1, T_2, and T_3). The intensities range from −1.5 (bright green) to 1.5 (bright red). The dendrogram on the top indicates the order of the sample grouping, with samples corresponding to the lower leaves being grouped together first. Similarly, the dendrogram on the left indicates the ordering of the protein spots. All transferrin (TF) and vitamin D-binding protein (GC) identifications are highlighted in blue and red, respectively.

hand, pooling the information across the gels to remove the common dye and gel effects strengthened the signal and reduced markedly the FDR-adjusted P-values. As explained in the previous paragraph, for spot 1659, the new P-value was close to.05.

3.8. Multivariate Analysis of Identified Proteins. Hierarchical clustering of the identified proteins (Figure 6) was conducted using the Euclidean distance metric and average linkage methods. For all subjects, other than subject 10, the first clustering step placed the three samples of that subject into one cluster. Subject 10 had two time points grouped together (x and z) with the third point (y) separated by Subject 3. Because the 78 identified proteins were the most differential between time and subjects, it is not unexpected to see clustering results that may not perfectly align all subjects or time points. For example, all protein spots identified as transferrin (TF) clustered together due to their similar expression patterns, while the vitamin D-binding protein (GC) spots were found in multiple clusters due to differences in expression patterns between the individuals. Multiple proteins may cluster together due to coregulation and similar functions, and in the case of APOA4 and APOA1 (Figure 6), coregulation has been reported [37, 42].

4. Conclusion

Statistical analysis of a 2-D DIGE experiment involving triplicate plasma samples from eleven human subjects taken

at three time points separated by several weeks demonstrated that the subject-to-subject variation exceeded the time-within-subject variation. The variation in the human plasma proteome reported here was greater than a previous technical variation study wherein one plasma sample was processed multiple times [13].

Here, for 70% of the high-quality protein spots, the coefficient of variation of the SLA was less than 30% across all subjects and time points, thus indicating that the baseline expression levels of those proteins are relatively stable in the population represented by the subjects in this study. Only 21 spots had larger than 50% CV, suggesting that these protein isoforms should be avoided as biomarker candidates. Many of these protein spots represent medium to high abundance plasma proteins. Since they are higher abundance, they might bias LC/MS datasets, but since the total number of these spots relative to the total plasma proteome is small, their total influence on a sample is likely also small. In addition, protein spots with gender-related differences should be considered separately for males and females. However, more thorough studies, including the use of a larger population set with additional time points over longer periods of time, are recommended to more fully address individual, longitudinal, and gender variability as related to biomarker discovery.

We noted that preprocessing the data by first removing the fixed effects of the gels and dyes was important in data analysis and improved the quality of the data. This step resulted in six protein spots showing a statistically significant gender effect at an FDR-adjusted 11% significance level. Without the preprocessing step, the smallest gender effect *P*-value was.19. While removing the gel and dye effects lead to stronger conclusions, the additional statistical normalization of the SLA had only marginal effects and did not alter the conclusions.

Spot matching confounds gel- and software-related protein differences with real biological effects. In the present study, we only considered spots that were matched on at least 75% of the gels. Spots with lower matching quality can be investigated separately, as they may correspond to proteins that are absent or have very low expression in certain individuals, but which may have biological significance. We envision that such studies will become more relevant as the field of personalized medicine matures, and as detection and matching algorithms continue to improve.

This study represents a first step toward quantitating the longitudinal and individual variation in the human plasma proteome, as measured on the 2-D DIGE platform. Interestingly, gender-related variations were also detected suggesting that gender variability should also be considered in biomarker discovery. Future, larger-scale experiments that include more subjects representative of various population segments, encompassing differences in ethnicity, age, gender, disease status, and other relevant factors, have the potential to define baseline proteomic similarities and differences in the human population, which will in turn facilitate improved biomarker discovery.

Acknowledgments

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Security, LLC under Contract DE-AC52-07NA27344, with support from the Department of Homeland Security and LLNL Laboratory Directed Research and Development funding UCRL-JRNL-229654. Todd H. Corzett and Imola K. Fodor contributed equally to this work.

References

[1] D. Nedelkov, "Population proteomics: addressing protein diversity in humans," *Expert Review of Proteomics*, vol. 2, no. 3, pp. 315–324, 2005.

[2] H. Hermjakob, "The HUPO proteomics standards initiative—overcoming the fragmentation of proteomics data," *Proteomics*, vol. 1, no. S2, pp. 34–38, 2006.

