REGULAR ARTICLE

A wide range of protein isoforms in serum and plasma uncovered by a quantitative intact protein analysis system

David E. Misek[1], Rork Kuick[1], Hong Wang[1, 5], Vladimir Galchev[1], Bin Deng[1, 5], Rong Zhao[1], John Tra[1], Michael R. Pisano[3], Ravi Amunugama[3], David Allen[3], Angela K. Walker[4], John R. Strahler[4], Philip Andrews[4], Gilbert S. Omenn[2] and Samir M. Hanash[1, 5]*

[1] Department of Pediatrics, University of Michigan Medical Center, Ann Arbor, MI, USA
[2] Department of Internal Medicine, University of Michigan Medical Center, Ann Arbor, MI, USA
[3] PRS Proteomics, Ann Arbor, MI, USA
[4] Department of Biological Chemistry, University of Michigan Medical Center, Ann Arbor, MI, USA
[5] Fred Hutchinson Cancer Research Center, Seattle, WA, USA

Received: February 22, 2005
Revised: April 8, 2005
Accepted: April 16, 2005

We have implemented an orthogonal 3-D intact protein analysis system (IPAS) to quantitatively profile protein differences between human serum and plasma. Reference specimens consisting of pooled Caucasian-American serum, citrate-anticoagulated plasma, and EDTA-anticoagulated plasma were each depleted of six highly abundant proteins, concentrated, and labeled with a different Cy dye (Cy5, Cy3, or Cy2). A mixture consisting of each of the labeled samples was subjected to three dimensions of separation based on charge, hydrophobicity, and molecular mass. Differences in the abundance of proteins between each of the three samples were determined. More than 5000 bands were found to have greater than two-fold difference in intensity between any pair of labeled specimens by quantitative imaging. As expected, some of the differences in band intensities between serum and plasma were attributable to proteins related to coagulation. Interestingly, many proteins were identified in multiple fractions, each exhibiting different p*I*, hydrophobicity, or molecular mass. This is likely reflective of the expression of different protein isoforms or specific protein cleavage products, as illustrated by complement component 3 precursor and clusterin. IPAS provides a high resolution, high sensitivity, and quantitative approach for the analysis of serum and plasma proteins, and allows assessment of PTMs as a potential source of biomarkers.

Keywords:
Biomarkers / Plasma / Protein isoforms / Proteome profiling / Serum

Correspondence: Professor Samir M. Hanash, M.D., Ph.D., Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North Mailstop M5-C800, P.O. Box 19024, Seattle, WA 98109-1024, USA
E-mail: shanash@fhcrc.org
Fax: +1-206-667-2537

Abbreviations: CLU, clusterin; **GVHD**, graft-*versus*-host disease; **IPAS**, intact protein analysis system

1 Introduction

Comprehensive profiling of human serum and plasma proteins has substantial relevance to the identification of circulating disease biomarkers [1]. However, circulating

* Additional corresponding author: Professor Gilbert S. Omenn, M.D., Ph.D.; **E-mail:** gomenn@umich.edu

proteins have a very wide range of protein concentrations and many occur in multiple isoforms that need to be separately quantified. Several approaches for comprehensive protein profiling have been applied to serum and plasma proteins [2]. Intrinsic limitations in resolving power and interference from albumin and several other high-abundance proteins, which together comprise approximately 90% of the protein mass of plasma, reduce the range of proteins that can be assessed by 2-D PAGE [3, 4]. Analysis of peptide subsets from complex digests as accomplished using MuDPIT [5] may not be sensitive to protein PTMs. We have combined immunodepletion of six highly abundant proteins [6], protein tagging, and an orthogonal, 3-D intact protein analysis system (IPAS) to profile serum and plasma [1, 7]. This quantitative analysis system resolves complex mixtures of Cy-dye-labeled proteins based on charge, hydrophobicity, and molecular mass. Proteins of interest are subjected to MS for their identification. In this study, we utilized three HUPO Plasma Proteome Project (PPP) reference specimens to evaluate the range of protein differences between serum and plasma and the sensitivity of the analytical approach to detect and identify protein isoforms.

2 Materials and methods

2.1 Samples

The HUPO PPP reference specimens were prepared by BD Biosciences without clot activator or protease inhibitors, as previously described [8]. We utilized 1.0 mL each of the BD-B1 Caucasian-American serum, EDTA-anticoagulated plasma, and citrate-anticoagulated plasma in this study.

2.2 Prefractionation immunodepletion chromatography

For each of the three specimens, the six most abundant serum/plasma proteins (albumin, IgG, IgA, transferrin, haptoglobin, and α-1-antitrypsin) were removed by immunodepletion chromatography (multiple affinity removal column HU-6, 4.6 mm id × 100 mm; Agilent, Wilmington, DE). Two columns were used in tandem to increase the loading capacity to 75 μL of plasma (~4.6 mg plasma protein) *per* run. A dual-pump HPLC system was installed in-house with UV detection at 280 nm. The columns were equilibrated in buffer A (Agilent) for 10 min at a flow rate of 0.5 mL/min, after which 75 μL of sample (*per* column run) diluted five-fold in buffer A was injected. Flow-through fractions were collected and stored at −80°C until use. Bound proteins were eluted from the column with buffer B (Agilent). Fourteen runs were required to process 1.0 mL, achieving a recovery of approximately 7 mg protein (11% of original load) in the flow-through fractions from each specimen.

