DOI 10.1002/pmic.200500358 *Proteomics* 2005, *5*, 3226–3245

REGULAR ARTICLE

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

Gilbert S. Omenn[1], David J. States[1], Marcin Adamski[1], Thomas W. Blackwell[1], Rajasree Menon[1], Henning Hermjakob[2], Rolf Apweiler[2], Brian B. Haab[3], Richard J. Simpson[4], James S. Eddes[4], Eugene A. Kapp[4], Robert L. Moritz[4], Daniel W. Chan[5], Alex J. Rai[5], Arie Admon[6], Ruedi Aebersold[7, 8], Jimmy Eng[8], William S. Hancock[9], Stanley A. Hefta[10], Helmut Meyer[11], Young-Ki Paik[12], Jong-Shin Yoo[13], Peipei Ping[14], Joel Pounds[15], Joshua Adkins[15], Xiaohong Qian[16], Rong Wang[17], Valerie Wasinger[18], Chi Yue Wu[19], Xiaohang Zhao[20], Rong Zeng[21], Alexander Archakov[22], Akira Tsugita[23], Ilan Beer[24], Akhilesh Pandey[5], Michael Pisano[25], Philip Andrews[1], Harald Tammen[26], David W. Speicher[27] and Samir M. Hanash[1, 28]

[1] University of Michigan, Ann Arbor, MI, USA*

Received: May 23, 2005
Accepted: May 24, 2005

HUPO initiated the Plasma Proteome Project (PPP) in 2002. Its pilot phase has (1) evaluated advantages and limitations of many depletion, fractionation, and MS technology platforms; (2) compared PPP reference specimens of human serum and EDTA, heparin, and citrate-anticoagulated plasma; and (3) created a publicly-available knowledge base (www.bioinformatics.med.umich.edu/hupo/ppp; www.ebi.ac.uk/pride). Thirty-five participating laboratories in 13 countries submitted datasets. Working groups addressed (a) specimen stability and protein concentrations; (b) protein identifications from 18 MS/MS datasets; (c) independent analyses from raw MS-MS spectra; (d) search engine performance, subproteome analyses, and biological insights; (e) antibody arrays; and (f) direct MS/SELDI analyses. MS-MS datasets had 15 710 different International Protein Index (IPI) protein IDs; our integration algorithm applied to multi-

* For other affiliations please see Addendum

Correspondence: Dr. Gilbert S. Omenn, Internal Medicine, University of Michigan, MSRB 1, 1150 W. Medical Center Dr. Ann Arbor, MI 48109-0656, USA
E-mail: gomenn@umich.edu
Fax: +1-734-647-8148

Abbreviations: AMT, accurate mass tag; **BD**, BD Diagnostics; **CAMS**, Chinese Academy of Medical Sciences; **CTAD**, citrate, theophylline, adenosine, dipryridamole; **EBI**, European Bioinformatics Institute; **EGF**, epithelial growth factor; **ETH**, Eidgenössische Technische Hochschule Zürich (Swiss Federal Institute of Technology in Zurich); **FFE-IEF**, free flow electrophoresis; **GO**, Gene Ontology; **HBV**, Hepatitis B virus; **HCV**, Hepatitis C virus; **HIV**, human immunodeficiency virus; **HTLV-1**, Human T-cell lymphotrophic virus; **IPI**, International Protein Index; **IQ**, IQ calmodulin; **ISB**, Institute for Systems Biology; **LCQ, LCQ-Deca-XP+**, LTQ, LTQ-linear IT-MS/MS, proprietary names of Thermo Finnigan LC/MS instruments, LCQ = 3 dimensional ion trap chamber; LTQ = linear ion trap chamber; **MS3**, MS-FT-ICR-MS, three MS stages; **MudPIT**, Multidimensional protein identification technology; **NIBSC**, National Institute of Biological Standards and Control; **PE-PGRS**, families of glycine rich proteins in mycobacterium tuberculosis; **PKD**, polycystic kidney disease; **PPP**, Plasma Proteome Project; **PRIDE**; PRoteomics IDEntifications database; **RING**, zinc finger protein; **SCX**, strong cation exchange; **SEQUEST Sf**, search algorithm, final score; **SERPINA3**, protein name, (not an abbreviation); **SNPs**, single nucleotide polymorphisms; **SQL**, Structured Query Language; **VCAM**, vascular cell adhesion molecule; **WAX**, weak anion exchange; **X!Tandem**, open source peptide search algorithm for mass spectrometry

ple matches of peptide sequences yielded 9504 IPI proteins identified with one or more peptides and 3020 proteins identified with two or more peptides (the Core Dataset). These proteins have been characterized with Gene Ontology, InterPro, Novartis Atlas, OMIM, and immunoassay-based concentration determinations. The database permits examination of many other subsets, such as 1274 proteins identified with three or more peptides. Reverse protein to DNA matching identified proteins for 118 previously unidentified ORFs.

We recommend use of plasma instead of serum, with EDTA (or citrate) for anticoagulation. To improve resolution, sensitivity and reproducibility of peptide identifications and protein matches, we recommend combinations of depletion, fractionation, and MS/MS technologies, with explicit criteria for evaluation of spectra, use of search algorithms, and integration of homologous protein matches.

This Special Issue of PROTEOMICS presents papers integral to the collaborative analysis plus many reports of supplementary work on various aspects of the PPP workplan. These PPP results on complexity, dynamic range, incomplete sampling, false-positive matches, and integration of diverse datasets for plasma and serum proteins lay a foundation for development and validation of circulating protein biomarkers in health and disease.

Keywords:
Database / HUPO Plasma Proteome Project / Plasma / Serum

1 Introduction

A comprehensive, systematic characterization of circulating proteins in health and disease will greatly facilitate development of biomarkers for prevention, diagnosis, and therapy of cancers and other diseases [1]. Proteomics technologies now permit extensive fractionation of proteins in complex specimens, analysis of peptides by MS, and matching of peptide sequences to protein "hits" through gene and protein databases generated directly and indirectly from the sequencing of the human genome [2, 3], as well as other methods for identifying proteins.

The HUPO, formed in 2001, aims to accelerate the development of the field of proteomics and to stimulate and organize international collaborations in research and education [4]. HUPO has launched major initiatives focused on the plasma, liver, and brain proteomes, proteomics standards and databases, and large-scale antibody production. The plasma proteome is linked with these other initiatives (see Fig. 1).

The long-term scientific goals of the HUPO Plasma Proteome Project (PPP) are (1) comprehensive analysis of the protein constituents of human plasma and serum; (2) identification of biological sources of variation within individuals over time due to physiology (age, sex, menstrual cycle, exercise, stress), pathology (various diseases, special cohorts), and treatments (common medications); and (3) determination of the extent of variation across individuals within populations and across populations due to genetic, nutritional and other factors. The pilot phase aims to (1) compare advantages and limitations of many technology platforms; (2) contrast reference specimens of human plasma (EDTA, heparin, or citrate-anticoagulated) and serum in terms of numbers of proteins identified and any interferences with various technology platforms; and (3) create a global, open-source knowledge base/data repository.

The collaborative nature of this Project permitted exploration of many variables and adoption during the study phase of emerging technologies. Planning proceeded expeditiously from the organizing meeting of HUPO in Bethesda in April 2002, to the first PPP meeting in Ann Arbor in September 2002, the expression of interest by numerous investigators at the 1st HUPO World Congress on Proteomics in Versailles in November 2002, and then the PPP Workshop for Technical Committees and participating laboratories in Bethesda in July 2003 to launch the pilot phase. PPP reference specimens were prepared and distributed, beginning in September 2003, and first data were submitted, analyzed,

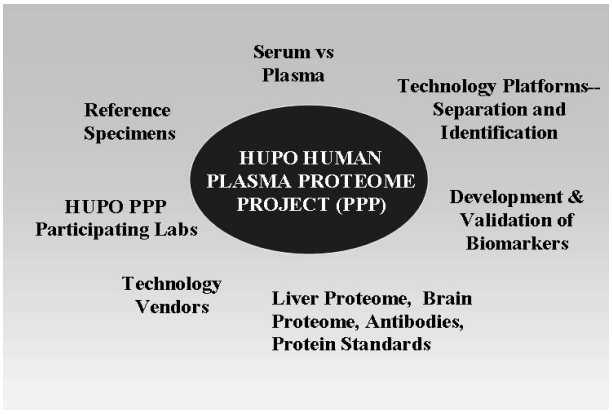

Figure 1. Schema showing relationship of HUPO Plasma Proteome Project (PPP) to other HUPO initiatives and components of the PPP.

and presented at a workshop at the 2nd HUPO World Congress in Montreal in November 2003. An intensive 4 day Jamboree Workshop was organized for Ann Arbor in June 2004, at which numerous work groups pursued cross-laboratory analyses and proposed further work. Investigators were advised to adopt more stringent criteria for high confidence peptide and protein identifications, and a commitment was made to collect raw spectra from the 18 laboratories that had submitted MS/MS or FT-ICR/MS datasets for independent analyses by three different groups. The datasets were moving targets, as some, but not all, labs submitted expanded or updated analyses, and about 15 laboratories completed "special projects" stimulated by HUPO PPP with a competition for small grants following the Montreal workshop.

The PPP provided participating laboratories with 1.0 mL of reference specimens of serum and plasma by three different methods of anticoagulation for plasma (EDTA, citrate, heparin) from specific donor pools. Investigators utilized their established and emerging technologies for fractionation and analysis of proteins. Investigators were encouraged to "push the limits" of their methods to detect and identify low abundance proteins. Comparisons of findings across laboratories provide a special opportunity for confirmation of protein identifications. Results were submitted to centralized bioinformatics functions at the University of Michigan and the European Bioinformatics Institute to create an integrated data repository from which PPP and other investigators could initiate further analyses and annotations. The approaches and core results have been presented at the US HUPO inaugural meeting in March 2005 and at other meetings. Certain of the results are highlighted in a companion manuscript (Omenn *et al.*, submitted).

Here we present a comprehensive account of the major findings from the pilot phase of the Human Plasma Proteome Project, including the many associated special projects.

2 PPP reference specimens

The primary specimens were sets of four reference specimens prepared under the direction of the HUPO PPP Specimens Committee by BD Diagnostics for each of three ethnic groups: Caucasian-American (B1), African-American (B2), and Asian-American (B3). Each pool consisted of 400 mL of blood each from one male and one post-menopausal female healthy, fasting donor, collected into 10 mL tubes in a prescribed sequence (see Supplementary Protocol) after informed consent. Very large pools were rejected as requiring too prolonged specimen handling and processing unlike the collection of individual specimens; even a protocol for two males and two females proved to require more than the 2 h limit we set. Equal numbers of tubes and aliquots were generated with appropriate concentrations of K_2-EDTA, lithium heparin, or sodium citrate for plasma or permitted to clot at room temperature for 30 min to yield serum (with micronized silica as clot activator). The additives were dry-sprayed on the inner walls of the tubes, except for 1.0 mL of 0.105 M buffered sodium citrate, which gave a final ratio of 9:1 for blood to citrate in a 10 mL final volume, causing an 11% dilution of the blood. No protease inhibitor cocktails were used. This procedure required 2 h, mostly at 2 to 6°C. After centrifugation, volumes from the male and female donors in each donor pair for each specimen type were pooled and then aliquoted into numerous 250 μL portions in vials which were frozen and stored at −70°C. The centrifugation conditions with citrate consistently produced platelet-poor plasma (platelet count $<10^3$/μL). Aliquots tested negative for HIV, HBV, HCV, HTLV-1, and syphilis. We supplied four × 250 μL aliquots for each of the four plasma/serum specimens in each set. These vials were shipped on dry ice *via* courier in early May 2003 (and later to additional laboratories which petitioned to join the project, some of which could no longer be supplied the B1 set). No reshipping was permitted.

The Chinese Academy of Medical Sciences (CAMS) used a variant of the BD protocol to generate similar reference serum and plasma specimens, as described by Li *et al.* [5] and He *et al.* [6]. Pools were prepared after review by the CAMS Ethics Committee and informed consent by ten male and ten female donors in Beijing. Donors were fasting and avoided taking medicines or drinking alcohol for the 12 h before sampling. A subsequent pooling of 20 mL from each of the male and female serum or plasma specimens created the C1-CAMS PPP reference specimens which were sent to the 15 laboratories requesting these specimens after storage at −80°C. They were shipped on dry ice using the same courier in September 2003. C1-CAMS specimens were centrifuged originally, and then again upon thawing, at 4°C [6].

Finally, the UK National Institute of Biological Standards and Control (NIBSC) made available to the PPP their lyophilized citrated plasma standard prepared for hemostasis and thrombosis studies from a pool of 25 donors [1].

A standard questionnaire was sent to all laboratories expressing interest. Of 55 laboratories that originally committed to participate, 41 received the BD B1 specimens, 27 the B2 and B3 specimens, 15 the CAMS specimens, and 45 the NIBSC specimens. Laboratories varied on how many of the specimens they actually analyzed.

3 Bioinformatics and technology platforms

As intended, laboratories used a wide variety of methods, including multiple LC-MS/MS instruments, MALDI-MS, and FT-ICR-MS; depletion of abundant proteins; fractionation of intact proteins on 2-D gels or with LC or IEF methods; protein enrichment or labeling methods; immunoassays or antibody arrays; and direct (SELDI) MS. They also varied on choice of search algorithm and database, and criteria for declaring high or lower confidence identification of peptide sequences and matching proteins (Table 1). In general, the numbers of proteins reported individually by the labs do not have the integration feature which was applied to the whole PPP dataset. In several cases, much more extensive analyses were reported. Thus, many of the individual papers in this special issue have additional protein identifications not included in the project-wide dataset(s).