[3] C. F. Taylor, "Minimum reporting requirements for proteomics: a MIAPE primer," *Proteomics*, vol. 1, no. S2, pp. 39–44, 2006.

[4] G. S. Omenn, "Exploring the human plasma proteome: editorial," *Proteomics*, vol. 5, no. 13, pp. 3223–3225, 2005.

[5] G. S. Omenn, Y.-K. Paik, and D. Speicher, "The HUPO plasma proteome project: a report from the Munich congress," *Proteomics*, vol. 6, no. 1, pp. 9–11, 2006.

[6] G. S. Omenn, D. J. States, M. Adamski, et al., "Overview of the HUPO plasma proteome project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database," *Proteomics*, vol. 5, no. 13, pp. 3226–3245, 2005.

[7] K. Cottingham, "HUPO plasma proteome project: challenges and future directions," *Journal of Proteome Research*, vol. 5, no. 6, p. 1298, 2006.

[8] J. M. Jacobs, J. N. Adkins, W.-J. Qian, et al., "Utilizing human blood plasma for proteomic biomarker discovery," *Journal of Proteome Research*, vol. 4, no. 4, pp. 1073–1085, 2005.

[9] S. M. Hanash, S. J. Pitteri, and V. M. Faca, "Mining the plasma proteome for cancer biomarkers," *Nature*, vol. 452, no. 7187, pp. 571–579, 2008.

[10] M. J. Han and D. W. Speicher, "Microscale isoelectric focusing in solution: a method for comprehensive and quantitative proteome analysis using 1-D and 2-D DIGE combined with MicroSol IEF prefractionation," *Methods in Molecular Biology*, vol. 424, pp. 241–256, 2008.

[11] S. Y. Cho, E.-Y. Lee, H.-Y. Kim, et al., "Protein profiling of human plasma samples by two-dimensional electrophoresis," *Methods in Molecular Biology*, vol. 428, pp. 57–75, 2007.

[12] B. A. Chromy, A. D. Gonzales, J. Perkins, et al., "Proteomic analysis of human serum by two-dimensional differential gel electrophoresis after depletion of high-abundant proteins," *Journal of Proteome Research*, vol. 3, no. 6, pp. 1120–1127, 2004.

[13] T. H. Corzett, I. K. Fodor, M. W. Choi, et al., "Statistical analysis of the experimental variation in the proteomic characterization of human plasma by two-dimensional difference gel electrophoresis," *Journal of Proteome Research*, vol. 5, no. 10, pp. 2611–2619, 2006.

[14] A. Alban, S. O. David, L. Bjorkesten, et al., "A novel experimental design for comparative two-dimensional gel analysis: two-dimensional difference gel electrophoresis incorporating

a pooled internal standard," *Proteomics*, vol. 3, no. 1, pp. 36–44, 2003.

[15] K. S. Lilley and D. B. Friedman, "All about DIGE: quantification technology for differential-display 2D-gel proteomics," *Expert Review of Proteomics*, vol. 1, no. 4, pp. 401–409, 2004.

[16] R. Marouga, S. David, and E. Hawkins, "The development of the DIGE system: 2D fluorescence difference gel analysis technology," *Analytical and Bioanalytical Chemistry*, vol. 382, no. 3, pp. 669–678, 2005.

[17] L. A. Echan, H.-Y. Tang, N. Ali-Khan, K. Lee, and D. W. Speicher, "Depletion of multiple high-abundance proteins improves protein profiling capacities of human serum and plasma," *Proteomics*, vol. 5, no. 13, pp. 3292–3303, 2005.

[18] R. C. Mahnke, T. H. Corzett, S. L. McCutchen-Maloney, and B. A. Chromy, "An integrated proteomic workflow for two-dimensional differential gel electrophoresis and robotic spot picking," *Journal of Proteome Research*, vol. 5, no. 9, pp. 2093–2097, 2006.

[19] W. N. Venables and D. M. Smith, *An Introduction to R: Notes on R, A Programming Environment for Data Analysis and Graphics, v.2.0.1*, Network Theory, Bristol, UK, 2004.

[20] B. A. Chromy, J. Perkins, J. L. Heidbrink, et al., "Proteomic characterization of host response to Yersinia pestis and near neighbors," *Biochemical and Biophysical Research Communications*, vol. 320, no. 2, pp. 474–479, 2004.

