REVIEW

THE HUPO Human Plasma Proteome Project

Gilbert S. Omenn

Internal Medicine, Human Genetics, Public Health and Center for Computational Medicine and Biology, University of Michigan, Ann Arbor, MI, USA

The Human Proteomics Organization (HUPO) Human Plasma Proteome Project (PPP) is a prominent example of the inherently collaborative nature of the overall community effort to characterize the proteome of humans in health and disease. The PPP Pilot Phase, called "Exploring the Human Plasma Proteome", engaged 55 laboratories, four technical committees, and vendors and sponsors on an international scale. Among other outcomes, the PPP generated a Core Dataset of 3020 proteins identified with two or more peptides, fully accessible at EBI/PRIDE, ISB/PeptideAtlas, and University of Michigan websites, a rich resource for follow-on analyses. The PPP provided extensive annotation, correlation of number of peptides with protein concentrations measured by immunoassay, an algorithm for choice of a representative protein for multiple proteins matching a given peptide, and independent analyses from the raw spectra. The next phase of the PPP will emphasize standardized procedures for specimen handling, potent new technology platforms for discovery and for targeted proteomics, and robust informatics efforts, including comparative analyses of other biofluids.

Received: April 11, 2007
Revised: May 23, 2007
Accepted: May 24, 2007

Keywords:
Annotation / Collaborative projects / HUPO / Mass spectrometry / Plasma proteome

1 Introduction

The HUPO Plasma Proteome Project (PPP) was initiated in 2002. It is part of a broad array of HUPO initiatives ([1–3]; www.hupo.org). Through its initiatives, its annual World Congress of Proteomics, the many active regional and national HUPO organizations, relationships with leading journals, and cooperation with EU and USA biomarker discovery and proteomics technology development initiatives, HUPO is having a major influence on acceleration of progress in proteomics.

Correspondence: Professor Dr. Gilbert S. Omenn, Internal Medicine, Human Genetics and Public Health, Center for Computational Medicine and Biology, University of Michigan, 100 Washtenaw Avenue, 2017F Palmer Commons, Ann Arbor, MI 48109-2218, USA
E-mail: gomenn@umich.edu
Fax: +1-734-615-6553

Abbreviations: **IPI**, International Protein Index; **MW**, molecular weight; **PPP**, Plasma Proteome Project

The long-term scientific goals of the PPP are (i) comprehensive analysis of the protein constituents of human plasma and serum; (ii) identification of biological sources of variation within individuals over time due to physiology (age, gender, menstrual cycle, exercise, stress, diet), pathology (various diseases, special cohorts), and treatments (common medications); and (iii) determination of the extent of variation across individuals within populations and across populations, due to genetic, nutritional, and other factors. The purpose of this report is to review the goals, development, and findings of the Pilot Phase of the PPP, to share some of the challenges and lessons of this significant collaborative effort, and to affirm the plans of HUPO to foster cross-analyses of organ and biofluid proteomes with the plasma or serum proteome.

The PPP Pilot Phase (Fig. 1) analyzed PPP reference specimens of human serum and EDTA-, heparin-, and citrate-anti-coagulated plasma; evaluated advantages and limitations of many depletion, fractionation, and MS technology platforms; and created a publicly available knowledge base at www.ebi.ac.uk/pride; www.peptideatlas.org and www.bioinformatics.med.umich.edu/hupo/ppp. Protocols

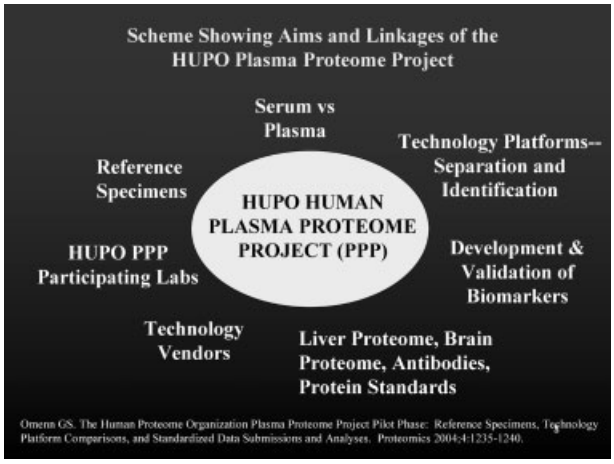

Figure 1. Scheme showing aims and linkages of the HUPO Plasma Proteome Project. Adapted from [2].