Table 1. Protein identifications by lab, by specimen, and by methods

Lab ID	Specimen	Depletion	Protein separation	Reduction/ alkylation	Peptide separation	Mass spectrum	Search software	3020 High confidence	3020 Lower confidence	Single peptide
1	b1-cit	aig	none	iam	rp/scx/rp	esi-ms/ms_decaxp	PepMiner	61	39	12
1	b1-edta	aig	none	iam	rp/scx/rp	esi-ms/ms_decaxp	PepMiner	35	30	14
1	b1-hep	aig	none	iam	rp/scx/rp	esi-ms/ms_decaxp	PepMiner	50	38	13
1	b1-serum	aig	none	iam	rp/scx/rp	esi-ms/ms_decaxp	PepMiner	21	6	5
1	b2-cit	aig	none	iam	rp/scx/rp	esi-ms/ms_decaxp	PepMiner	57	37	12
1	b2-hep	aig	none	iam	rp/scx/rp	esi-ms/ms_decaxp	PepMiner	58	30	12
1	b2-serum	aig	none	iam	rp/scx/rp	esi-ms/ms_decaxp	PepMiner	59	31	12
1	b3-serum	aig	none	iam	rp/scx/rp	esi-ms/ms_decaxp	PepMiner	17	6	7
2	b1-cit	none	cho affinity	iam	scx/rp	esi-ms/ms_qtof	SEQUEST	165	79	94
2	b1-serum	none	cho affinity	iam	scx/rp	esi-ms/ms_qtof	SEQUEST	136	48	38
2	nibsc	none	cho affinity	iam	scx/rp	esi-ms/ms_qtof	SEQUEST	171	121	85
11	b1-cit	none	cho affinity	iam	rp	esi-ms/ms_decaxp	SEQUEST	59	4	9
11	b1-edta	none	cho affinity	iam	rp	esi-ms/ms_decaxp	SEQUEST	64	6	4
11	b1-hep	none	cho affinity	iam	rp	esi-ms/ms_decaxp	SEQUEST	62	9	15
11	b1-serum	none	cho affinity	iam	rp	esi-ms/ms_decaxp	SEQUEST	64	3	16
12	b1-cit	aig	none	iam	rp/scx/rp	esi-ms/ms_deca	SEQUEST	111	0	113
12	b1-edta	aig	none	iam	rp/scx/rp	esi-ms/ms_deca	SEQUEST	111	0	101
12	b1-hep	aig	none	iam	rp/scx/rp	esi-ms/ms_deca	SEQUEST	127	0	130
12	b1-serum	aig	none	iam	rp/scx/rp	esi-ms/ms_deca	SEQUEST	123	0	111
17	b1-serum	aig	1s sds	iam	rp	esi-ms/ms_lcq	SEQUEST	50	19	7
21	b1-cit	top6	rotofor-ief/rp/1d-sds	iam	rp	esi-ms/ms_qtof	MASCOT	40	0	1
21	b1-cit	top6	rotofor-ief/rp/1d-sds	none	none	maldi-ms/ms^{a}bi4700	MASCOT	51	0	3
21	b1-cit	top6	rotofor-ief/rp/1d-sds	none	rp	esi-ms/ms_qtof	MASCOT	39	0	1
21	b1-edta	top6	rotofor-ief/rp/1d-sds	iam	rp	esi-ms/ms_qtof	MASCOT	40	0	1
21	b1-edta	top6	rotofor-ief/rp/1d-sds	none	none	maldi-ms/ms^{a}bi4700	MASCOT	51	0	3
21	b1-edta	top6	rotofor-ief/rp/1d-sds	none	rp	esi-ms/ms_qtof	MASCOT	39	0	1
21	b1-serum	top6	rotofor-ief/rp/1d-sds	iam	rp	esi-ms/ms_qtof	MASCOT	40	0	1
21	b1-serum	top6	rotofor-ief/rp/1d-sds	none	none	maldi-ms/ms^{a}bi4700	MASCOT	51	0	3
21	b1-serum	top6	rotofor-ief/rp/1d-sds	none	rp	esi-ms/ms_qtof	MASCOT	39	0	1
22	b1-serum	top6	1s sds	iam	rp/scx/rp	esi-ms/ms_decaxp	SEQUEST	277	0	161
24	b1-serum	a	rp	iam	rp	esi-ms/ms_qtrap	MASCOT	7	12	1
24	b1-serum	none	rp	iam	rp	esi-ms/ms_qtrap	MASCOT	17	21	3
26	b2-cit	none	rotofor-ief/1d-sds	iam	rp	esi-ms/ms_qtof	MASCOT	160	44	12
28	b1-cit	ig	none	none	rp	esi-fticr	VIPER	218	45	208
28	b1-serum	ig	none	none	rp	esi-fticr	VIPER	223	50	239
28	b2-cit	ig	none	none	rp	esi-fticr	VIPER	255	140	346
28	b2-serum	ig	none	none	rp	esi-fticr	VIPER	244	181	405
28	b3-cit	ig	none	none	rp	esi-fticr	VIPER	214	188	359
28	b3-serum	ig	none	none	rp	esi-fticr	VIPER	218	193	384
29	b1-cit	top6	none	iam	scx/rp	esi-ms/ms_decaxp	SEQUEST	19	129	136
29	b1-cit	top6	none	iam	scx/rp/2mz	esi-ms/ms_decaxp	SEQUEST	51	160	181
29	b1-edta	top6	none	iam	scx/rp	esi-ms/ms_decaxp	SEQUEST	50	199	264
29	b1-edta	top6	none	iam	scx/rp/2mz	esi-ms/ms_decaxp	SEQUEST	82	491	557
29	b1-hep	top6	none	iam	scx/rp	esi-ms/ms_decaxp	SEQUEST	26	97	122
29	b1-serum	top6	none	iam	scx/rp	esi-ms/ms_decaxp	SEQUEST	90	338	432
29	c1-cit	top6	none	iam	scx/rp/2mz	esi-ms/ms_decaxp	SEQUEST	82	449	517
29	c1-edta	top6	none	iam	scx/rp/2mz	esi-ms/ms_decaxp	SEQUEST	72	555	610
29	c1-hep	top6	none	iam	scx/rp/2mz	esi-ms/ms_decaxp	SEQUEST	82	227	283
29	c1-serum	top6	none	iam	scx/rp/2mz	esi-ms/ms_decaxp	SEQUEST	97	519	570
29	nibsc	top6	none	iam	scx/rp/2mz	esi-ms/ms_decaxp	SEQUEST	82	371	432
33	nibsc	top6	ffe/rp	none	rp/ziptip	maldi-ms/ms_qstar	Digger	54	0	0
33	nibsc	top6	ffe/rp	none	rp/ziptip	maldi-ms/ms_qstar	MASCOT	58	0	3
34	b1-hep	top6	zoom-ief/1d-sds	iam	rp	esi-ms/ms_decaxp	SEQUEST	123	148	146
34	b1-serum	top6	zoom-ief/1d-sds	iam	rp	esi-ms/ms_ltq	SEQUEST	427	741	1172
40	b1-hep	none	aig affinity/rp	iam	scx/rp	esi-ms/ms_lcq	Sonar	160	253	185

Table 1. Continued

Lab ID	Specimen	Depletion	Protein separation	Reduction/ alkylation	Peptide separation	Mass spectrum	Search software	3020 High confidence	3020 Lower confidence	Single peptide
41	b1-cit	none	gradiflow/tca	none	scx/rp	esi-ms/ms_qstar	SEQUEST	72	0	34
41	b1-edta	none	gradiflow/tca	none	scx/rp	esi-ms/ms_qstar	SEQUEST	62	0	16
41	b1-hep	none	gradiflow/tca	none	scx/rp	esi-ms/ms_qstar	SEQUEST	51	0	7
41	b1-serum	none	gradiflow/tca	none	scx/rp	esi-ms/ms_qstar	SEQUEST	76	0	27
41	nibsc	none	gradiflow/tca	none	scx/rp	esi-ms/ms_qstar	SEQUEST	53	0	1
43	b1-cit	aig	none	iam	rp	esi-ms/ms_qtof	MASCOT	26	0	0
43	b1-edta	aig	none	iam	rp	esi-ms/ms_qtof	MASCOT	31	0	0
43	b1-hep	aig	none	iam	rp	esi-ms/ms_qtof	MASCOT	37	0	0
43	b1-hep	aig	none	iam	rp	maldi-ms/ms[a]bi4700	MASCOT	26	0	0
43	b1-serum	aig	none	iam	rp	esi-ms/ms_qtof	MASCOT	24	0	0
43	nibsc	aig	none	iam	rp	esi-ms/ms_qtof	MASCOT	21	0	0
46	c1-serum	top6	none	iam	rp	esi-ms/ms_ltq	SEQUEST	185	522	571
55	b1-cit	none	sax	iam	rp	esi-ms/ms_ltq	SEQUEST	216	48	73

High and lower confidence

1. PepMiner results: score >80/100
2. ProteinProphet: high p $\geq$ 0.95; lower 0.95 > p $\geq$ 0.2
11. $X_{corr} \geq$ 1.5/2.0/2.5 for charge states +1/+2/+3. Tryptic cleavage rules. High confidence: two or more peptide ids or single peptide ID manually inspected; spectrum must show high signal and top 3 ions must be assigned either b or y. Otherwise, lower confidence
12. PeptideProphet high confidence p $\geq$ 0.35. All IDs reported as high confidence.
17. SEQUEST results: no-enzyme searches, acceptance criteria not stated. (For the automatic interpretation of fragment ion spectra the SEQUEST algorithm is used screening the NCBI protein database (weekly updated version)). The chosen parameters are: aver
21. MASCOT result; high confidence only: probability $\geq$ 98%, numerous isoforms identified
22. SEQUEST result: $X_{corr} \geq$ 1.9/2.5/3.75 for charge states +1/+2/+3, no manual inspection, no other criteria used
24. MASCOT result. High confidence: if two or more peptides, each of them has to have MASCOT score $\geq$ 20; if single peptide, it has to have MASCOT score $\geq$ 30.
26. High confidence fully bryptic peptides: MASCOT individual peptides score $\geq$21 or total score $\geq$80; if single peptide hit, score $\geq$60; if lower scores, manually inspected to check fragment ions and mass error.
28. Confidence is based on reproducibility of identification in triplicate analyses of a sample. High confidence = identification of AMT peptides for a given ORF in two or three of triplicate FT-ICR analyses. Lower confidence = identification of AMT peptides in only one of three FT-ICR analyses. VIPER and Q-Rollup software were used to match FT-ICR accurate masses to the AMT database
29. High confidence: $X_{corr} \geq$ 1.9/2.2/3.75 (for charges +1/+2/+3), deltaCn $\geq$ 0.1, and $R_{sp} \leq$ 4. Lower confidence: $X_{corr} \geq$ 1.5/2.0/2.5 (for charges +1/+2/+3), deltaCn $\geq$ 0.1
33. High confidence: Digger nxc $\geq$ 0.3; MASCOT score $\geq$ 15
34. High and lower confidence both used PPP stringent segment parameters of $X_{corr} \geq$1.9, 2.2 and 3.15; deltaCN $\geq$0.1; Rsp $\leq$4; high-two or more peptides; lower-one peptide.
40. Sonar results. High confidence: protein expect value < 1; lower confidence: protein expect value $\geq$ 1
41. DTA Select results, criteria not stated, manually inspected
43. MASCOT results: protein p-value $\leq$ 0.05 and at least one peptide with MASCOT score $\geq$ 20.
46. High confidence: $X_{corr} \geq$ 1.9/2.2/3.75 (for charges +1/+2/+3), deltaCn $\geq$ 0.1, and $R_{sp} \leq$ 4; lower confidence: X_{corr} >1.5/2.0/2.5.
55. Identical sets of .dta files were searched using SEQUEST, Sonar and X!Tandem. SEQUEST criteria: X_{corr} > 1.8/2.0/2.5 for charge states +1/+2/+3, deltaCn $\geq$ 0.1, Sp $\leq$ 200. X!Tandem criteria: expectation value $\leq$ 0

3.1 Constructing a PPP database for human plasma and serum proteins

Data management for this project included guidance and protocols for data collection, then centralized integration, analysis, and dissemination of findings worldwide *via* a communications infrastructure. As described in great detail by Adamski *et al.* [7, 8], key challenges were integration of heterogeneous datasets, reduction of redundant information to minimal identification sets, and data annotation. Multiple factors had to be balanced, including when to "freeze" on a particular release of the ever-changing database selected for the PPP and how to deal with "lower confidence" peptide identifications. Freezing of the database was essential to conduct extensive comparisons of complex datasets and annotations of the dataset as a whole. However, it complicates the work of linking findings of the current study to evolving knowledge of the human genome and its annotation. Many of the entries in the protein sequence database(s) available at the initiation of the project or even the analytical phase were revised, replaced, or withdrawn over the course of the project, and continue to be revised. Our policies and practices anticipated the guidelines issued recently by Carr *et al.* [9], as documented by Adamski *et al.* [7].