2.3 Cy-dye labeling of the flow-through protein fractions

The flow-through immunodepleted fractions were concentrated in Centricon YM3 (3 kDa M_r cutoff) concentrators according to manufacturer (Millipore, Bedford, MA), protein-quantitated, volume-adjusted to 1.5 mL in lysis buffer (30 mM Tris (pH 8.5), 8 M urea, 4% CHAPS), and then placed on ice in a 15 mL centrifuge tube. Cy5, Cy3, and Cy2 minimal labeling (labeling on the ε-amino group of lysines) dyes were used to label the serum, EDTA-plasma, and citrate-plasma, respectively. Each Cy dye was resuspended in DMF at 1 mM and reacted at 4 nmol Cy dye/mg protein for 30 min in the dark on ice with an estimated 3% of the lysyl ε-amino groups, and then quenched with 50-fold molar excess of L-lysine (relative to dye) for 10 min on ice.

2.4 Protein separations

A mixture of the three labeled samples was fractionated according to p*I* by liquid-based IEF for 6 h at 10°C using a 50 mL Rotofor (Bio-Rad, Hercules, CA) and pH 4–8 ampholines (Resolyte, BDH). Each of 20 resulting fractions was individually injected onto a MacroTrap column (3.0 mm id × 8 mm with polymeric RP-HPLC column packing; Michrom Bioresources, Auburn, CA, USA). The sample was washed with water and buffer A, then resolved on a POROS R2/10 RP-HPLC column (4.6 mm id × 100 mm, with 10 μm crosslinked polystyrene-divinylbenzene; Applied Biosystems, Foster City, CA, USA) during a 60 min gradient elution run at 1.5 mL/min. The mobile phases used were A: 98% H_2O + 2% ACN + 0.1% TFA and B: 10% H_2O + 90% ACN + 0.1% TFA. A total of 52 second-dimension fractions (designated 13–39 and 40–66) were collected for each run. After boiling for 5 min, RP-HPLC fractions were dissolved in 20 μL sample buffer (120 mM Tris-HCl (pH 6.6), 10% SDS, 20% glycerol, 3% DTT, 0.03 M bromophenol blue), and loaded onto 18 × 16 cm (0.75 mm spacers) 7.5–15% gradient polyacrylamide gels. The designation of each of the third-dimension gels consisted of their first-dimension fraction number followed by either A or B (*e.g.*, gel 9A held second-dimension fractions 13–39 of first-dimension fraction 9).

2.5 Image analysis and spot-picking

Following fixation, three images of each gel were obtained using a 9400 Typhoon Scanner (Amersham Biosciences, Piscataway, NJ, USA), each corresponding to one of the three Cy dyes used to label the samples. Images were quantitatively analyzed with the DeCyder Software Package (version 5.0, Amersham Biosciences). This software was designed for 2-D gel analysis, thus bands on the images were typically detected as two spots. For each spot, a boundary was determined and a background-corrected, integrated intensity was computed for each of the three images. Intensities from Cy3

and Cy5 images were normalized to the Cy2 image by dividing by the median of the ratios of spot-integrated intensities for spots larger than 1000 U in both images. Spots with normalized ratios between two images for the same gel that were larger than 2.0 (or smaller than 0.5) were preferentially selected for spot-picking using landmark disks on the left and right of each gel and an Ettan Spot Picker (Amersham Biosciences). The spot-picking robot excised 2.0 mm circular regions at the designated locations of the gel into individual wells of 96-well plates. Molecular weight (MW) estimates for each band were calculated by fitting a cubic polynomial to the log of MW *versus* relative mobility for a set of nine MW standards on gel 5A. The relative mobilities on each gel were calculated as the fraction of the distance between the bottom of the stacking gel and the leading edge of labeled proteins.

2.6 MS

2.6.1 LC-MS/MS

Two sets of samples were subjected to LC-MS/MS analysis (Table 1). For one set (from fractions 5A and 15A), excised gel pieces were washed successively in H_2O, NH_4HCO_3, and ACN, vacuum-dried, and incubated in a 10 μL trypsin solution (6.3 ng/μL in 25 mM NH_4HCO_3, pH 8.0) for 16 h at 37°C. The peptide mixtures were analyzed using nanoflow capillary HPLC (C18, 7 cm × 75 μm id, at flow rate 200–300 nL/min with a 1 h gradient elution) by ESI-MS/MS on a Q-TOF-micro (Micromass, Manchester, UK). MS/MS spectra were acquired in the automated MS to MS/MS switching mode, with an *m/z*-dependent set of collision offset values. Doubly- and triply-charged ions were selected and fragmented with argon as the collision gas. The acquired spectra were automatically processed and searched against the nonredundant Swiss-Prot protein sequence database using ProteinLynx Global Server 1.1 (Micromass) or against NCBI-nr using MASCOT (www.matrixscience.com).

Another set of samples (from fractions 10A and 14A) was subjected to proteolytic digestion on a ProGest workstation in NH_4HCO_3, reduced with DTT, alkylated with iodoacetamide, and incubated with trypsin at 37°C for 4 h. Peptide hydrolysate (15 μL) was processed on a 15 cm × 75 μm C18 column at a flow-rate of 200 nL/min for analysis by nanoLC-MS/MS on a Micromass Q-TOF 2. MS/MS data were searched with MASCOT for tryptic peptides, permitting oxidation (M), acetyl *N*-terminal, *N*-terminal Q, and *N*-terminal E with unrestricted protein mass, ±100 ppm peptide mass tolerance, ±0.1 Da fragment mass tolerance, and a maximum of one missed cleavage.