were developed by technical committees, and discussed and adopted at an investigators' workshop in July 2003. Many potential reference specimens were considered, ranging from aliquots of a pooled specimen from as many as 10 000 individuals (American Red Cross) to a single specimen from one or a few individuals. We chose to prepare sets of serum and plasma specimens from male-female donor pairs from four different ethnicities, to begin to address interests of the various investigators. The reference specimens were prepared by BD and its contractor and were distributed globally, beginning in September 2003. Each laboratory received 1.0 mL of the reference specimens requested. They used their established and emerging technologies for fractionation and analysis of proteins. They were encouraged to "push the limits" of their methods to detect and identify low abundance proteins.

The PPP Pilot Phase was highly collaborative. Thirty-five participating laboratories in 13 countries submitted datasets to the Data Coordinating Center at the University of Michigan. A small grants program facilitated follow-on analyses by 15 laboratories. An intensive 4-day Jamboree Workshop was held in June 2004 for real-time cross-analyses and identification of further analyses to be done. Working groups addressed specimen stability and protein concentrations; protein identifications from MS/MS datasets from 18 laboratories; independent analyses from the raw MS/MS spectra; search engine performance, subproteome analyses, and biological insights; antibody arrays; and direct MS/SELDI analyses across 10 laboratories. The 3020 proteins in the Core Dataset were characterized with Gene Ontology, InterPro, Novartis Atlas, OMIM, and immunoassay-based concentration determinations. The findings were reported in depth in 28 articles of a special issue of PROTEOMICS in August 2005 [4] and a Wiley book in 2006 [5].

As discussed in other articles in this issue, proteomic analyses of plasma or serum must cope with the challenges of high complexity, extreme dynamic range of concentrations, physiological and genetic variation, and still-evolving methods and databases for identification of peptides and matching to protein sequences.

2 Findings from the HUPO Plasma Proteome Project

2.1 MS/MS results

In all, 55 laboratories around the world requested the HUPO PPP reference specimens of serum and plasma from Caucasian-American, African-American, Asian-American, Chinese, and UK populations [4]. Eighteen of the participating laboratories conducted a wide range of depletion and fractionation protocols combined with MS/MS or FTICR-MS. They submitted 42 306 protein identifications using various search engines and databases to handle spectra and generate peptide sequence lists from the specimens analyzed. Peptides with $\geq$6 amino acids matched to 15 519 non-redundant proteins in the International Protein Index (IPI) of the European Bioinformatics Institute in Hinxton, UK [6]. IPI version 2.21 (July 2003) was the standard reference database for this Project. We designed an integration algorithm that selected one representative protein among multiple proteins (homologs and isoforms) to which the identified peptides gave 100% sequence matches [7]. This integration process resulted in 9504 proteins in the IPI v2.21 database. The PPP database counts homologous proteins and all isoforms of particular proteins just once, unless the sequences actually differentiated additional matches. We have reported details of the depletion, fractionation, and analytical methods and the cut-points or filters used by each investigator with the various search engines linked to different MS/MS instruments [4]. No equivalency rules were applied across all the search algorithms for all the cut-points. However, Kapp *et al.* [8] provided such a cross-algorithm analysis for three specified false-positive rates using one laboratory dataset and five search algorithms.

Data management for this Project comprised guidance and protocols for data collection, centralized integration and analysis, and dissemination of findings worldwide. Key challenges were integration of heterogeneous datasets, reduction of redundant information, dataset annotation, and how to keep straight progressive submissions and revisions of datasets from a laboratory analyzing multiple reference specimens with various combinations of technology platforms or successive thresholds for the search engine. Multiple factors had to be balanced, including when to "freeze" on a particular release of the ever-changing database selected for the PPP. Freezing of the database was essential for conducting 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 IPI protein sequence database version available at the initiation of the Project were revised, replaced, or withdrawn over the course of the Project. This fact complicates all cross-study comparisons; careful attention to the version of each database used and interpolation to a common version are essential for such comparisons. Our policies and practices anticipated the guidelines issued recently by Carr *et al.* [9], as documented by Adamski *et al.* [7].