The 18 participating laboratories using MS/MS or FT-ICR-MS submitted a total of 42 306 protein identifications using various search engines and databases to handle spectra and generate peptide sequence lists from the specimens analyzed. These reports matched to 15 710 non-redundant entries (of which 15 519 were based on peptides with six or more amino acids) in the International Protein Index, which had been chosen as the standard reference database for this Project (IPI version 2.21, July 2003) [9]. We designed an integration algorithm which selected one representative protein among multiple proteins (homologs and isoforms) to which identified peptides gave 100% sequence matches. This integration process resulted in 9504 proteins in the IPI v2.21 database identified with one or more peptides. From this point of view, the PPP database is conservative, counting homologous proteins and all isoforms of particular proteins (and their corresponding genes) just once, unless the sequences actually differentiated any additional matches. We included at this stage proteins identified by matches to one or more peptide sequences of "high" or "lower" confidence according to cutpoints utilized with the various search engines used by different MS/MS instruments. Table 1 shows the details of the cutpoints or filters used by each investigator and the numbers of "high" and "lower" confidence protein IDs. All laboratories utilizing SEQUEST were asked to reanalyze their results using the PPP specified filters of X_{corr} values $\geq$ 1.9, 2.2, and 3.75 for singly, doubly, and triply charged ions, with deltaCN value $\geq$ 0.1 and $R_{sp} \geq 4$ for fully tryptic peptides for "high confidence" identifications; most did so. No equivalency rules were applied across all the search algorithms for all the cutpoints.

However, Kapp *et al.* [11] provide such a cross-algorithm analysis for three specified false-positive rates using one laboratory dataset. Since the approaches and analytical instruments used by the various laboratories (Table 1) were far too diverse to utilize a standardized set of mass spec/search engine criteria, we created a relatively stringent defined set of protein IDs from the 9504 above by requiring that the same protein be identified with at least a second peptide. In a peptide chromatography run for MS, not all peaks are selected for MS/MS analysis, and the identification of peptide fragment ions is a low-percentage sampling process. Thus, additional analyses in the same lab and in other labs would be expected to enhance the yield of peptide IDs. Consequently, MS data from the individual laboratories were combined to increase the probability of peptide and protein identification. The use of different instrumentation with proprietary software and different search engines for identification made it unfeasible to apply a standard set of parameters to peptide sequences. Therefore, we required a minimum of two distinct peptides to be inferred from mass spectra and matched 100% to the database protein sequence, as a uniform criterion for a given protein to be considered identified.

Of this total of 9504 protein IDs, 6484 were based on one peptide, while 3020 were based on two or more peptides (Table 2). That process generated the list of 3020 proteins (5102 before integration) which is utilized as our Core Protein Dataset for the HUPO PPP knowledge base. Full details with unique IPI accession numbers for each protein are accessible for examination and re-analysis at http://www.bioinformatics.med.umich.edu/hupo/ppp and www.

Table 2. Protein identifications by lab and specimen, based on two or more peptides for each protein match, generating the PPP 3020 protein core dataset

Lab Id	nibsc	b1-cit	b1-edta	b1-hep	b1-serum	b2-cit	b2-hep	b2 serum	b3-cit	b3-serum	c1-cit	c1-edta	c1-hep	c1-serum	plasma	serum	both
1	0	100	65	88	27	94	88	90	0	23	0	0	0	0	197	108	220
2	292	244	0	0	184	0	0	0	0	0	0	0	0	0	399	184	469
11	0	63	70	71	67	0	0	0	0	0	0	0	0	0	102	67	120
12	0	111	111	127	123	0	0	0	0	0	0	0	0	0	277	123	348
17	0	0	0	0	69	0	0	0	0	0	0	0	0	0	0	69	69
21	0	78	78	0	78	0	0	0	0	0	0	0	0	0	78	78	78
22	0	0	0	0	277	0	0	0	0	0	0	0	0	0	0	277	277
24	0	0	0	0	51	0	0	0	0	0	0	0	0	0	0	51	51
26	0	0	0	0	0	204	0	0	0	0	0	0	0	0	204	0	204
28	0	263	0	0	273	395	0	425	402	411	0	0	0	0	565	572	693
29	453	323	724	123	428	0	0	0	0	0	531	627	309	616	1576	867	1839
33	60	0	0	0	0	0	0	0	0	0	0	0	0	0	60	0	60
34	0	0	0	271	1168	0	0	0	0	0	0	0	0	0	271	1168	1251
40	0	0	0	413	0	0	0	0	0	0	0	0	0	0	413	0	413
41	53	72	62	51	76	0	0	0	0	0	0	0	0	0	113	76	137
43	21	26	31	43	24	0	0	0	0	0	0	0	0	0	51	24	52
46	0	0	0	0	0	0	0	0	0	0	0	0	0	707	0	707	707
55	0	264	0	0	0	0	0	0	0	0	0	0	0	0	264	0	264
	nibsc	b1-cit	b1-edta	b1-hep	b1-serum	b2-cit	b2-hep	b2-serum	b3-cit	b3-serum	c1-cit	c1-edta	c1-hep	c1-serum	plasma	serum	both
	679	1016	876	838	1749	568	88	470	402	419	531	627	309	1124	2580	2353	3020

ebi.ac.uk/pride. Figure 2 shows the numbers of proteins identified with $\geq$ n peptides with the percentage of those IDs confirmed in a second laboratory. Of these peptides, the vast majority were ten or more amino acids in length, with a median of 12.9 and a minimum of six amino acids in this dataset; the distribution of lengths is shifted to the right compared with the theoretical tryptic peptides from the total IPI database. The 3020 proteins represent a very broad sampling of the IPI proteins in terms of characterization by p*I* and by molecular weight of the transcription product (often a "precursor" protein).

The PPP database permits future users to choose their own cut-points for subanalyses, including 2857 proteins identified at least once with "high confidence" criteria; 1555 proteins based on two or more peptides, at least one of which was reported as high confidence (from the intersection of the 3020 and the 2857); and 1274 proteins based on matching to three or more peptides.

Figure 3 shows the methods used and the log of the number of proteins identified by the various laboratories. At the top of the figure are results with MALDI-MS. Four labs reported MALDI-MS without MS/MS for certain specimens. For example, Lab 22 analyzed all four samples of each of the B1, B2, and B3 specimens by MALDI-MS, and then used in-depth ESI-MS/MS Deca-xp for B1 serum only. Altogether there were 367 distinct protein IDs by MALDI-MS, of which 226 were confirmed by MS/MS or FT-ICR/MS in the core dataset of 3020 IPI proteins, while 141 were not so confirmed. The mean and median numbers of peptides for the confirmed proteins were significantly higher than for those not confirmed. The MALDI-MS data were not used in identifying the 3020 protein dataset or creating Fig. 2.

The capillary LC-FT-ICR-MS results (Lab 28) were included. This method (Adkins *et al.* [12]) depends upon previous ion-trap MS/MS studies to generate a database of highly accurate mass and normalized elution time parameters for each peptide. Proteins in new specimens cannot be recognized if those proteins were not already detected and characterized in creating (and updating) the AMT database. Only 22% of 722 proteins identified across the six PPP specimens had more than one peptide match; ProteinProphet clustered these 722 into 377 non-redundant proteins. The LC-MS/AMT method has the potential to expedite analysis of large numbers of specimens once the mass tolerance is tightened, the elution times are made highly reproducible, and the AMT parameters are known for a very substantial number of true-positive peptides. Even then, however, samples of differing origin and complexity may have different PTMs and different elution times, limiting the usefulness of the AMT tags. At present, peptide coverage seems to be quite limited. However, powerful MS-FT-ICR-MS (MS3) combinations are being introduced [13]. Lab 28 contributed valuable data on serum/plasma comparisons. Adkins *et al.* [12] also demonstrated that their approach gives a rough quantitative estimation of protein concentrations based on average ion current for all the peptides identified for 18 particular proteins, correlated in log-log plots with nephelometric immunoassay results.

The most striking difference in MS was the comparison of LCQ-Deca XP+ ion trap (IT) and LTQ linear IT MS/MS instruments by Lab 34. The analyses were of two different specimens from the BD B1 set, using similar depletion, protein array pixelation prefractionation, and tryptic peptide fractionation (Tang *et al.* [14] this issue). LCQ analysis of B1-heparin-plasma yielded 575 IDs, while LTQ analysis of B1-serum yielded 2890 protein IDs, both with the PPP high-stringency SEQUEST filters. Many low abundance proteins in the low ng/mL to pg/mL range were identified. The comparison is complicated, however, by the fact that the protein identifications used different amounts of starting material. Depletion was applied to 193 µL (14.5 mg) of plasma and 415 µL (35.3 mg) of serum. After the fractionation steps, fractions

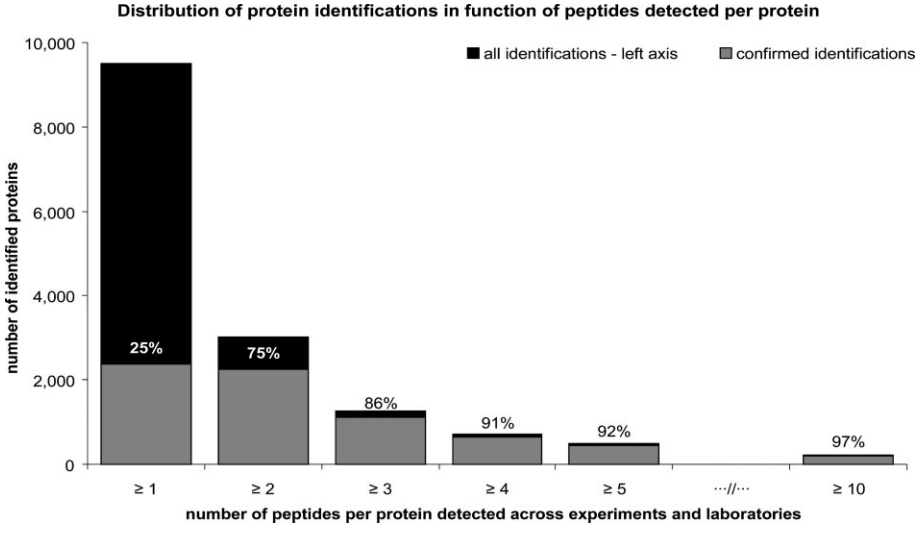

Figure 2. Number of proteins identified as a function of number of peptides matched.

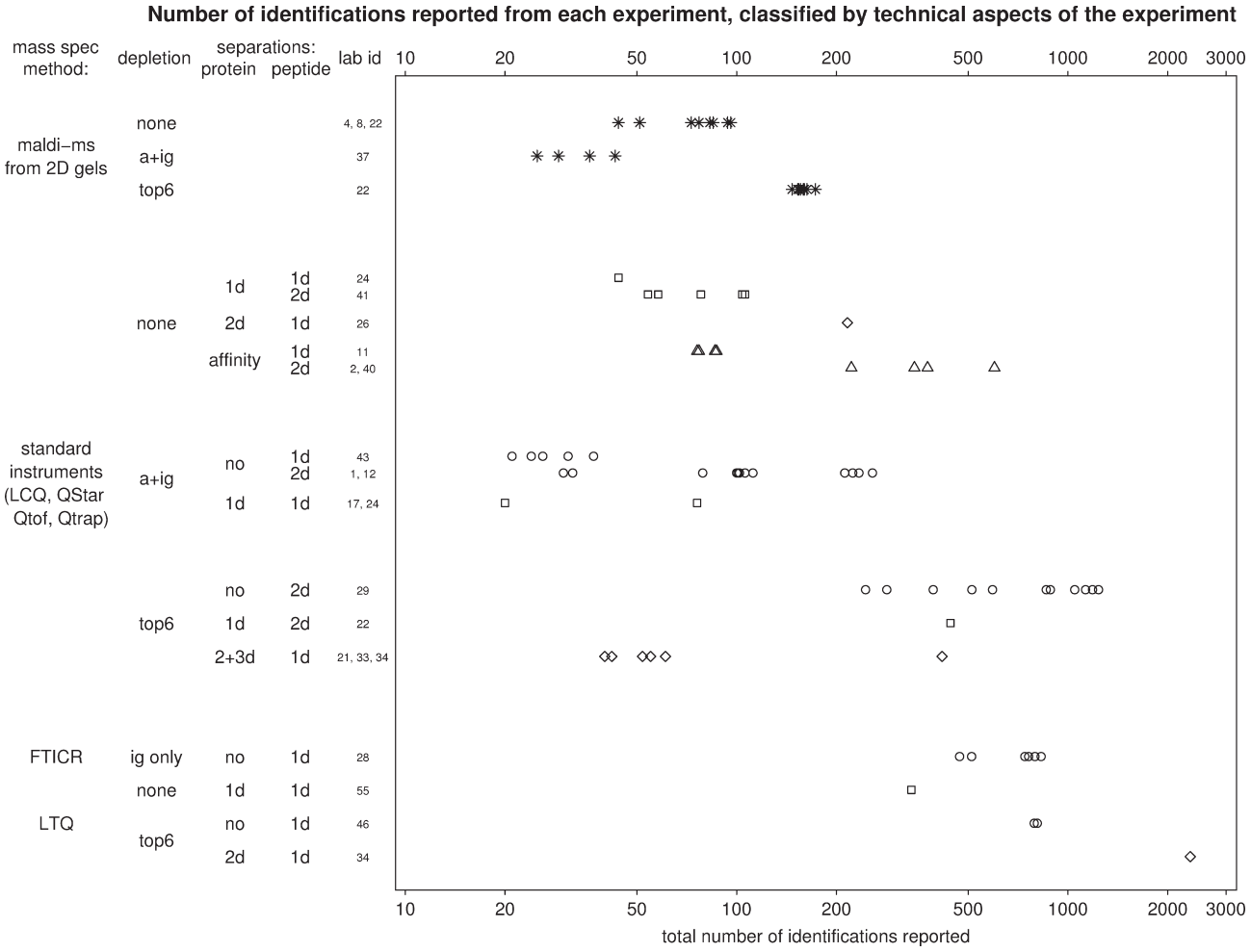

Figure 3. Categorization of depletion, fractionation, and MS methods and yield of proteins identified (log scale).

equivalent to 0.6 μL (45 μg) of the plasma and 2.4 μL (204 μg) of the serum were analyzed in the LC-MS/MS. Thus, some or possibly most of the difference in yield may be attributable to a larger volume analyzed. There were some other differences, as well including use of protease inhibitors with the depletion buffer, higher DTT concentration, fewer MicroSol-IEF fractions, and data-dependent MS/MS scans of the three most abundant ions with the LCQ instead of ten ions in the LTQ B1-serum experiment. There were also some differences in the searching of databases with one (serum) *versus* two (plasma) missed cleavage sites permitted. Tang *et al.* [14] describe extensive sensitivity analyses of experimental parameters that affect the tradeoff between numbers of high confidence protein IDs and analysis time. For example, gas phase fractionation to analyze different segments of the *m/z* range in each run was judged to be inefficient.