Table 1. Summary of digests submitted for identification by three MS instruments

MS method	Number of digests	Number of protein hits	Number of distinct genes	Gels
TOF/TOF	482	505	52	1–3, 6–9, 11, 13, 15–20A, 7–13, 20B
LC-MS/MS#1	81	93	44	5A, 15A
LC-MS/MS#2	92	191	40	10A, 14A
Total	655	789	87 (union)	

2.6.2 TOF/TOF

The TOF/TOF was used to analyze plugs from 16A gel and 8B gel patterns (Table 1). Gel pieces were washed twice in 100 μL water, then in 100 mM NH_4HCO_3, and finally in 50 μL ACN for 15 min. The samples were dried under vacuum for 5 min, then digested with 10 μL solution of 6 mg/mL urea, 100 mM NH_4HCO_3, 10% ACN, and 6 ng/μL modified bovine trypsin (Roche Applied Science, Indianapolis, IN, USA) for 4 h at 37°C or overnight at room temperature. Extracted peptides were cleaned up using C18 micro-bed ZipTip (Millipore). An Applied Biosystems 4700 Proteomics Analyzer (MALDI-TOF/TOF) was used to obtain MS fingerprint spectra using 1500–2000 laser shots; five to eight MS peaks were automatically subjected to MS/MS, typically using 3500–4000 laser shots. Spectra were processed using the 4700 Console instrument control software (Applied Biosystems, A); separate searches for MS peak lists and sets of MS/MS peak lists were performed using GPS explorer software (Applied Biosystems) (MASCOT search engines). Searches were conducted against the entire NCBInr protein database with 25 ppm tolerances for MS searches and a maximum of 65 peaks from the 1000 to 4000 Da mass range. MS/MS searches permitted 50 ppm errors on the parent ion mass and 0.3 or 0.6 Da errors on the parent ions, using a maximum of 100 MS/MS peaks *per* spectrum. In cases of failure to obtain internal calibration of MS peak masses, parent ion errors of 0.6 Da were permitted in MS/MS searches. The GPS explorer software delivers a confidence based on the MASCOT score similar to the threshold computed by MASCOT (for 95% confidence). Protein identifications with confidence levels exceeding 98% were accepted after excluding nonhuman proteins or human keratins. In 8% of cases, identifications with lower confidence (95–98%) were accepted when (1) the same protein obtained good scores in both MS and MS/MS runs or in two MS/MS runs for the same sample, (2) a series of samples from neighboring bands on the gels gave good scores for the same protein, or (3) an identification from MS/MS spectra had the most intense peaks accounted for in the candidate sequence (typically those requiring fragmentation *N*-terminal to proline and/or *C*-terminal to aspartic acid).

Spectra, peak lists, and details of search algorithm output are available from the HUPO PPP at http://www.bioinformatics.med.umich.edu/proteome/hupo/ppp and the PPP core data sets will be available in the PRIDE database at http://www.ebi.ac.uk/pride.

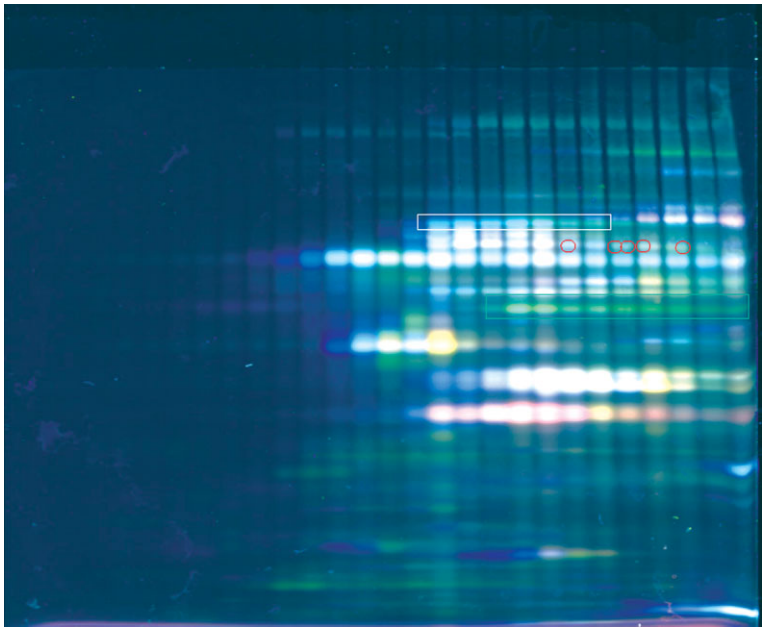

Figure 1. Quantitative differences between serum, EDTA, and citrate-plasma. Shown is gel 13A loaded with second-dimension fractions 13–39 from first-dimension fraction 13. Cy5 dye (serum) is shown as red, Cy3 (EDTA sample) as green, and Cy2 (citrate sample) as blue. Locations of gel plugs giving MS/MS identifications for complement C3 in gel 13A are shown as red circles. Bands in the green rectangle correspond to bands from first-dimension fraction 15 identified as fibrinogen gamma, where they were also most intense in the Cy3 channel. Blue bands in the white rectangle were identified as prothrombin (F2, coagulation factor 2), and were less than half as intense in the Cy5 channel (serum) for these bands.

3 Results

The combination of immunoaffinity depletion of six abundant proteins, three-dye labeling of the serum, EDTA-anticoagulated plasma, and citrate-anticoagulated plasma from a single donor pool, and 3-D IPAS fractionation gave a rich pattern of proteins in the third-dimension SDS-PAGE gels (Fig. 1). Protein bands were more numerous in some fractions than in others, and generally were more numerous in the second-dimension fractions 13–39 (run on "A" gels) than fractions 40–66 ("B" gels). More than 5000 bands in total exhibited greater than two-fold difference in intensity between any pair of labeled specimens by quantitative imaging with DeCyder software. The identity of proteins in 173 gel plugs analyzed by LC-MS/MS was determined with confidence (with a total of 284 "hits"); there were 505 hits with 482 different gel plugs analyzed by MALDI-TOF/TOF (Table 1). The corresponding bands varied in intensity by more than 1000-fold, although success at obtaining confident identifications among the gel plugs analyzed increased with increased total intensity of the band (sum of intensities for all three dyes). For example, for the faintest bands attempted, a total of seven protein identifications were obtained for the 25 digests analyzed by LC-MS/MS, whereas none of the 92 digests with a similar intensity level was identified by TOF/TOF analysis. For some digests, multiple protein identifications were obtained within a single digest. In our experience, the Cy-dye-labeling technique allows the visualization of proteins that are below the detection level of SYPRORuby or silver staining.