Since the approaches and analytical instruments used by the various laboratories were too diverse to utilize a single standardized set of mass spec/search engine criteria, we created a defined subset of protein IDs from the 9504 above by requiring that the same protein be identified with at least a second peptide. Since the identification of peptide fragment ions is a low-percentage sampling process, additional analyses in the same lab and in other labs were expected to enhance the yield of peptide IDs. As demonstrated by lab 34, use of an LTQ instrument identified numerous proteins with two or more peptides that had been identified using an LCQ instrument with only one (high-confidence) peptide [10]. Consequently, MS data from individual laboratories were combined to increase the probability of peptide identification and protein assignment. Many labs reported results with a variety of methods for depletion of abundant proteins; one revealing study found new spots on 2-D gels after top-6 depletion, but few new proteins. The new spots were primarily additional isoforms of already-identified moderately abundant proteins [11].

Of the 9504 protein IDs, 3020 were based on two or more peptides. The list of 3020 proteins (5102 before integration) has been utilized as the Core Protein Dataset for the HUPO PPP knowledge base. Full details with unique IPI accession numbers for each protein are accessible for examination and new analyses at the University of Michigan (www.bioinformatics.med.umich.edu/hupo/ppp), European Bioinformatics Institute (www.ebi.ac.uk/pride), Institute for Systems Biology (www.peptideatlas.org), www.TheGPMdb.org, and www.hprd.org Human Proteinpedia websites. As a sample of use [12], during the period 5 to 31 January 2005, there were 5000 hits and 1000 downloads of PPP data from PRIDE. The 3020 proteins represent a very broad sampling of the IPI proteins in terms of characterization by p*I* and by molecular weight (MW) of the transcription product (often a "precursor" protein). The publicly available PPP database permits future users to choose their own cut-points for subanalyses, by minimal number of peptides matching, by confidence criteria, and especially by re-analysis from the raw spectra. Beer *et al.* [13] independently identified 2895 proteins with two or more peptides using PepMiner and Sequest for major PPP datasets, and Beavis matched 5816 IPI proteins with two or more peptides using X!Tandem [4]. They had 865 and 913 proteins in common with the PPP-3020, respectively. Kapp [8] and Deutsch [14] used Digger and ProteinProphet/PeptideAtlas on smaller PPP datasets [4].

2.2 Subproteomes

Two laboratories reported glycoprotein enrichment, one with hydrazide chemistry and the other with binding to three lectins [15, 16]. Together they had 254 protein IDs, of which 164 were reported also by other laboratories, while 90 were identified only after glycoprotein enrichment. These glycoprotein findings lay a foundation for the expected very large expansion of glycoprotein analyses with the N-glycosite peptide resource under preparation by the Aebersold laboratory [17]. Glycoprotein findings can be enhanced also with attention to the glycans [18, 19].

2.3 Direct MS/SELDI analyses

Ten laboratories requested PPP specimens for analyses with the then-popular SELDI chip fractionation, MS analysis, and algorithm-based differentiation of *m/z* peaks across specimens. Rai *et al.* [20] reported the cross-laboratory evaluation of eight submitted datasets, of which five were considered appropriate for comparison of plasma results and four for serum results. Correlations across labs were 0.7 or higher for 37 of 42 spectra with S/N$>$5. More detailed analyses identified just one protein, haptoglobin, with variation in intensity of its subunits in different reference specimens. Enhancement of the technology to actually identify proteins was recommended, along with stringent standardization of preanalytical variables, as for all proteomics technology platforms.