Labs 46 and 55 also employed LTQ instruments and obtained large numbers of identifications for reference specimens C1-serum and B1-citrate-plasma, respectively (Tables 1 and 2, Fig. 3).

3.2 Analysis of confidence of protein identifications

High false-positive rates are acknowledged to be a major problem in protein identification. Estimates can be generated, at least in relatively homogeneous datasets, by probabilistic methods using PeptideProphet and ProteinProphet, by matching to reversed-sequence databases [15–20]. The alternative of careful manual inspection of the spectra becomes a huge task and is subjective. The spectrum may represent a mixture of different peptides with almost equal parent masses and elution times. The biological specimen may have allelic variants or a contaminant not recorded in the database. Even if the sequence is correct, PTMs may take the sequence outside the scope of the match. However, true positives may be a problem, too, especially when the database sequence is simply not the same as that of the biological specimen analyzed.

To estimate the confidence of protein identifications across our heterogeneous database, we compared the observed data on number of peptide matches *per* identifica-

tion to a model in which identifications are randomly distributed. False-positive and true positive peptide identifications should show opposite behavior when numbers of identifications become large. We expect false-positive IDs to accumulate roughly proportional to the total, so that the chance of two or more false-positive identifications coinciding on the same database entry should be the product of their random probabilities. In contrast, a protein which is present in detectable concentration will produce many tryptic peptides in nearly stoichiometric quantities. Increased sampling, therefore, should increase the number of distinct peptides mapping to the same (correct) database entry. This model results in a Poisson distribution of number of peptides matched *per* sequence. Two parameters are needed to specify the model, the total number of proteins (Ndb) and the expected proportion of false peptide matches *per* database entry (lambda, ranging in this case from 0.211 to 0.146). The IPI 2.21 database contains 49 924 sequences after adjustment for redundancy. The upper bound for lambda corresponds to the assumption that every identified protein has at least one false-positive matching peptide; this bound eliminates all single-peptide hits. The lower bound accepts as correct all 1956 protein identifications based on a high confidence single peptide report, but treats all the 4528 lower confidence single peptide identifications as false. Throughout this range of values of lambda, proteins with four or more supporting peptides are predicted to be correct with better than 0.99 confidence; with exactly three peptides, 0.95–0.98; and with exactly two peptides 0.70 to 0.85 (Fig. 4). We based our annotations on the 3020 identifications made with two or more peptides project-wide to avoid a bias toward highly abundant proteins, if we had limited annotation to proteins based on three or more peptides. Furthermore, a substantial majority of protein IDs based on exactly two peptides is probably correct. Independent conclusions from manual review of a large number of spectra led one of our investigators to estimate at least 20% of one-peptide hits appear to be true positives. In addition, MacCoss *et al.* [21] concluded that the chance that multi-peptide proteins are false-positives declines exponentially with the number of peptides identified.

3.3 Quantitation of protein concentrations

A critical parameter for detection and identification of proteins is the abundance or concentration of the protein and its isoforms. We generated a calibration curve for a set of sentinel proteins for which quantitative immunoassays were available. Four different immunoassay and antibody microarray methods were performed by four independent laboratories (DadeBehring, Genomics Institute of Novartis Foundation, Molecular Staging, and Van Andel Research Institute). A total of 323 assays measured 237 unique analytes (Haab *et al.* [22]). In the cases of multiple assays, we cannot be certain that the same epitopes were targeted. This approach permits assessment of systematic variation in con-

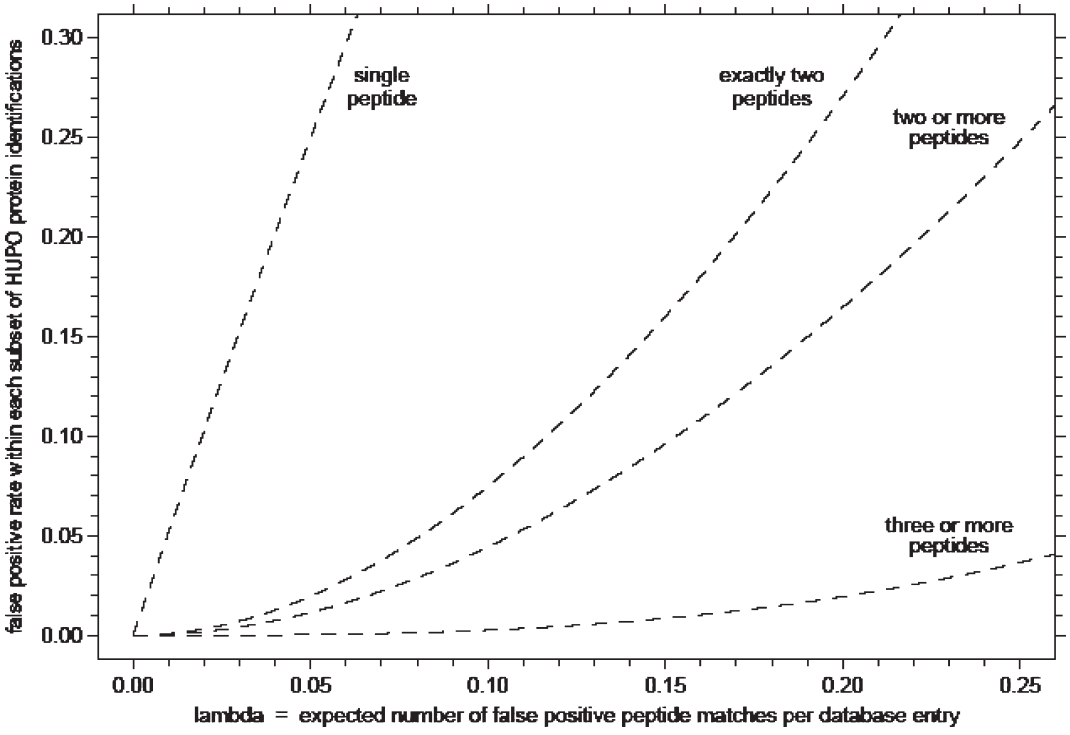

Figure 4. Plot of estimated error rate for subsets of PPP proteins based on one, two, or three or more peptides, Poisson model.

centration of proteins associated with blood preparation methods (serum and the three anticoagulation methods for plasma in each specimen set) and, after matching to IPI identifiers, facilitates an analysis of dependence on concentration for MS-based protein identifications using the HUPO PPP specimens. Some proteins were at such low concentrations that they were even undetectable with immunoassay or microarray methods. After extensive curation, we matched 76 IPI proteins among the 9504 dataset (based on one or more peptides) and 49 proteins among the 3020 protein dataset (based on two or more peptides) to quantitative analytes. Figure 1 in Haab *et al.* [22] shows four parameters used to determine the sensitivity of detection of these proteins as a function of immunoreactive concentration: number of labs reporting that protein, number of peptides on which protein IDs were based, percent coverage of the protein sequence, and score. The correlation coefficient for the total number of peptides matching that protein is r = 0.86 for the 3020 dataset and r = 0.90 for the 9504 protein dataset:

$$\log10(N) = 0.365*\log10(\text{conc}) - 0.711.$$

As expected, the most abundant proteins are the most readily detected, with essentially 100% agreement; with much less abundant proteins, only the laboratories with protocols and instruments capable of much more sensitive detection identified these proteins. Among the 49 proteins matched to the 3020 protein dataset, 12 are biologically interesting proteins identified with measured concentrations from 200 pg/mL to 20 ng/mL (Table 3).

Table 3. Least abundant proteins identified with two or more peptides (included in core dataset) with measured concentrations in the range of 200 to 20 000 pg/mL serum or plasma

Protein	Concentraion (pg/mL)
Alpha fetoprotein	2.9E+02
TNF-R-8	3.3E+02
TNF-ligand-6	1.5E+03
PDGF-R alpha	4.6E+03
Leukemia inhibitory factor receptor	5.0E+03
MMP-2/gelatinase	8.8E+03
EGFR	1.1E+04
TIMP-1	1.4E+04
IGFBP-2	1.5E+04
Activated leukocyte adhesion mol	1.6E+04
Selectin L	1.7E+04

4 Comparing the specimens

4.1 Choice of specimen and collection and handling variables

Pre-analytical variables can alter the analysis of blood-derived samples. Publications and protocols are generally deficient in this regard. Besides preparing the reference specimens of serum and plasma for direct comparisons, we undertook special studies on choice of sample type, stability during storage, use of protease inhibitors, and criteria for clinical standardization. The Specimens Committee concluded (Rai *et al.* [23]) that plasma is preferable to serum, due to less degradation *ex vivo* (as shown specifically by Tammen *et al.* [24] and Misek *et al.* [25]). Nevertheless, there is a view that standardization of proteomics assays with serum may be desirable, since archived specimens are so frequently sera.

They concluded that platelet-depletion of plasma may be desirable to avoid platelet activation with release of proteins, especially if there is a 4C step in the preparation. BD explained that 4C was chosen for centrifugation and holding of the tubes prior to aliquoting to aid in stabilizing labile biomarkers. For investigators concerned about platelet contamination, options include filtration of the plasma through a 0.2 μ low protein binding filter; double centrifugation of the specimen; and use of additives that minimize platelet activation, such as CTAD, a mixture of citrate, theophylline, adenosine, and dipyridamole. Samples should be aliquoted and stored frozen with minimization of thaw/re-freeze cycles, preferably in liquid nitrogen, though −80°C seems to be very nearly as good. Protease inhibitors would be desirable, but present cocktails introduce complications due to peptide inhibitors that may interfere in the MS and small molecule inhibitors that form covalent bonds with proteins, shifting the isoform pattern. The Committee recommends diligent tracking of pre-analytical variables, and development and use of certified reference materials for quality control and quality assurance.

Haab *et al.* [22] extensively analyzed the concentrations of assayable proteins in the PPP specimen sets. They noted a systematic 15% lower value for many proteins in citrate-plasma, compared with other specimens; it turns out that this can be attributed to dilution and osmotic effect with the citrate solution, without any impairment in detection of proteins compared with the other specimens. However, David Warunek and Bruce Haywood of BD advised us that results with citrate-anticoagulated plasma can be quite sensitive to the blood:additive ratio and the subject's hematocrit. EDTA, meanwhile, is a much better chelator of calcium and more effective at platelet inactivation.

The sets of four specimens from a given donor pool yielded rather similar numbers of proteins when analyzed by the same lab and same techniques (see Table 2). Naturally, the agreement on identification of specific proteins was greater for higher abundance proteins. Since the laboratories exercised considerable discretion in deciding how many and which of the reference specimens to request and how many to actually analyze, as well as how intensively to analyze them, comparisons across the specimen results is of limited validity in this exploratory phase of the PPP. However, comparisons within several laboratories (1, 2, 11, 12, 28, 29, 41, 43 in Table 2) show quite close values for numbers of proteins identified, with deficiencies for B1-serum and B3-serum in Lab 1, and B1-heparin and possibly B3-heparin in

Lab 29). It is curious that several laboratories chose citrate-plasma if they analyzed only one plasma specimen (Tables 1 and 2). Lab 28 shows greater similarity within each of the three donor pools for three citrate-plasma *versus* serum comparisons, than for citrate-plasma or sera across the three pools. The values for total number of proteins within each pair were quite close, whereas the B1 specimens yielded significantly fewer identifications than the B2 and B3 pairs. For B2 serum and B2 citrate plasma, they reported 365 proteins in common, of 542 and 572 identified in each. Ion current estimation of concentrations put 275 of the 365 within ±2-fold; 59 proteins had plasma/serum values >2X and 31 had P/S values <0.5X (Adkins *et al.* [12], this issue). Lab 34 is a special case, because different instruments were used for B1-heparin (LCQ) and B1-serum (LTQ), as noted above (Section 3.1).

Table 2 summarizes the protein IDs by lab and specimen. As noted above, the numbers of proteins identified in the consolidated database may be different from those in the individual papers in this special issue due to the integration procedure applied to the Core Dataset and the expanded analyses for these papers. The most analyzed specimen, B1-serum (Caucasian American) had 1749 IDs among the 3020. The three anticoagulated B1 specimens yielded a total of 1904 unduplicated IDs, of which 1023 were in common with the proteins identified in the B1-serum. The total number of unique IDs in the four B1 specimens that meet the two or more peptides criterion in either plasma or serum is 2630. A similar analysis of the combined C1 (Chinese) pooled specimens in just two labs yielded 1693 proteins, of which 1416 were identified in the B1 pool. With the exception of Labs 26 and 28, no very extensive analyses of the B2/B3 African-American and Asian American specimens were submitted. Combining all datasets, including the lyophilized NIBSC citrate-plasma specimen, we reached the 3020 protein dataset.