HGNC gene symbols (http://www.gene.ucl.ac.uk/nomenclature) for each protein identification were utilized as a means of simplifying the data, since search engines often produce lists of essentially equivalent hits to sets of protein sequences from the same gene. The results are summarized as a Venn diagram in Fig. 2. There were at least two cases in which more than one gene could explain the matching peptides obtained from MS search engines (HSPA8, where another HSP70 family member was possible; ACTB, where ACTG1 was also possible). While many more protein hits were obtained from the larger set of digests analyzed by MALDI-TOF/TOF, the number of distinct proteins and genes identified was not proportionately greater.

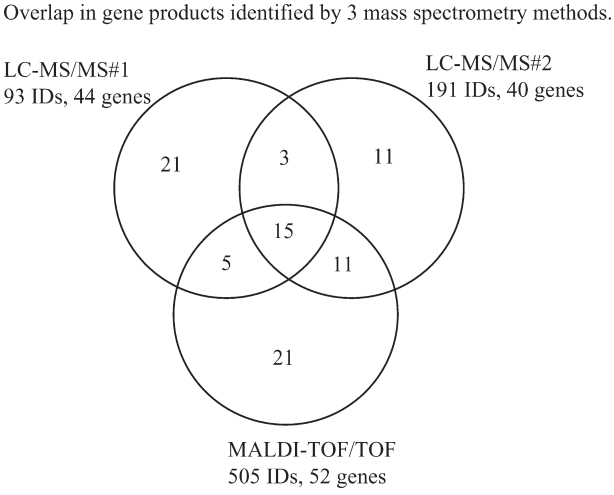

Figure 2. Venn diagram showing the overlap in identifications made by the three MS methods, after the reduction of identifications from protein accession numbers to HGNC gene symbols. As noted in Section 2, the three participating MS laboratories examined different fractions, so this diagram is descriptive, not comparative. Overlaps, therefore, represent different isoforms of the relevant proteins identified by gene accession number.

Analysis of intact proteins allows identification of protein variation resulting from alternative splicing, cleavage, and other types of PTM. A sampling of proteins that exhibited quantitative differences between serum and plasma, following separation of intact proteins based on charge, hydrophobicity, and MW, yielded five identified proteins that varied in band intensity by more than 1000-fold (Table 2), while less stringent fold changes exhibited progressively more identified proteins. As expected, several differences in band intensities between serum and plasma, including fibrinogens alpha, beta, and gamma, and vitronectin, were attributable to proteins related to coagulation (Table 3). Interestingly, many of the same proteins were identified in multiple fractions, exhibiting different p*I*, hydrophobicity, or molecular mass, reflective of the expression of different protein isoforms or specific protein cleavage products (Table 3). Several of the proteins that were identified multiple times were involved in the coagulation cascade. These were likely sampled more often than other proteins due to the marked difference in the intensity of their corresponding bands between plasma and serum specimens, as is illustrated in Fig. 1 for the coagulation factor 2 protein (prothrombin).

Table 2. Number of bands detected (with proteins subsequently identified by MS) according to fold change differences in Cy-dye labeling

Fold change differences	Number of bands detected
>2-fold	591
>3-fold	433
>5-fold	270
>10-fold	142
>100-fold	13
>1000-fold	5

As listed in Table 3, the protein most frequently detected as the top-scoring protein in the digest was the product of the APOA1 gene, detected 69 times, and found in digests from the most acidic to the most basic first-dimension fractions. In all, 26 different proteins were detected more than ten times each, representing relatively abundant proteins. In many cases, several different protein products from the same gene were found in bands at different MWs and p*I* locations, as exemplified by complement component 3 precursor (C3) and clusterin (CLU) (Fig. 3).

The C3 precursor is a 1663 amino acid chain (mass 180 kDa), which is known to undergo a variety of biologically important cleavage events [9, 10]. The 41 Kb C3 gene (gi:1165941) encodes a beta chain of mass 60 kDa and an alpha chain of mass 120 kDa, held together following cleavage by a disulfide linkage. Figure 3A indicates the distribution of 39 picked gel plugs in which C3 was identified. Peptide sequences obtained from bands estimated to be 72 kDa suggested that these bands were most likely C3 beta (amino acids 23–667). The slightly higher MW bands were unrelated to C3 beta, and instead contained a subset of peptides from C3 alpha (amino acids 672–1663), and are presumably C3 alpha after cleavage at amino acid 1303 by factor I. The highest MW band identified as C3 (about 120 kDa) contained C3 alpha peptides, probably corresponding to C3 alpha (amino acids 749–1663) following release of the C3a anaphylotoxin fragment (amino acids 672–748). Matching peptides from MS/MS searches for the lowest MW C3 bands were contained in a processed fragment of the alpha chain referred to as C3dg (amino acids 955–1303).