[21] N. Bolshakova and F. Azuaje, "Cluster validation techniques for genome expression data," *Signal Processing*, vol. 83, no. 4, pp. 825–833, 2003.

[22] J. C. Dunn, "Well separated clusters and optimal fuzzy partitions," *The Journal on Systemics, Cybernetics and Informatics*, vol. 4, pp. 95–104, 1974.

[23] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 25, pp. 14863–14868, 1998.

[24] L. Kaufmann and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons, New York, NY, USA, 1990.

[25] R. R. Sokal and C. D. Michener, "A statistical method for evaluating systematic relationships," *University of Kansas Science Bulletin*, vol. 38, no. 6, pp. 1409–1438, 1958.

[26] R. Tibshirani, G. Walther, and T. Hastie, *Estimating the Number of Clusters in a Dataset via the Gap Statistic*, Department of Biostatistics, Stanford University, Stanford, Calif, USA, 2000.

[27] L. Eriksson, N. Kettaneh-Wold, J. Trygg, and S. Wold, *Multi- and Megavariate Data Analysis*, vol. 533, Umetrics Academy, Umeå, Sweden, 2001.

[28] I. T. Jolliffe, *Principal Component Analysis*, Springer, Berlin, Germany, 2002.

[29] "Umetrics SIMCA,10.5," Sweden.

[30] H. Wold, "Estimation of principal components and related models by iterative least squares," in *Multivariate Analysis*, P. Krishnaiah, Ed., pp. 391–420, Academic Press, New York, NY, USA, 1966.

[31] J. C. Pinheiro and D. M. Bates, *Mixed-Effects Models in S and S-PLUS*, Statistics and Computing, Springer, New York, NY, USA, 2000.

[32] R. O. Kuehl, *Design of Experiments: Statistical Principles of Research Design and Analysis*, Duxbury Press, Belmont, Calif, USA, 2nd edition, 2000.

[33] X. Cui and G. A. Churchill, "Statistical tests for differential expression in cDNA microarray experiments," *Genome Biology*, vol. 4, no. 4, article 210, 2003.

[34] I. K. Fodor, D. O. Nelson, M. Alegria-Hartman, et al., "Statistical challenges in the analysis of two-dimensional difference gel electrophoresis experiments using DeCyder™," *Bioinformatics*, vol. 21, no. 19, pp. 3733–3740, 2005.

[35] R. Pedreschi, M. L. A. T. M. Hertog, S. C. Carpentier, et al., "Treatment of missing values for multivariate statistical analysis of gel-based proteomics data," *Proteomics*, vol. 8, no. 7, pp. 1371–1383, 2008.

[36] I. R. Dickson, M. Bagga, and C. R. Paterson, "Variations in the serum concentration and urine excretion of α2HS-glycoprotein, a bone-related protein, in normal individuals and in patients with osteogenesis imperfecta," *Calcified Tissue International*, vol. 35, no. 1, pp. 16–20, 1983.

[37] O. Schamaun, B. Olaisen, B. Mevag, et al., "The two apolipoprotein loci apoA-I and apoA-IV are closely linked in man," *Human Genetics*, vol. 68, no. 2, pp. 181–184, 1984.

[38] I. Gigli, I. von Zabern, and R. R. Porter, "The isolation and structure of C4, the fourth component of human complement," *Biochemical Journal*, vol. 165, no. 3, pp. 439–446, 1977.

[39] E. M. Press and J. Gagnon, "Human complement component C4. Structural studies on the fragments derived from C4b by cleavage with C3b inactivator," *Biochemical Journal*, vol. 199, no. 2, pp. 351–357, 1981.

[40] X. Zhang, Y. Guo, Y. Song, et al., "Proteomic analysis of individual variation in normal livers of human beings using difference gel electrophoresis," *Proteomics*, vol. 6, no. 19, pp. 5260–5268, 2006.

[41] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society Series B*, vol. 57, no. 1, pp. 289–300, 1995.

[42] L. Vergnes, T. Taniguchi, K. Omori, M. M. Zakin, and A. Ochoa, "The apolipoprotein A-I/C-III/A-IV gene cluster: ApoC-III and ApoA-IV expression is regulated by two common enhancers," *Biochimica et Biophysica Acta*, vol. 1348, no. 3, pp. 299–310, 1997.