Tammen *et al.* [24] focused on the "peptidome" with mass <15 kDa. Peptides may be fragments of higher M_r proteins, or hormones, growth factors, and cytokines with specific biological functions. Their findings are not included in the Core Dataset since they used differential peptide display, plotting m/z ratios against retention time, with RP-HPLC-MALDI-TOF-MS. They do use nESQ-qTOF-MS/MS or MALDI-Tof-Tof-MS to confirm some peptide identifications. They did not actually attempt to identify proteins from the peptides. However, they made observations highly relevant to specimen processing. A large number of peptides, including many abundant peptides, are present only in serum, presumably due to the multi-protease events of clotting (AP-FXIII), enzyme activities (kallikrein), or peptides derived from cellular components, especially platelets, or the clot itself (thymosin beta-4, zyxin). In fact, at least 40% of the peptides detected in serum were serum-specific. Clotting is unpredictable due to influences of temperature, time, and medications, which are hard to standardize. These observations with serum may be highly relevant to the interpretation of SELDI results. They reported altered elution behavior of peptides in the presence of heparin, due to the polyanion nature of polydisperse low M_r heparin. Heparin acts through activation of antithrombin III, while citrate and EDTA inhibit coagulation and other enzymatic processes by chelate formation with ion-dependent enzymes. They recommend platelet-depleted EDTA or citrate-plasma, which gave consistent and similar results. They do not recommend addition of protease inhibitors, especially aprotinin, which requires μg/mL concentrations that interfere with analysis.

4.2 Depletion of abundant proteins followed by fractionation of intact proteins

Reducing the complexity of protein mixtures by depletion and fractionation of intact proteins greatly simplifies the task for MS/MS analysis. There are essentially three patterns of depletion in Table 2 and Fig. 3: no depletion of the most abundant proteins, depletion only of albumin or Ig or both, and depletion of the top-6 proteins, which are albumin, IgG, IgA, haptoglobin, alpha-1 anti-trypsin, and transferrin (Agilent column). There is clear evidence from the main database and from a series of special project studies by PPP investigators that depletion makes it significantly more feasible to visualize, detect, and then identify lower abundance proteins (Echan *et al.* [26], Li *et al.* [5], Zolotarjova *et al.* [27], Huang *et al.* [28], Tang *et al.* [14], Misek *et al.* [25], Yang *et al.* [29], Barnea *et al.* [30], Moritz *et al.* [31], Cho *et al.* [32], Kim *et al.* [33]). However, when only 2-DE is employed, the many "new" spots detected after depletion are unmasked isoforms of medium-abundance proteins, rather than lower abundance proteins [5, 26]. There is a counterbalancing problem, namely non-target or inadvertent removal of other proteins [6], which could be due to peptides and proteins bound to the target proteins, especially albumin; cross-reactivity with the bound antibodies; or non-specific binding to the column or resin or dye. Details of the protocols, proprietary buffers, column capacity, and previous use of the columns may be important variables. With older and much less expensive albumin-removal agents, such as Cibacron Blue dye, there is thought to be binding to the dye (as well as any binding to the albumin).

Moritz *et al.* [31] provide a preliminary report using free-flow electrophoresis (FFE-IEF) and rapid (6 min) RP-HPLC to fractionate citrate-plasma (Lab 33). They analyzed both bound and flow-through fractions from immunoaffinity depletion of the top-6 proteins. From 15 of 96 FFE fractions, with 72 780 MS/MS spectra analyzed with MASCOT and Digger and subjected to manual validation, they obtained 55 proteins based on two or more peptides and 23 more based on one peptide, across a mass range of from 4 to 190 kDa; these included several with estimated concentrations of 0.5–1 ng/mL. They highlight the identification in the bound fraction of a 35-amino acid serine protease protection peptide (CRISPP) that is cleaved from the *C*-terminus of alpha-1 anti-trypsin, non-covalently complexed with alpha-1 anti-trypsin, and not included in the IPI 2.21 database.

They detected protein complexes by using non-denaturing, non-reducing buffers. They enhanced their yield by building a data-dependent exclusion list to prevent re-identifying abundant peptides.

Tang *et al.* [14] investigated many experimental parameters of depletion, fractionation, and such MS variables as gas phase fractionation. They combined solution isoelectrofocusing and 1-D SDS gel electrophoresis to generate "pixels" of proteins with defined p*I* and M_r ranges, then fractionated tryptic digests with 2-D LC, followed by LCQ-Deca-XP+ or LTQ-linear IT-MS/MS for B1-heparin-plasma and B1-serum reference specimens, respectively. These methods yielded 575 and 2890 high-confidence protein identifications (see Section 3.1) using the stringent HUPO PPP SEQUEST parameters; they did not remove potential homologous database entries; 319 of the 575 plasma proteins were identified in the serum specimen. Of these 319, half are single-peptide proteins in plasma, but many more are multiple-peptide proteins in serum, with the LTQ instrument, and have rich MS/MS fragmentation patterns. They estimated that proteins in the low ng/mL range were detected from 45 μg of plasma protein using the LCQ-Deca XP+, whereas proteins in the low pg/mL range were detected from 204 μg of serum using the LTQ. They uniquely utilized a SEQUEST Sf score, which combines X_{corr}, deltaCn, Sp, R_{sp}, and ions scores using a neural network to reflect the strength of peptide assignment on a scale of 0 to 1; scores $\geq$0.7 were considered to have a high probability of being correct, regardless of other parameters; when Sf scores replaced $R_{sp} \geq 4$, they obtained 744 and 4377 non-redundant protein identifications from the plasma and serum specimens, respectively.

Misek *et al.* [25] identified many isoforms and compared relative abundance of proteins in serum, EDTA-plasma, and citrate-plasma labeled, respectively, with the fluorescent dyes Cy3, Cy5, and Cy2 after top-6 immunoaffinity depletion. The three labeled, depleted samples were subjected to three-dimensional protein fractionation by p*I*, hydrophobicity, and M_r. About 3000 bands on 1-D SDS gels with $\pm>$ two-fold differences in intensity of fluorescence in dye pairs were excised and analyzed by MS/MS, yielding a total of only 82 non-redundant proteins; 28 proteins were identified in ten or more different fractions. Complement C3 and clusterin are presented as examples of proteins whose biologically significant cleavage products can be identified with this method. Not surprisingly, the yield in MS/MS was greater for proteins with higher intensity (abundance). Multiple isoforms reduce the concentration of a protein in any particular spot or fraction and may react very differently with antibodies used to quantify the proteins or detect the proteins, as on microarrays.

Subfractionation of the complex mixtures that are plasma and serum can be performed chemically or with capture agents. A very good example is the glycoprotein subproteome. Labs 2 and 11 (Tables 1 and 2) utilized hydrazide chemistry and binding with three lectins, respectively, to enrich for glycoproteins. The chemical method, which captures *N*-linked glycoproteins subsequently treated with PNGase F, was published by Zhang *et al.* in 2003 [34]. Yang *et al.* [29] used wheat germ agglutinin, Jacalin lectin, and Con A together on agarose to isolate and characterize approximately 150 glycoproteins in PPP serum and plasma reference specimens after analysis by LCQ-MS/MS, with confirmation in some cases using a linear IT LTQ instrument. There was close similarity for the composition of the glycoproteome across the plasma and serum specimen sets, except for fibrinogen, which was absent from serum (after clotting). Samples from the individuals from three different ethnic groups showed only a few individual differences. Together the two laboratories identified 254 glycoproteins, of which 164 were identified by other laboratories in this collaboration. That means that 90 were found only in the glycoprotein-enriched studies. Glycoprotein has an important incidental benefit in that the non-glycosylated albumin protein should be excluded; in fact, some albumin remains, given its very high abundance and its tendency to bind glycoproteins.

Cho *et al.* [32] combined immunoaffinity depletion of the top-6 proteins with free-flow electrophoresis or 2-DE of fractions, and MALD-TOF-MS PMF; they found only minor differences across the donor and specimen preparation variables. With 2-DE they found few non-target proteins in the immunoaffinity bound fraction.

Kim *et al.* [33] sought to identify and eliminate false-positive peptide identifications and subsequent protein matches by analyzing molecular weight on 1-D SDS gels after immunoaffinity depletion. Of 494 proteins identified with 2-D-LC/ESI-MS/MS of 28 1-D fractions, using SEQUEST with stringent PPP filters, 202 were excluded as single-peptide hits as well as estimated M_r too deviating from theoretical M_r, but 166 one-peptide matches were retained based on good M_r match. This approach requires careful curation for biologically cleaved proteins. Their method actually increased the number of accepted proteins, since only 128 (26% of 494) were based on two or more peptides among the total of 292 protein identifications claimed for the B1-serum specimen.

Echan *et al.* [26] compared the immunoafffinity top-6 depletion column and corresponding spin cartridge from Agilent with a prototype ProteoPrep dual anti-albumin/anti-IgG antibody column from Sigma Aldrich, with five commercially available kits using Cibacron Blue for albumin and/or Protein A or G for immunoglobulin depletion, and with no depletion. These variables correspond to the categories depicted in Figure 3. The polyclonal antibody column gave nearly complete depletion, showed low non-specific binding, based on 2-DE profiles, and permitted many new spots to be visualized. However, the number of new proteins was quite small, due to the emergence of newly visualized spots representing numerous isoforms of the now-most abundant remaining proteins. They estimated that silver staining on 2-D gels should have been able to detect proteins originally present in the serum or plasma at 40 ng/mL or

higher, while the protein identified with lowest known concentration is at about 30 μg/mL, before accounting for heterogeneity of isoforms. The two-protein column had more capacity for albumin and IgG removal, but also removed many non-target proteins, which may be improved with optimized buffers. Apparently, buffer variables are very influential with all of the antibody columns. Given published reports of up to 63 proteins bound to albumin [35], secondary binding conditions can introduce major variability in results. Clearly, more potent technology combinations are required to adequately evaluate the non-target binding of proteins during immunoaffinity depletion, as well as to reach down to the ng/mL to pg/mL concentration range. Echan *et al.* [26] point out that the inexpensive and convenient dye and protein A/G methods can be used for fractionation rather than depletion. They also note the potential to specifically deplete many more proteins with expanded immunoaffinity columns.

Additional papers by Zolotarjova *et al.* [27] and by Huang *et al.* [28], scientists at Agilent and at GenWay Biotech, respectively, present laboratory results with their immunoaffinity products. The polyclonal rabbit antibody column from Agilent removes albumin, IgG, IgA, haptoglobin, transferrin, and alpha-1 anti-trypsin. The polyclonal chicken IgY antibodies on microbeads from GenWay remove six (albumin, IgG, IgA, IgM, transferrin, and fibrinongen) or 12 (also alpha-1 anti-trypsin, alpha-2 macroglobulin, haptoglobin, apolipoproteins A-I and A-II, and orosomucoid/alpha-1 acid glycoprotein). Both groups report highly effective removal and little to no non-target binding. These products were introduced during the conduct of the PPP pilot phase and were made available to investigators.

One way to maximize identifications is to analyze bound fractions as well as pass-through fractions, as done by He *et al.* [6] and by Labs 29 and 46 (Tables 1 and 2). He *et al.* [6] report large numbers of proteins in the top-6 immunoaffinity bound fraction when extensive LTQ-MS/MS is applied, utilizing the stringent PPP SEQUEST filters. They may not have used the full system optimized by the column manufacturer.

4.3 Comparing technology platforms

Li *et al.* [5] analyzed the PPP C1-serum specimen with five different proteomics technology combinations after immunoaffinity depletion of the top-6 proteins. In all, 560 unique proteins were identified, 165 with two or more peptides. Only 32 proteins were identified by all five approaches and 37 by 2-DE, 2-D HPLC, and shotgun approaches, primarily due to finding only 78 unique proteins among 1128 spots excised, digested, and analyzed with method 1, WAX-2-DE-MALDI-TOF-MS-MS. Protein 2-D-HPLC fractionation + RP-HPLC/microESI-MS-MS gave 179 proteins; an online SCX shotgun strategy ("bottom-up") gave 131, an offline SCX shotgun strategy gave 224, and an offline shotgun-nanospray strategy yielded 330 proteins. High and medium abundance proteins are found by all methods, while low abundance proteins are complementary, reflecting both different methods and inherent incompleteness of sampling and identifying peptide ions. Different technology combinations give different useful information; for example, the 2-DE method 1 provided more information about p*I*-altered isoforms and relative abundance of identified proteins. The offline strategies sharpen the peaks and improve separation of peptides, submit more fractions to the MS instrument, and allow the MS enough time to acquire the qualified spectra of more eluting peptides. Nanoflow accentuates the same advantages, permitting ultrahigh sensitivity. Overall, electrophoresis and chromatography, coupled respectively with MALDI-TOF/TOF-MS and ESI-MS/MS, identified complementary sets of serum proteins. Like Aebersold and Mann [2], they conclude that no single analytical approach will identify all the major proteins in any proteome. Others have recently used similar 2-D separation of peptides offline, intact protein fractionation prior to MS, or sensitive ESI-MS/MS analysis of fractionated peptides [36–39]. As far as cost-effectiveness, the 2-D HPLC approach required much more time and labor and was much less suited to automation than the other strategies; it has the advantage of being able to process large volumes of sample, when that is available and desired. Handling fractions also introduces more evidence of contamination; epidermal keratins are seldom found with the shotgun methods. Low abundance proteins are not only masked by medium abundance proteins on gels, but inefficient extraction of peptides from gels is a limitation for low abundance proteins.