Figure 3B shows the distribution of 14 picked gel plugs in which CLU was identified. The CLU precursor (gi:42740907) of 449 amino acids (mass about 40 kDa) has a 22 amino acid signal sequence that is cleaved, as in com-

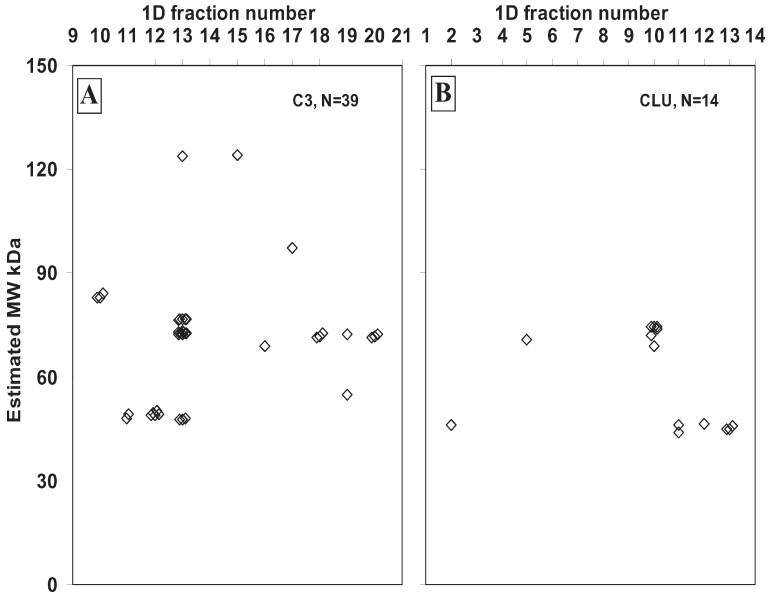

Figure 3. First-dimension fraction numbers and estimated MW of the identified proteins. Left (A): 39 gel locations where complement C3 (C3) was identified. Bands for 1-D fraction number 13 are shown in Fig. 1. Right (B): 14 gel locations where CLU was identified.

Table 3. Summary of protein identifications obtained from all mass spectrometric methods utilized, and listed according to HGNC gene symbol. For all identifications a representative protein is indicated. Lowest and highest fraction numbers refer to the 20 first-dimension fractions in IEF (fractions 1–20 based on p*I*). Number of times a band was found to differ by ±3-fold (or greater) between pairs of labeled specimens (Cy2, Cy3, Cy5) is indicated. Only the top-scoring protein for each digest was considered

Protein name	Accession	HGNC gene symbol	Times identified	Highest fraction where found	Lowest fraction where found	Highest MW where found	Lowest MW where found	Times best hit	Citrate/serum 3-fold	Citrate/serum −3-fold	EDTA/serum 3-fold	EDTA/serum −3-fold	EDTA/citrate 3-fold	EDTA/citrate −3-fold
Apo-A-I protein	178777	APOA1	69	20	1	190	9	63	20	17	13	12	15	10
Kininogen	4504893	KNG	55	13	2	223	57	54	19	6	8	3	7	5
Fibrinogen alphaA	223918	FGA	51	20	1	152	33	45	22	5	21	2	4	7
Alpha2-HS glycoprotein	2521983	AHSG	39	12	6	250	54	34	10	11	8	2	13	4
Complement component 3	4557385	C3	39	20	10	124	48	34	3	13	5	9	8	1
Fibrinogen gamma	223170	FGG	28	20	5	69	37	27	9	2	18	1	11	0
HPX protein	13529281	HPX	31	20	1	98	41	27	8	11	2	3	4	5
Fibrinopeptide B	229188	FGB	22	20	1	71	55	20	6	4	7	1	5	0
Alpha-1-acid glycoprotein 2	29170378	ORM2	22	20	1	97	35	20	2	12	2	7	6	2
Ceruloplasmin (ferroxidase)	4557485	CP	30	15	2	188	60	17	0	9	1	1	11	1
Inter-alpha-trypsin inhibitor	1082547	ITIH4	25	15	13	147	42	17	2	6	2	3	2	0
Alpha-1-anti-chymotrypsin	443345	SERPINA3	16	10	1	70	61	16	7	3	2	2	0	5
Serum albumin precursor	P02768	ALB	14	5	5	149	60	14	0	1	0	0	0	0
Apolipoprotein C-III	4557323	APOC3	14	11	3	13	7	14	7	2	6	0	2	4
Leucine-rich alpha-2-glycoprotein	21707947	LRG1	14	10	7	192	51	14	5	2	4	4	0	2
Alpha-1-B-glycoprotein	69990	A1BG	21	14	1	186	54	13	4	4	4	0	2	0
Sulfated glycoprotein-2	338057	CLU	14	13	2	75	44	12	6	2	3	1	1	2
Alpha-2-macroglobulin	177872	A2M	11	20	13	222	62	10	0	3	1	0	1	0
Afamin	4501987	AFM	17	15	13	89	82	10	0	5	0	8	0	8
Complement component 4A	7671645	C4A	11	20	5	129	18	10	2	3	2	2	2	1
Complement factor H	4504375	HF1	13	15	5	200	143	10	2	1	3	0	1	1
Apolipoprotein A-IV	114006	APOA4	19	20	7	200	37	9	1	6	1	1	3	1
Vitronectin	18201911	VTN	10	13	1	89	57	9	4	0	5	0	0	0
Orosomucoid 1	20070760	ORM1	8	2	1	54	51	8	2	3	1	4	3	4
Complement component 1 inhibitor	4557379	SERPING1	8	5	1	102	87	8	0	4	0	2	1	0
Angiotensinogen	532198	AGT	7	13	13	64	63	7	1	3	0	3	0	0
Serum amyloid P component precursor	4502133	APCS	7	19	13	37	35	7	0	3	0	2	2	0
Complement factor B	13278732	BF	10	20	8	102	91	7	1	1	1	0	0	0
Antithrombin III	999513	SERPINC1	10	20	13	188	60	7	2	3	2	1	2	0
Alpha-2-antiplasmin precursor	112907	SERPINF2	11	13	13	74	72	7	2	0	2	0	0	0
Transthyretin	5107791	TTR	8	19	1	98	16	7	3	1	1	2	1	2
Glial fibrillary acidic protein	P14136	GFAP	6	5	5	84	52	6	0	1	0	0	2	0
Apolipoprotein A-II	4502149	APOA2	6	11	10	192	10	5	1	1	2	2	2	3
Gelsolin	4504165	GSN	7	20	14	97	89	5	0	1	0	3	0	1
Apolipoprotein D	619383	APOD	4	10	5	68	40	4	2	1	2	0	1	0
HGF activator	4504383	HGFAC	6	14	14	99	97	4	3	1	3	0	0	0
Plasminogen	130319	PLG	4	20	19	107	93	4	0	1	0	0	0	2