Barnea *et al.* [30] expanded on their original submission as Lab 1 (Tables 1 and 2) with an analysis of several protein fractionation and several MS/MS methods on PPP reference specimen B2-serum. Albumin and IgG were depleted with are Bio-Rad mini-kit based on Affi-Gel Blue and Affi-Gel protein A, respectively. The aim was to increase the concentrations of individual proteins and then their tryptic peptides in each fraction submitted for MS/MS analysis, seeking to reach the threshold for detection. Combining pre-proteolysis fractionation with post-digestion fractionation was more effective than more extensive fractionation of the peptides. Each method has some advantages of avoiding loss of proteins with particular characteristics (p*I*, M_r, other). The base case was MudPIT analysis of unfractionated, digested proteins; then SDS-PAGE, SCX, and Rotofor fractionations were coupled with LC-MS/MS or with MudPIT. In each pair, MudPIT gave more protein IDs than LC-MS/MS. SCX gave the most IDs among the fractionation methods.

He *et al.* [6] analyzed ten pooled male and ten pooled female C1-sera, using top-6 depletion, tryptic digestion, then RP-HPLC, ESI-MS/MS shotgun analysis. They reported 944 non-redundant proteins under stringent PPP criteria based on [40], combining separate analyses of male (594) and female (622) sera; there were 206 with two or more peptides. Some lower abundance proteins were detected, including complement C5 and CA125. Instead of one analysis of serum, here there are eight analyses: male and female,

bound and unbound, and a duplicate of each. The reproducibility of the duplicates is 40–50%; the overlap of bound and unbound is 16–18%, and of male and female 40–50% (*i.e.*, same as duplicates). They used four databases: IPI 2.20 (June 2003), IPI 2.32 (May 2004), Swiss-Prot 43 (March 2004), and NCBI (Dec 2003) and obtained quite similar protein groups for the first three and also for NCBI, though the pre-grouping numbers of proteins were 2.5 times larger for NCBI, demonstrating the known redundancy in the NCBI database.

4.4 Alternative search algorithms for peptide and protein identification

One of the important challenges for collaborative proteomic studies is the variety of search algorithms embedded in mass spectrometers. Some of these search algorithms are proprietary with key elements undescribed in the open literature or even for the user laboratory. Each investigator has many options in the choice of parameters for the software search to identify peptides from the mass spectra of ion fragments and then to deduce the best protein match from yet another broad array of gene and protein databases, including different versions of each evolving database. Expert curation of such collaborative datasets is required. In the PPP Jamboree Workshop of June 2004, the offer to generate cross-algorithm analyses with PPP data was strongly endorsed, and many months of effort were invested.

Kapp *et al.* [11] report a unique analysis of alternative search algorithms. They used one raw file from the Pacific Northwest National Laboratory LCQ-MS/MS data on serum depleted only of IgG published by Adkins *et al.* [41], which served as a basis for the later FT-ICR-MS analyses for the PPP (Lab 28). The same spectra were subjected to analyses with MASCOT, SEQUEST (with and without PeptideProphet), Sonar, Spectrum Mill, and X!Tandem by experts familiar with the use of each. Careful manual inspection was applied, as well, though it is always a challenge to understand what exactly were the criteria used in manual inspection. The paper provides a useful description and categorization of the features of each search engine into heuristic algorithms and probabilistic algorithms. The authors then present and compare their performance identifying peptides and proteins, benchmarking them based on a range of specified false-positive rates. In all, 600 peptides were identified, of which 355 were found with very high confidence (estimated error rate 1%) by all four of MASCOT, SEQUEST, Spectrum Mill, and X!Tandem. The authors concluded that no one of these algorithms outperforms the rest. Spectrum Mill and SEQUEST performed well in terms of sensitivity, but performed less well than MASCOT, X!Tandem, and Sonar in terms of specificity. Thus, they recommend using at least two search engines for consensus scoring, though the scheme for creating combined scores awaits further work. The probabilistic algorithm, MASCOT, correctly identified the most peptides, while the re-scoring algorithm, PeptideProphet, enhanced the overall performance of SEQUEST. This paper utilizes reversed-sequence searches, as well as probabilistic estimates of false-positive rates. Unfortunately, the spectra in this dataset were dominated by high abundance proteins, such that the 600 peptides were matched to only 40–60 proteins using a trypsin-constrained search.

4.5 Independent analyses of raw spectra or peaklists

After the original data submission protocol had been established, built upon peptide sequences and protein identifications, three groups emerged as having capability for centralized, independent analyses that would bypass the peculiarities of the search engine software embedded in particular MS instruments and the criteria applied by individual investigators in establishing thresholds for high and lower confidence identifications or applying manual inspection of the spectra.

Beer at IBM/Haifa developed PepMiner software [42], which processes very large numbers of raw spectra to generate clusters of spectra and then SEQUEST-like analysis and scoring for peptide and protein IDs. Beer *et al.* [43] applied this method to the spectra from laboratories 1, 2, 17, 22, 28, 29, 34, and 40. The data from laboratory 1 included those submitted for the Core Dataset(s) as well as those in the Barnea *et al.* [39] special project paper. They identified 14 296 peptides, which were assigned to 4985 proteins with one or more peptides, 2895 proteins with two or more peptides, and 1646 with three or more peptides. The 4985 IDs had 2245 in common with the 15 519 unintegrated and 1983 in common with the 9504 integrated PPP IDs. The 2895 based on two or more peptides compares with our 2868 based on two or more peptides for the same eight laboratories, with 865 in common with our Core Dataset.

Deutsch *et al.* [44], at the Institute for Systems Biology in Seattle, US, utilized SEQUEST with PeptideProphet/ProteinProphet software developed by the Eng group to estimate error rates and probability of correct assignment of spectra to peptide sequences and then to protein IDs [15, 45]. Analyzing the PPP datasets from laboratories 2, 12, 22, 28, 29, 34 (B1-heparin only), 37, and 40 with the PeptideAtlas process [46], they observed 6929 distinct peptides with a probability score $\geq$0.90, including 6342 which mapped to 1606 different EnsEMBL proteins and 1131 different EnsEMBL genes. Reduction of multiple mappings yielded 960 different proteins, of which 479 have matches in the PPP 3020.

Kapp *et al.* [11] at the Ludwig Institute in Melbourne are utilizing MASCOT and Digger software developed at Ludwig on submissions from 14 laboratories; incomplete analyses show more than 500 high-confidence, non-redundant proteins with trypsin-constrained searches.

In addition, Beavis at the Manitoba Centre for Proteomics created a dataset with 16 191 EnsEMBL proteins from the PPP raw spectra using X!Tandem [47], of which 9497 matched to IPI v2.21, 3903 to our unintegrated list, and 2828 to our 9504 proteins based on one or more peptides. Of

5816 IPI proteins with two or more peptides, 1259 matched to the 5102 unintegrated and 913 to the 3020 Core Dataset.

Martens *et al.* [48] noted the value of these independent analyses in overcoming numerous sources of variation from the search algorithm, the database, and the investigator. They recommend that *m/z* peaklists routinely be made publicly available, while deferring on the raw data, which currently lack standardized formats, let alone the required infrastructure for centralized storage and distribution. However, a plan to assure access to the raw spectra, as well as the peaklists, can facilitate wide dissemination and utilization of complex datasets, as we have demonstrated in this collaboration by both the participating laboratories and the independent analysts, the incorporation into PRIDE by EBI, into PeptideAtlas by ISB and ETH, and into the Global Proteome Machine DataBase by Beavis.

It is striking that these independent analyses not only differed in the proteins that they identified, but also in the peptides identified from the same MS/MS spectra that were the basis for the protein matches. Further improvements in software and analytical methods are needed, given the many sources of error in peptide identification [49]; automated *de novo* sequencing can help, and chemical synthesis of peptides to determine the spectra directly can be employed selectively.

4.6 Comparisons with published reports

Table 4 shows the numbers of proteins reported in human plasma or serum in the literature, the number of those proteins in the IPI database, and the congruence with our PPP 9504 and PPP 3020 protein lists. Our lists are integrated (see Section 3.1), while the others generally are not, and do not use the same methods. It is clear that the number and nature of proteins identified in serum and plasma depend greatly on the sample preparation and fractionation and on MS methods and analytical tools.

Table 4. Comparison of PPP integrated protein identification lists with published datasets for human plasma or serum

Published data	Total IDs	# IPI proteins	PPP_9504 dataset	PPP_3020 dataset
Anderson *et al.* [50]	1175	990	471	316
Shen *et al.* [38]	1682	1842	526	213
Chan *et al.* [54]	1444	1019	402	257
Zhou *et al.* [35]	210	107	68	51
Rose *et al.* [55]	405	287	159	142

Anderson *et al.* [50] published a compilation of 1175 nonredundant proteins reported in at least one of four sources (literature review plus three recent experimental datasets [51, 41, 52]); only 46 proteins were reported in all four sources, suggesting high false-positive rates from reliance on single-peptide hits [49]. The experimental papers used multidimensional chromatography, 2-DE, and MS; MudPIT analysis of a tryptic digest; or MudPIT of a tryptic digest of low-M_r plasma fractions. Of the 990 of these proteins which have IPI (version 2.21) identifiers, 316 are found in our 3020 protein Core Dataset. When we relaxed the integration requirement (5102 IPI IDs), as was the case for [50], this figure rose only to 356 matches. Using the full 9504 dataset, the corresponding matches were 471 with integration and 539 without integration (15 710 protein IPI IDs).

Shen *et al.* [38] used high-efficiency nanoscale RP LC and strong cation exchange LC in conjunction with ion-trap MS/MS and then applied conservative SEQUEST peptide identification criteria (with or without considering chymotryptic or elastic peptides) and peptide LC normalized elution time constraints. Between 800 and 1682 human proteins were identified, depending on the criteria used for identification, from a total of 365 μg of human plasma. With their cooperation, we re-ran their raw spectra using HUPO PPP SEQUEST parameters (high confidence: $X_{corr} \geq 1.9/2.2/3.75$ (for charges $+1/+2/+3$), deltaCn ≥ 0.1, and $R_{sp} \geq 4$; lower confidence: $X_{corr} \geq 1.5/2.0/2.5$ (for charges $+1/+2/+3$), deltaCn ≥ 0.1) and obtained 1842 IPI protein matches. Of these, 526 and 213 were found in the PPP 9504 and 3020 datasets, respectively.

Chan *et al.* [53] resolved trypsin-digested serum proteins into 20 fractions by ampholyte-free liquid phase IEF. These 20 peptide fractions were submitted to strong cation-exchange chromatography, then microcapillary RP-LC-MS/MS. They identified 1444 unique proteins in serum. When we mapped these proteins against the IPI v2.21 database, there were 1019 distinct proteins. From this set, 402 and 257 proteins matched with the 9504 and 3020 datasets, respectively.

Zhou *et al.* [35] identified an aggregate of 210 low M_r proteins or peptides after multiple immunoprecipitation steps with antibodies against albumin, IgA, IgG, IgM, transferrin, and apolipoprotein, followed by RP-LC-MS/MS. Only 107 proteins were mapped with IPI identifiers, of which 68 and 51 were found in the 9504 and 3020 PPP protein lists, respectively.

Finally, Rose *et al.* [54] reported fractionation in an industrial-scale approach, starting with 2.5 liters of plasma from healthy males, depleted of albumin and IgG, then smaller proteins and polypeptides separated into 12 960 fractions by chromatographic techniques. From thousands of peptide identifications, 502 different proteins and polypeptides were matched, 405 of which were included in the publication. Of the 287 which mapped to IPI identifiers, 159 and 142 are included in our 9504 and 3020 protein dataset, respectively.

Thus, across studies, as well as across the PPP participating laboratories, incomplete sampling of proteins is a dominant feature. A substantial depth of analysis is achieved with depletion of highly abundant proteins, fractionation of intact proteins followed by digestion and two or more MS/MS runs for each fraction. Standardized, statistically sound

criteria for peptide identification and protein matching, and estimation of error rates are necessary features for comprehensive profiling studies.

4.7 Direct MS (SELDI) analyses

Ten laboratories requested PPP specimens for analyses with SELDI chip fractionation, MS analysis, and algorithm-based differentiation of m/z peaks across specimens. Rai *et al.* [56] report the cross-laboratory evaluation of eight submitted datasets, of which five were judged appropriate for comparison of plasma results and four for serum results. Intra-laboratory CV varied from 15 to 43%. Correlations across labs were 0.7 or higher for 37 of 42 spectra with signal/noise ratios >5. More detailed analyses were done to actually identify one protein, haptoglobin, and variation in the intensity/concentration of its subunits in the different PPP reference specimens. They recommend stringent standardization and pre-fractionation to increase the usefulness of this method.

4.8 Annotation of the HUPO PPP core dataset(s)

From the inception, HUPO has intended that the Plasma Proteome Project facilitate extensive and innovative annotation of the human plasma and serum proteome. A large element of the Jamboree Workshop was focused on collaborative annotation. Several papers in this issue report on those collaborations.