Table 3. Continued

Protein name	Accession	HGNC gene symbol	Times identified	Highest fraction where found	Lowest fraction where found	Highest MW where found	Lowest MW where found	Times best hit	Citrate/serum 3-fold	Citrate/serum −3-fold	EDTA/serum 3-fold	EDTA/serum −3-fold	EDTA/citrate 3-fold	EDTA/citrate −3-fold
Peroxisomal acyl-coenzyme A thioester hydrolase 1	O14734	PTE1	4	5	5	250	123	4	0	0	0	0	0	1
Attractin	3676347	ATRN	5	14	14	200	190	3	0	3	0	0	3	0
Zinc-alpha-2-glycoprotein	4699583	AZGP1	3	13	10	104	51	3	0	1	0	1	0	1
Complement 9	2258128	C9	3	10	10	78	75	3	3	0	2	0	0	1
Fibronectin precursor	31397	FN1	4	15	14	250	200	3	0	0	0	0	0	0
Complement factor I	116133	IF	3	15	13	47	46	3	0	0	0	0	0	0
Lumican	642534	LUM	7	14	5	99	73	3	2	0	1	0	0	0
Alpha-1-microglobulin/bikunin precursor	4502067	AMBP	2	12	12	43	41	2	0	2	0	2	0	0
Beta-2-glycoprotein	6435718	APOH	2	20	19	60	58	2	1	0	0	0	0	1
Chain A, Complement Protein C8gamma	21730336	C8G	2	20	19	25	23	2	0	1	1	0	2	0
Prothrombin	1335344	F2	4	14	13	84	81	2	0	1	0	0	0	0
Vitamin D-binding protein	32483410	GC	6	15	13	200	59	2	0	1	0	0	1	0
Histidine-rich glycoprotein	4504489	HRG	2	15	13	49	45	2	1	0	1	0	0	0
Heat shock 70 kDa protein 8 isoform 1	5729877	HSPA8	2	8	8	73	73	2	1	0	1	0	0	0
Inter-alpha inhibitor, H2	4504783	ITIH2	4	14	11	190	115	2	1	0	0	0	0	0
Inhibitor, H3 polypeptide	10092579	ITIH3	2	8	8	125	125	2	1	0	1	0	0	0
Lysozyme precursor	4557894	LYZ	2	11	8	47	32	2	1	0	1	0	0	0
Microtubule-associated protein 1A	19861596	MAP1A	3	5	5	250	250	2	1	0	1	0	0	0
Fenestrated-endothelial linked structure protein	12963353	PLVAP	2	5	5	250	250	2	1	0	0	0	0	0
Pigment epithelial-differentiating factor	423038	SERPINF1	2	17	15	62	52	2	0	1	0	0	1	0
Alpha-1-acid glycoprotein 1	P02763	A1AG	1	5	5	52	52	1	0	0	0	1	0	1
Leucine-rich alpha-2-glycoprotein precursor	P02750	A2GL	1	5	5	56	56	1	0	1	0	0	1	0
BAG-family molecular chaperone regulator-2	O95816	BAG2	1	5	5	83	83	1	1	0	1	0	0	0
Complement component 4 binding protein, alpha	4502503	C4BPA	1	20	20	74	74	1	0	0	0	1	0	0
Complement component 8, beta polypeptide	4557391	C8B	1	20	20	65	65	1	0	0	0	0	0	0
Creatine kinase-B	180555	CKB	1	5	5	51	51	1	0	0	0	0	0	0
Dermcidin precursor	P81605	DCD	1	5	5	250	250	1	0	0	0	0	0	0
Electron transfer flavoprotein-ubiquinone oxidoreductase	Q16134	ETFDH	1	5	5	149	149	1	1	0	1	0	1	0
Flightless-I homolog	2135121	FLII	2	5	5	250	126	1	1	0	0	0	0	0

Table 3. Continued

Protein name	Accession	HGNC gene symbol	Times identified	Highest fraction where found	Lowest fraction where found	Highest MW where found	Lowest MW where found	Times best hit	Citrate/serum 3-fold	Citrate/serum −3-fold	EDTA/serum 3-fold	EDTA/serum −3-fold	EDTA/citrate 3-fold	EDTA/citrate −3-fold
Haptoglobin precursor	67586	HP	1	10	10	60	60	1	1	0	0	0	0	0
Lactotransferrin precursor	P02788	LTF	1	15	15	88	88	1	0	0	0	0	0	0
Proteoglycan 4	5031925	PRG4	1	5	5	250	250	1	0	0	0	0	0	0
dJ14N1.2 (novel S-100/ICaBP type calcium binding domain protein	12314268	S100A13	1	10	10	59	59	1	0	0	0	0	1	0
S100 calcium-binding protein A9; calgranulin B	4506773	S100A9	1	11	11	99	99	1	0	0	0	0	0	0
Nesprin-2 gamma	17861384	SYNE2	1	5	5	250	250	1	1	0	0	0	0	0
Tropomyosin 4	12803959	TPM4	1	10	10	68	68	1	1	0	1	0	0	1

plement C3. The resultant protein is subsequently cleaved into alpha and beta chains of 205 and 222 amino acids, respectively, that are joined by five disulfide bonds [11]. We identified CLU with a molecular mass of approximately 45 kDa, a total of seven times. In three instances, peptides were found to match only to the alpha chain, while only peptides from the beta chain were found in the other four bands. Protein bands for the beta chain had a slightly higher apparent MW than protein bands for the alpha chain. Both alpha and beta chains are reported to be approximately 40 kDa as a result of each chain having three putative *N*-linked glycosylation sites and perhaps additional PTMs. Peptides matching both alpha and beta chains were identified in the remaining seven bands with an apparent MW of 70 kDa. These proteins may be dimers due to incomplete reduction of all of the disulfide bonds.