Ping *et al.* [56] emphasize use of peptide identification results from MS/MS to reveal cleavage of signal peptides, proteolysis within hydrophobic stretches in transmembrane protein sites, and PTMs. Using 2446 of the 3020 PPP from IPI that matched to EnsEMBL gene products, they highlight subproteomes comprised of glycoproteins, low M_r proteins and peptides, DNA binding proteins, and coagulation pathway, cardiovascular, liver, inflammation, and mononuclear phagocyte proteins. Surprises include 216 proteins matched by Gene Ontology to DNA binding and 350 to the nucleus, including histone proteins, suggesting detection of proteins released by apoptosis or other means of cell degradation. Using the Novartis Atlas of mRNA expression profiles for 79 human issues, liver dominated as the source of the majority of proteins, although many of these proteins are also produced in other tissues. Many classic protein markers of leukocytes were not detected, including markers of B-cell, T-cell, granulocyte, platelet, and macrophage lineages, presumably all at low abundance with little shedding. In contrast, some quite low abundance proteins were found repeatedly, such as VCAM-1 and especially IL-6.

Signal peptide cleavage sites are generally predicted based on presence of a hydrophobic stretch of amino acids flanked at one end by basic amino acids. Seeking experimental evidence for such cleavage sites, these authors focused on semi-tryptic peptides, presuming that the signal cleavage event does not involve trypsin *in vivo*. Such evidence may override database predictions, as, apparently, in the cited example of SERPINA3/alpha-1-antichymotrypsin. They also identified two previously unreported proteins that undergo regulated intramembrane proteolysis, one of which releases an extracellular immunoglobulin domain - a reason not to reject all immunoglobulin matches. The MS/MS spectra can be examined for evidence of unrecognized PTMs. Using the Osprey tool, they found an average of nearly six protein-protein interactions *per* protein for a subset of 652 proteins; if they are circulating as multi-protein complexes, they will be less likely to be cleared through the kidney glomeruli.

Berhane *et al.* [57] focused on 345 proteins of particular interest for cardiovascular research. They classified the proteins into eight categories, most of which have relevance to other organ systems, as well: markers of inflammation in cardiovascular disease, vasoactive and coagulation proteins, signal transduction pathways, growth and differentiation-associated, cytoskeletal, transcription, channels and receptors, and heart failure and remodeling-related proteins. Of particular interest were the detection for the first time in plasma of the ryanodine receptor, part of the intracellular calcium channel in cardiac (and skeletal) muscle, and smoothelin, a structural protein restricted to smooth muscle cells, co-localized with actin. They used a number of identified peptides as an indicator of abundance of the protein (as in Section 3.3, above); for the first two categories, about 50% of proteins were identified with less than ten peptides, whereas no proteins among transcription factors had more than ten peptides and 56% had the minimum of two peptides. No cardiac contractile proteins were identified, even though they are far more abundant than transcription factors or signaling proteins in the heart, suggesting that necrotic cell death and uncontrolled cell rupture had no part in the appearance of any of the detected proteins in the healthy donors studied.

Muthusamy *et al.* [58] utilized a Java 2 Platform literature search tool to facilitate manual curation of functional classes of proteins, starting with the PPP set of 3020 IPI proteins (2446 genes). They subjected protein and nucleotide sequences in NCBI to BLAST queries to identify splice isoforms; they report that 51% of the genes encoded more than one protein isoform (a total of 4932 products). A total of 11 381 single nucleotide polymorphisms involving protein-coding regions were mapped onto protein sequences.

The Core Dataset of 3020 proteins was annotated with use of Gene Ontology for subcellular localization, molecular processes, and biological functions, showing very broad representation of cellular proteins. Subcellular component classification of the 1276 IPI-3020 proteins included in GO showed a relatively high proportion of proteins from membrane compartments (26%), nuclei (19%), cytoskeleton (11%), and other cell sites (23%), compared with the expected predominance of secreted proteins ("traditional plasma proteins") (14%). GO analyses of molecular processes showed 39% binding, 28% catalytic, 7% signal transducer, 6% transporter, 4% transcription regulator, and 3%

enzyme regulator. GO analyses of biological functions revealed 36% metabolism, 25% cell growth and maintenance, 5% immune response, 1% blood coagulation and 1% complement activation. Examination of specific Gene Ontology terms against a random sample of 3020 from the human genome (Supplementary Fig. 1) shows some proteins >3 SD from the expected line. Categories over-represented include extracellular, immune response, blood coagulation, lipid transport, complement activation, and regulation of blood pressure, as expected; on the other hand, surprisingly large numbers of cytoskeletal proteins, receptors and transporters also were identified.

An InterPro analysis similarly compared the 3020 protein dataset with the fine-grained protein families and domains described for the full IPI v2.21 56 530 human proteins dataset (Supplementary Fig. 2). Over-represented domains include EGF, intermediate filament protein, sushi, thrombospondin, complement C1q, and cysteine protease inhibitor, while underrepresented include Zinc finger (C2H2, B-box, RING), tyrosine protein phosphatase, tyrosine and serine/threonine protein kinases, helix-turn-helix motif, and IQ calmodulin binding region, compared with frequencies in the entire human genome.

Of the 1297 of the 3020 protein dataset that had identifiers in Swiss-Prot 44, 230 were annotated as transmembrane proteins. Another 25 have mitochondrial transit signals, and an *N*-terminal signal sequence occurred in 373 proteins. Putative PTMs were noted for 254, including 85 with phosphorylation and 45 with glycosylation sites. A separate analysis of nearly twice as many proteins based on EnsEMBL matches using the Human Protein Reference Database (www.hprd.org; Muthusamy *et al.*, [58]) found 628 with a signal sequence, 405 with transmembrane domains, 153 with a total of 1169 phosphorylation events, and 112 with a total of 555 glycolysation events.

One of the aims of the HUPO initiatives, as noted in the Section, is to link organ-based proteomes (liver, brain) with detection of corresponding proteins in plasma, and with proteins that are mediators, or at least, biomarker candidates, of inherited or acquired diseases. Using the Online Mendelian Inheritance in Man (OMIM), we found 338 of our 3020 IPI proteins that match EnsEMBL genes in OMIM, including RAG 2 for severe combined immunodeficiency (SCID)/Omenn syndrome, polycystin 1 for polycystic kidney disease (PKD), and BRCA 1, BRCA 2, p53, and APC for inherited cancer syndromes.

In the final article of this special issue, Martens *et al.* [59] describe the development and usefulness of the EBI PRoteomics IDEntifications database (PRIDE). The HUPO PPP dataset was the first large dataset to populate this database. The aim is to make publicly available data publicly accessible, in contrast to voluminous lists in printed articles or, more often now, in journals' websites, with custom layouts not suited to computer-based re-analysis. PRIDE offers an Application Programming Interface. In contrast, tables in PDF are described as notoriously difficult to extract. As noted, the PPP established a short-term solution with a relational database using a Microsoft Structured Query Language (SQL) server, which centralized all data collection and served as the testbed for the centralized, project-independent database that is now PRIDE. In turn, PRIDE has been designed with several features intended to facilitate future collaborative studies.

4.9 Identification of novel peptides using whole genome ORF search

A fascinating annotation from the PPP database has been used by States to enhance the annotation of the human genome itself [60]. The mass spectra data obtained by PPP investigators represent a resource for identifying novel and cryptic genes that may have been missed in previous annotations of the human genome. A total of 583 proteins in the 3020 protein set, including 185 identifications supported by three or more peptides, is not associated with genes in EnsEMBL. These are confident to highly confident experimental observations. The fact that they are not associated with known genes demonstrates that the annotation of the human genome remains incomplete.

To test the feasibility of this approach, we searched all ORFs using peak list data from six PPP laboratories (17, 30, 37, 41, 52, 55). NCBI human genome sequence build 33 was translated in all three reading frames and both strands; all non-redundant ORFs were assembled into chromosome specific sequence collections. The open source tool X!Tandem [61] was used in these analyses, with requirements for multiple mass spectra and a threshold hyperscore of 30 to accept peptide matches and greatly reduce the likelihood of false positive matches to ORFs. In all, 118 novel peptides were identified as highly probable matches to ORFs in the human genome not previously known to have protein products. This kind of protein-to-DNA mapping of the human genome is a notable bonus of the Plasma Proteome Project.

4.10 Identification of microbial proteins in the circulation

Microbial organisms populate all orifices and surfaces of many organs in the body, and their proteins may enter the blood intact or after degradation, as well as through contamination during venepuncture. We separately matched our peak lists for six small datasets against microbial genomes in the NCBI Microbial (non-human) GenBank (June 2004 release), using X!Tandem for RefSeq protein sequence identification. In this preliminary analysis, we found matches to several *E. coli* proteins (including elongation factor EF-Tu, outer membrane protein 3a, and glutamate decarboxylase isozyme) and mycobacterial proteins (members of glycine-rich PE-PGRS family) based on at least three peptide matches. No peptides for these proteins were found in the IPI human database, so these sequences are independent of the human gene and protein collections.

5 Discussion

This Special Issue of PROTEOMICS presents papers integral to the collaborative analysis, plus many reports of supplementary work on various aspects of the PPP workplan. The Core Dataset of 3020 proteins based on two or more peptide matches provides an anchor for future studies and for meta-analyses of the growing literature. These PPP results advance our understanding of complexity, dynamic range, biomarker potential, variation, incomplete sampling, false-positive matches, and integration of diverse datasets for plasma proteins. These results lay a foundation for development and validation of circulating protein biomarkers in health and disease. For the present, we recommend use of EDTA-plasma or citrate-plasma as the specimen of choice. Few labs actually compared these two alternative methods for plasma (Tables 1 and 2).

There are many opportunities for the HUPO Plasma Proteome Project going forward. First, these papers document our present understanding and reveal several open questions which require more focused studies: (a) to generate guidelines and standardized operating procedures for specimen collection, handling, archiving, and post-archive processing, including the protease inhibitor issue; (b) to use high-resolution methods to optimize specific immunoaffinity depletion of abundant proteins with minimal non-target losses; (c) to combine separation platforms and MS capabilities with an aim to expand the portion of the plasma proteome that can be profiled with confidence; (d) to achieve quantitative comparisons across specimens, not just compositional analyses; (e) to achieve high concordance in repeat analyses of the same specimen with the same methods; and (f) to overcome the extremely low overlap between protein identification datasets within a large collaboration of this type and, of course, across the literature, especially addressing the discrepancies due to post-MS/MS spectral analysis and peptide and protein database matching.

Other challenges are not specific to the plasma proteome, so we should discuss them together with other HUPO initiatives: (a) the limitations of present sequence databases, which are incomplete, redundant, and constantly being updated with corrections and new splice variants and SNPs; (b) the need to improve the true-positive to false-positive ratio, which requires explicit optimization; (c) the lack of reference specimen materials, which should be prepared with specific objectives and user communities in mind; (d) the need for independent corroboration of initial findings; and (e) organized strategies to validate proteomic discoveries and lead to microarray analyses with well-characterized antibodies, so that many specimens from clinical trials and epidemiological studies can be assayed. A new generation of studies will be considered at the Munich 4th HUPO Congress on Proteomics.

Second, there is an opportunity for the HUPO PPP to play a leading role in the continuing development and analysis of datasets arising from all quarters, in collaboration with the HUPO Protein Standards Initiative led by EBI [62] and other leading bioinformaticians, many of whom have contributed to this pilot phase of the PPP [62]. An immediate role for PPP is the cross-initatives analysis of Human Liver Proteome and Human Brain Proteome datasets with the PPP datasets, explicitly including experimental analyses of plasma samples from the same people and animals whose liver and brain specimens are studied. Several of the challenges listed above which involve search engine performance and integration of peptide identifications and protein matches with different databases deserve systematic investigation. Furthermore, quantitative analyses of concentrations, interactions, and networks will be increasingly important and feasible [63].

Third, there is an opportunity for HUPO to facilitate, and possibly organize, major disease-related studies of candidate biomarkers for earlier diagnosis, better stratification of newly diagnosed patients, appropriate pathways-based monitoring of targeted therapies, and design of preventive interventions. There is great anticipation of the application of ever-improving proteomics technologies for disease studies [64, 65].

For the overriding strategic question of gaining much higher throughput, at least four options have emerged in preliminary discussions:

(a) LC-MS with highly accurate mass and elution time parameters for peptide identification. A combination of specific depletion of abundant proteins, slow (2 h) nano-flow LC for elution time standardization, and highly accurate mass determination ($<$1 ppm) may make it feasible to base identifications solely on enhanced mass fingerprints once a high-quality accurate mass x elution time database with adequate sequence coverage of proteins to differentiate variants due to splicing, SNPs, and protein processing is in place. Additions to the database would require prior MS/MS identification.

(b) High accuracy LC-MS/MS/MS for peptide identifications. At the HUPO 3rd World Congress on Proteomics in Beijing, Mann described remarkable mass precision and very good efficiency of analysis with MS3, comprising MS/FT-ICR/MS. Applications to intracellular localization and discovery-phase identification of PTMs have already been achieved. It is likely, as with other methods, that an MS/MS or MS/MS/MS-based discovery phase would be converted into a different methodology, such as protein capture microarrays for high-throughput analysis of large numbers of plasma (or serum) specimens once the biomarkers were validated.

(c) Protein affinity micro-arrays. Humphery-Smith [66] proposed that affinity ligands be designed and produced to recognize conserved regions in each Open Reading Frame for signal enrichment. The ligands could be antibodies, receptions, aptamers, or other capture agents. The conserved regions might be sequences uncomplicated by PTMs, not subject to cleavage, and exposed at the surface. Enhanced chemiluminescence, rolling circle amplification, isotopic labeling, light scattering, or other methods could serve as

read-out technologies. This approach could improve protein identifications over a wide dynamic range.