Other identified proteins also yielded multiple isoforms. For Apo A-I, the majority of bands had an estimated MW of 30 kDa, corresponding to the NCBI-nr value of 30 745 and SwissProt value of 30 759. Some bands at lower MW may represent cleavage products. For Apo C-III, most bands were in the 7–10 kDa range, corresponding to the NCBI values of 8759 for Apo C-III and 10 846 for Apo C-III precursor. Vitronectin bands were at approximately 65 kDa, in apparent contrast with an NCBI value of 54 kDa. Vitronectin has been reported as a 75 kDa glycoprotein that circulates as a folded monomer in plasma [12]. SERPINA3 (also known as alpha-1 antichymotrypsin, ACT) had estimated MWs of 60–70 kDa, contrasted with an NCBI value of 46 kDa. Hwang *et al.* [13] have reported SERPIN A3/ACT in the brain with MW of 60–65 kDa and in the liver at 66–75 kDa, all of which were reduced to the 46 kDa theoretical MW associated with the polypeptide backbone after deglycosylation with *N*-glycosidase F.

Processing of serum and plasma proteins, including cleavages, might contribute unique fingerprints for various disease states [14–16]. Alternatively, protein cleavage and degradation might result from breakdown of cellular elements and activation of proteases during blood collection and aliquoting. These phenomena may be more accentuated by certain methods of sample processing compared to others. It was therefore of interest to explore the extent of protein modifications in serum compared to plasma. In this study, we investigated whether bands that were more intense in the serum than in citrate- and EDTA-anticoagulated plasma samples tended to be of lower MW than those for which the serum did not give the most intense band. Using data for 25 genes whose protein products were identified as the best match at least six times, and that also showed variation in their apparent MW (largest MW at least 10% greater than the smallest MW), we performed Rank-Sum tests to determine whether the bands that were more intense in serum were the lower MW bands for that gene (Table 4). For four of the genes (C3, KNG, CLU, A1BG) we obtained test statistics with p-values < 0.01, whereas we expect only 0.25 ($= 25 \times 0.01$) such genes by chance. For only one gene did the test statistic give $p < 0.05$ in the other direction, corresponding to higher MW bands being more intense in the serum (and no genes gave $p < 0.01$), whereas we expect at least one test this significant by chance alone. A similar exercise categorizing bands based on whether intensity was larger or smaller in the citrate *versus* EDTA samples gave no genes with p-values < 0.01.

4 Discussion

We have recently described implementation of an orthogonal 3-D IPAS, coupled with immunodepletion of abundant proteins and tagging of the remaining proteins, to quantitatively profile the human plasma proteome [1]. We applied this strategy to profile the plasma proteome for changes related to acute graft-*versus*-host disease (GVHD), following allogeneic

Table 4. Genes identified as best match at least six times, for which highest and lowest MW estimates differed by at least 10%. We excluded genes for which the number of protein bands that were most intense in the serum were all or none. *p*-Values are small when bands with smaller MW were most intense in the serum sample for that gene product

Gene symbol	Number of bands with this gene as best match	Number of times largest in serum	*p*-Value from Rank-Sum test
C3	34	23	0.003
KNG	54	14	0.005
CLU	12	3	0.006
A1BG	13	4	0.007
ITIH4	17	11	0.054
APOC3	14	1	0.086
APOA1	63	21	0.088
CP	17	6	0.211
SERPINC1	7	3	0.240
ORM2	20	14	0.255
BF	7	2	0.349
AHSG	34	9	0.400
C4A	10	5	0.458
TTR	7	1	0.500
SERPINA3	16	5	0.567
HF1	10	5	0.623
FGB	20	6	0.629
ALB	14	10	0.664
HPX	27	14	0.669
A2M	10	5	0.699
LRG1	14	4	0.802
FGG	27	1	0.848
GFAP	6	2	0.918
FGA	45	7	0.921
APOA4	9	7	0.980

bone marrow transplantation [7]. Using capillary HPLC-ESI Q-TOF MS, we identified a large number of proteins with expected concentrations across a very broad range that exhibited quantitative differences between the preGVHD and postGVHD samples. These proteins included several acute phase reactants and many immunologically active proteins that likely reflect the inflammatory processes of GVHD in these bone marrow transplant patients. IPAS thus provides a combination of comprehensive profiling and quantitative analysis, with a substantial dynamic range, as required for disease-related applications.

The present study was primarily aimed at detecting differences between serum and plasma preparations from a single pooled donor blood specimen (BD-B1 HUPO PPP reference specimens [8]). Proteins expected to be depleted during the coagulation process were generally much lower in serum compared with EDTA-plasma or citrate-plasma (see Table 3). With the evolving status of the DeCyder software and the complication that the Cy dyes were not designed to be completely equivalent for hydrophobicity (they were designed only for charge and MW), we are cautious in drawing detailed conclusions about differences in the fluorescence ratios across the second-dimension fractions.

The most remarkable aspect of this study was the detailed analysis of certain proteins for p*I* and especially MW variation among identified isoforms. The examples of complement C3 and CLU have been highly informative. The estimated MWs obtained for complement C3 fit remarkably well the published MWs of the precursor, the disulfide-bound beta plus alpha chains following cleavage of the *N*-terminal signal peptide, the individual beta and alpha chains, and the individual subsets of the alpha chain. While the associations with MWs were simply indications of potential matches to cleavage isoforms, the specific peptide sequences determined by LC-MS/MS or MALDI-TOF/TOF exactly match the specific significant isoforms of complement C3 and CLU. The same approach could be applied to many other proteins, as indicated with Apo A-I, Apo C-III, and vitronectin.

These findings have profound implications. They suggest that for most proteins detectable in serum or plasma that may represent potential markers for disease, the specific forms in which they occur may not be directly predictable either from their full length sequence or from the form(s) in which they occur in their tissues of origin. This raises the possibility that cleaved or modified forms of proteins may be depleted or enriched from tissues as they are released into the circulation or may be modified following release. Therefore, the occurrence or changes in the levels of such forms in disease is only detectable in extracellular fluids and not in their tissues of origin. This necessitates direct analysis of biological fluids for their identification, as there is no direct relationship between their levels in tissues and biological fluids. The current interest in plasma- and serum-based biomarkers necessitates the development of protocols that are optimized to avoid protein artifacts that may result from breakdown of cellular elements, degradation, and other types of modifications of circulating proteins. Numerous studies have pointed to protein modifications as an important source for biomarkers. For example, Marshall *et al.* [14] have observed that, in myocardial infarction, a spectral pattern originated from the cleavage of complement C3 alpha chain to release the C3f peptide and cleavage of fibrinogen to release peptide A. The fibrinogen peptide A and complement C3f peptide were, in turn, progressively truncated by aminopeptidases to produce two families of fragments that formed a characteristic spectral pattern of myocardial infarction. The peptide patterns observed reflected the balance of disease-specific-protease *in vivo* and aminopeptidase activity *ex vivo* [14, 17].

It was therefore of interest to explore the extent of protein modifications in serum compared to plasma. We investigated whether bands that were more intense in the serum than in plasma samples tended to be of lower MW than those for which the serum did not give the most intense band.

Using data for 25 genes whose protein products were identified as the best match at least six times and which also showed some variation in their apparent MW, we obtained test statistics with p-values < 0.01 for four of the genes (C3, KNG, CLU, A1BG), which is significant. A similar analysis categorizing bands based on whether intensity was larger or smaller in the citrate *versus* EDTA samples gave no genes with p-values < 0.01. These findings, though of a limited nature, point to a potentially greater source of variability in serum patterns, as compared to plasma patterns, one that may be attributable to sample preparation and resulting modifications. The IPAS approach implemented in this study would be well-suited to investigate the range of modifications associated with various sample preparation procedures, as well as the range of modifications that may be disease specific.

5 References

[1] Hanash, S., *Nature* 2003, *422*, 226–232.

[2] Wang, H., Hanash, S., *Mass Spectrom. Rev.* 2005, *24*, 413–426.

[3] Pedersen, S. K., Harry, J. L., Sebastian, L., Baker, J., Traini, M. D., McCarthy, J. T., Manoharan, A. *et al.*, *J. Proteome Res.* 2003, *2*, 303–311.

[4] Zuo, X., Speicher, D. W., *Proteomics* 2002, *2*, 58–68.

[5] Washburn, M. P., Wolters, D., Yates, J. R., 3rd, *Nat. Biotechnol.* 2001, *19*, 242–247.

[6] Pieper, R., Su, Q., Gatlin, C. L., Huang, S. T., Anderson, N. L., Steiner, S., *Proteomics* 2003, *3*, 422–432.

[7] Wang, H., Clouthier, S. G., Galchev, V., Misek, D. E., Duffner, U., Min, C.-K., Zhao, R. *et al.*, *Mol. Cell. Proteomics* 2005, *4*, 618–625.

[8] Omenn, G. S., *Proteomics* 2004, *4*, 1235–1240.

[9] de Bruijn, M. H., Fey, G. H., *Proc. Natl. Acad. Sci. USA* 1985, *82*, 708–712.

[10] Fong, K. Y., Botto, M., Walport, M. J., So, A. K., *Genomics* 1990, *7*, 579–586.

[11] Rosenberg, M. E., Silkensen, J., *Int. J. Biochem. Cell. Biol.* 1995, *27*, 633–645.

[12] Underwood, P. A., Kirkpatrick, A., Mitchell, S. M., *Biochem. J.* 2002, 365, 57–67.

[13] Hwang, S. R., Steineckert, B., Kohn, A., Palkovits, M., Hook, V. Y., *J. Biol. Chem.* 1999, *274*, 1821–1827.

[14] Marshall, J., Kupchak, P., Zhu, W., Yantha, J., Vrees, T., Furesz, S., Jacks, K. *et al.*, *J. Proteome Res.* 2003, *2*, 361–372.

[15] Zhang, Z., Bast, R. C., Jr., Yu, Y., Li, J., Sokoll, L. J., Rai, A. J., Rosenzweig, J. M. *et al.*, *Cancer Res.* 2004, *64*, 5882–5890.

[16] He, Q. Y., Lau, G. K., Zhou, Y., Yuen, S. T., Lin, M. C., Kung, H. F., Chiu, J. F., *Proteomics* 2003, *3*, 666–674.

[17] Zhang, R., Barker, L., Pinchev, D., Marshall, J., Rasamoelisolo, M., Smith, C., Kupchak, P. *et al.*, *Proteomics* 2004, 4, 244–256.