(d) Isotope coded peptide standards for quantitative protein identification. Aebersold [67] proposed going from discovery using MS to "browsing" using unique chemically-synthesized peptides tagged with heavy isotope for each gene and even each protein isoform. This standard peptide mixture could be combined with specimen fractions on sample plates for MS. The double peaks would be examined with precise differential mass determination, using an ordered peptide array. This method would combine quantitation with identification, but the limits of dynamic range would persist.

In closing, the PPP Executive Committee expresses its appreciation to all the investigators and their associates, to the Technical Committee members, and to the government and corporate sponsors who have contributed greatly to the progress of the HUPO Plasma Proteome Project.

The HUPO Plasma Proteome Project received funding support under a trans-NIH grant supplement 84942 administered by the National Cancer Institute with participation from the National Institutes of Aging, Alcohol & Alcohol Abuse, Cancer (Prevention and Treatment Divisions), Diabetes, Digestive & Kidney Diseases, Neurological Diseases & Stroke, and Environmental Health Sciences. The Michigan Core had support from the Michigan Life Sciences Corridor grant MEDC-238. Corporate sponsors/partners provided funding, technology, specimens, datasets, and/or technical advice; we thank Johnson & Johnson, Pfizer, Abbott Laboratories, Novartis, Invitrogen, Procter & Gamble, BD Biosciences, Ciphergen, Agilent, Amersham, Bristol Myers Squibb, DadeBehring, Molecular Staging, Sigma-Aldrich, and BioVisioN.

6 References

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

[2] Mann, M., Aebersold, R., *Nature* 2003, *422*, 198–207.

[3] Hanash, S., *Drug Discov. Today* 2003, *7*, 797–801.

[4] Hanash, S. M., Celis, J. E., *Mol. Cell. Proteomics* 2002, *1*, 413–414.

[5] Li, X., Gong, Y., Wang, Y., Wu, S. *et al.*, *Proteomics* 2005, *5*, DOI: 10.1002/pmic.200400425.

[6] He, P., He, H-Z., Dai, J., Wang, Y. *et al.*, *Proteomics* 2005, *5*, DOI: 10.1002/pmic.200400422.

[7] Adamski, M., Blackwell, T., Menon, R., Martens, L. *et al.*, *Proteomics* 2005, *5*, DOI: 10.1002/pmic.200500186.

[8] Adamski, M., States, D. J., Omenn, G. S., Data Standardization and Integration in Collaborative Proteomics Studies, In: Srivastava, S. (Ed), *Informatics in Proteomics*. New York, Marcel Dekker, 2004, Chapter 8, pp. 169–194.

[9] Carr, S., Aebersold, R., Baldwin, M., Burlingame, A. *et al.*, *Mol. Cell. Proteomics* 2004, *3*, 351–353.

[10] Kersey, P. J., Duarte, J., Williams, A., Karavidopoulou, Y. *et al.*, *Proteomics* 2004, *4*, 1985–1988; http://www.ebi.ac.uk/IPI/IPIhelp.html.

[11] Kapp, E. A., Schutz, F., Connolly, L. M., Chakel, J. A. *et al.*, *Proteomics* 2005, *5*, DOI: 10.1002/pmic.200500126.

[12] Adkins, J. N., Monroe, M. E., Auberry, K. J., Shen, Y. *et al.*, *Proteomics* 2005, *5*, DOI: 10.1002/pmic.200400633.

[13] Olsen, J. V., Mann, M., *Proc. Natl. Acad. Sci. USA* 2004, *101*, 13417–13422.

[14] Tang, H-Y., Ali-Khan, N., Echan, L. A., Levenkova, N. *et al.*, *Proteomics* 2005, *5*, DOI: 10.1002/pmic.200401099.

[15] Nesvizhskii, A. I., Keller, A., Kolker, E., Aebersold, R., *Anal. Chem.* 2003, *75*, 4646–4658.

[16] Sadygov, R. G., Liu, H., Yates, J. R., *Anal. Chem.* 2004, *76*, 1664–1671.

[17] Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J. *et al.*, *J. Proteome Res.* 2003, *2*, 43–50.

[18] Tabb, D. L., Saraf, A., Yates, J. R., III, *Anal. Chem.* 2003, *75*, 6415–6421.

[19] Keller, A., Purvine, S., Nesvizhskii, A. I., Stolyar, S. *et al.*, *OMICS* 2002, *6*, 207–212.

[20] Sadygov, R. G., Yates, J. R., III, *Anal. Chem.* 2003, *75*, 3792–3798.

[21] MacCoss, M. J., Wu, C. C., Yates, J. R., *Anal. Chem.* 2002, *74*, 5593–5599.

[22] Haab, B. B., Geierstanger, B. H., Michailidis, G., Vitzthum, F. *et al.*, *Proteomics* 2005, *5*, DOI: 10.1002/pmic.200400470.

[23] Rai, A. J., Gelfand, C. A., Haywood, B. C., Warunek, D. *et al.*, *Proteomics* 2005, *5*, DOI: 10.1002/pmic.200400537.

[24] Tammen, H., Schulte, I., Hess, R., Menzel, C. *et al.*, *Proteomics* 2005, *5*, DOI: 10.1002/pmic.200400419.

[25] Misek, D. E., Kuick, R., Wang, H., Galchev, V. *et al.*, *Proteomics* 2005, *5*, DOI: 10.1002/pmic.200500103.

[26] Echan, L. A., Tang, H.-Y., Ali-Khan, N., Lee, K., Speicher, D. W., *Proteomics* 2005, *5*, DOI: 10.1002/pmic.200400518.

[27] Zolotarjova, N., Martosella, J., Nicol, G., Bailey, J. *et al.*, *Proteomics* 2005, *5*, DOI: 10.1002/pmic.200402021.

[28] Huang, L., Harvie, G., Feitelson, J. S., Herold, D. A. *et al.*, *Proteomics* 2005, *5*, DOI: 10.1002/pmic.200400420.

[29] Yang, Z., Hancock, W. S., Richmond-Chew, T., Bonilla, L., *Proteomics* 2005, *5*, DOI: 10.1002/pmic.200400411.

[30] Barnea, E., Sorkin, R., Ziv, T., Beer, I., Admon, A., *Proteomics* 2005, *5*, DOI: 10.1002/pmic.200400412.

[31] Moritz, R. L., Clippingdale, A. B., Kapp, E. A., Eddes, J. S. *et al.*, *Proteomics* 2005, *5*, DOI: 10.1002/pmic.200500096.

[32] Cho, S. Y., Lee, E.-Y., Chun, Y. W., Lee, J.-S. *et al.*, *Proteomics* 2005, *5*, DOI: 10.1002/pmic.200400497.

[33] Kim, J. Y., Lee, J. H., Park, G. W., Cho, K. *et al.*, *Proteomics* 2005, *5*, DOI: 10.1002/pmic.200400413.

[34] Zhang, H., Li, X. J., Martin, D. B., Aebersold, R., *Nat. Biotechnol.* 2003, *21*, 660–666.

[35] Zhou, M., Lucas, D. A., Chan, K. C., Issaq, H. J. *et al.*, *Electrophoresis* 2004, *25*, 1289–1298.

[36] Vollmer, M., Horth, P., Nagele, E., *Anal. Chem.* 2004, *76*, 5180–5185.

[37] Marshall, J., Jankowski, A., Furesz, S., Kireeva, I. *et al.*, *J. Proteome Res.* 2004, *3*, 364–382.

[38] Shen, Y., Jacobs, J. M., Camp, D. G., Fang, R. *et al., Anal. Chem.* 2004, *76*, 1134–1144.

[39] Qian, W. J., Liu, T., Monroe, M. E., Strittmatter, E. F. *et al., J. Proteome Res.* 2005, *4*, 53–62.

[40] Washburn, M. P., Wolters, D., Yates, J. R., *Nat. Biotechnol.* 2001, *19*, 242–248.

[41] Adkins, J. N., Varnum, S. M., Auberry, K. J., Moore, R. J. *et al., Mol. Cell. Proteomics* 2002, *1*, 947–952.

[42] Beer, I., Barnea, E., Ziv, T., Admon, A., *Proteomics* 2004, *4*, 950–960.

[43] Beer, I., Barnea, E., Admon, A., *Proteomics* 2005, *5*, DOI: 10.1002/pmic.200400457.

[44] Deutsch, E. W., Eng, J. K., Zhang, H., King, N. L. *et al., Proteomics* 2005, *5*, DOI: 10.1002/pmic.200500160.

[45] Keller, A., Nesvizhskii, A. I., Kolker, E., Aebersold, R., *Anal. Chem.* 2002, *74*, 5383–5392..

[46] Desiere, F., Deutsch, E. W., Nesvizhskii, A. I., Mallick, P. *et al., Genome Biol.* 2005, *6*, R9.

[47] Craig, R., Beavis, R. C., *Rapid Commun. Mass Spectrom.* 2003, *17*, 2310–2316.

[48] Martens, M., Nesvizhskii, A. I., Hermjakob, Adamski, M. *et al., Proteomics* 2005, *5*, DOI: 10.1002/pmic.200400376.

[49] Johnson, R. S., Davis, M. T., Taylor, J. A., Patterson, S. D., *Methods* 2005, *35*, 223–236.

[50] Anderson, N. L., Polanski, M., Pieper, R., Gatlin, T. *et al., Mol. Cell. Proteomics* 2004, *3*, 311–316.

[51] Pieper, R., Gatlin, C. L., Makusky, A. J., Russo, P. S. *et al., Proteomics* 2003, *3*, 1345–1364.

[52] Tirumalai, R. S., Chan, K. C., Prieto, D. A., Issaq, H. J. *et al., Mol. Cell. Proteomics* 2003, *2*, 1096–1103.

[53] Chan, K. C., Lucas, D. A., Hise, D., Schaefer, C. F. *et al., Clin. Proteomics* 2004, *1*, 101–225.

[54] Rose, K., Bougueleret, L., Baussant, T., Bohm, T. *et al., Proteomics* 2004, *4*, 2125–2150.

[55] Rai, A. J., Stemmer, P. M., Zhang, Z., Adam, B.-L. *et al., Proteomics* 2005, *5*, DOI: 10.1002/pmic.200400606.

[56] Ping, P., Vondriska, T. M., Creighton, C. J., Gandhi, T. K. B. *et al., Proteomics* 2005, *5*, DOI: 10.1002/pmic.200500140.

[57] Berhane, B., Zong, C., Liem, D. A., Huang, A. *et al., Proteomics* 2005, *5*, DOI: 10.1002/pmic.200401084.

[58] Muthusamy, B., Hanumanthu, G., Suresh, S., Rekha, B. *et al., Proteomics* 2005, *5*, DOI: 10.1002/pmic.200400588.

[59] Martens, M., Hermjakob, H., Jones, P., Adamski, M. *et al., Proteomics* 2005, *5*, DOI: 10.1002/pmic.200400647.

[60] Stein, L.D., *Nature* 2004, *431*, 915–916.

[61] Fenyö, D., Beavis, R.C., *Anal. Chem.* 2003, *75*, 768–774.

[62] Orchard, S., Hermjakob, H., Apweiler, R., *Mol. Cell. Proteomics* 2005, *4*, 435–440.

[63] Marko-Varga, G., Fehniger, T.E., *J. Proteome Res.* 2004, *3*, 167–178.

[64] Celis, J. E., Korc, M., *Mol. Cell. Proteomics* 2005, *4*, 345–593.

[65] de Hoog, C.L., Mann, M., *Annu. Rev. Genomics Hum. Genet.* 2004, *5*, 267–293.

[66] Humphery-Smith, I., *Proteomics* 2004, *4*, 2519–2521.

[67] Aebersold, R., *Nature* 2003, *422*, 115–116.

7 Addendum

[2] European Bioinformatics Institute, Hinxton, UK; [3] Van Andel Research Institute, Grand Rapids, MI, USA; [4] Ludwig Institute, Melbourne, Australia; [5] Johns Hopkins Univ, Baltimore, MD, USA; [6] Technion, Haifa, Israel; [7] Federal Institute of Technology (ETH), Zurich, Switzerland; [8] Institute for Systems Biology, Seattle, WA, USA; [9] Northeastern Univ, Boston, MA, USA; [10] Bristol Myers Squibb, NJ, USA; [11] Ruhr Unversity Bochum, Germany; [12] Yonsei Research Center, Seoul, Korea; [13] Korea Basic Science Institute, Seoul, Korea; [14] UCLA, Los Angeles, CA, USA; [15] Pacific Northwest National Lab, Richland, WA, USA; [16] Institute of Radiation Medicine, Beijing, China; [17] Mt Sinai School of Medicine, New York, NY, USA; [18] University of New South Wales, Australia; [19] Institute of Biological Chemistry, Taiwan; [20] Chinese Academy of Medical Sciences, Beijing, China; [21] Shanghai Institutes for Biological Sciences, Shanghai, China; [22] Institute of Biomedical Chemistry, Moscow, Russia; [23] NEC Proteomics Research Laboratory, Tsukuba Japan; [24] IBM, Haifa, Israel; [25] Proteomics Research Services Inc, Ann Arbor, MI, USA; [26] BioVisioN AG, Hannover, Germany; [27] Wistar Institute, Philadelphia, PA, USA; [28] Fred Hutchinson Cancer Research Center, Seattle, WA, USA