2.4 Quantitation of selected proteins

A critical parameter for detection and identification of proteins is the abundance or concentration of the protein and its isoforms. Haab *et al.* [21] generated a calibration curve for a set of sentinel proteins for which quantitative immunoassays were available. Four different assay methods were performed for the PPP by DadeBehring, Genomics Institute of Novartis Foundation, Molecular Staging, and Van Andel Research Institute. A total of 323 assays measured 237 unique analytes, although we cannot be certain that different assays for the presumed same protein targeted the same epitopes. The results were used to estimate dependence on concentration for proteins identified by MS. After extensive curation, Haab *et al.* matched 76 IPI proteins among the 9504 dataset and 49 proteins among the 3020-protein dataset. The results showed that the number of peptides identified for a protein in this collaborative dataset correlates highly with the measured concentration of the presumed same protein by immunoassays (correlation = 0.90 for 76 proteins in the 9504 dataset and 0.86 for 49 proteins in the 3020 dataset). As expected, the most abundant proteins are the most readily detected, with essentially 100% agreement; for much less abundant proteins only the laboratories with protocols and instruments capable of more sensitive detection identified these proteins. Biologically interesting proteins with measured concentra-

tions from 20 ng/mL down to 200 pg/mL include selectin L, activated leukocyte adhesion molecule, IGFBP-2, TIMP-1, EGFR, MMP-2/gelatinase, leukemia inhibitory factor receptor, PDGF-R-alpha, TNF-ligand-6, TNF-R-8, and alpha-fetoprotein.

2.5 Annotation of the HUPO PPP Core Dataset

From its inception, the HUPO PPP was designed to facilitate extensive and innovative annotation of the human plasma and serum proteome. A large element of the June 2004 Jamboree Workshop was collaborative annotation, leading to several papers in the Proteomics Special Issue [4].

Ping *et al.* [22] searched for evidence of cleavage of signal peptides, proteolysis within hydrophobic regions of transmembrane sites, and PTMs. Using the 2446 of the 3020 proteins that matched to Ensembl gene products, they highlighted subproteomes comprised of glycoproteins, low-MW proteins and peptides, DNA-binding proteins, and coagulation pathway, cardiovascular, liver, inflammation, and mononuclear phagocyte proteins. Notable were 216 proteins matched by Gene Ontology to DNA binding and 350 to the nucleus, including histone proteins. Liver dominated as the probable source of proteins, based on Novartis Atlas mRNA expression profiles for 79 human tissues. Many classic protein markers of leukocyte, platelet, or macrophage lineages were not detected, suggesting little contamination of the plasma. They also highlighted the biological significance of including semi-tryptic peptides, which might reveal intramembrane proteolysis. Berhane *et al.* [23] focused on 345 proteins of interest for cardiovascular research, classified into eight categories, most of which relate to other organ systems, as well: inflammation, vasoactive and coagulation proteins, signal transduction pathways, growth and differentiation-associated, cytoskeletal, transcription, channels and receptors, and heart failure and remodeling-related proteins. Muthusamy *et al.* [24] subjected protein and nucleotide sequences in NCBI for 2446 genes to BLAST queries, finding that 51% of the genes encoded more than one protein isoform. In addition, they mapped 11 381 single-nucleotide polymorphisms involving protein-coding regions onto protein sequences.

With Gene Ontology for subcellular localization, molecular processes, and biological functions, we found a very broad array of proteins identified. For example, subcellular classification of the 1276 IPI-3020 proteins included in Gene Ontology showed 26% from membrane components, 19% from nuclei, 11% from cytoskeleton, 23% from other cell types, and just 14% from secreted proteins ("traditional plasma proteins"). Examination of specific GO terms against a random sample of 3020 from the human genome [4, 12] showed over-represented proteins (>3 SD from the expected line) in categories of extracellular, immune response, blood coagulation, lipid transport, complement activation, and regulation of blood pressure, as expected, plus cytoskeletal proteins, receptors, and transporters. Corresponding InterPro analyses showed that domains associated with EGF, intermediate filament protein, sushi, thrombospondin, complement C1q, and cysteine protease inhibitor were over-represented (>3 SD) compared with random occurrence, while zinc finger RING protein, tyrosine protein phosphatase, tyrosine and serine/threonine kinases, helix-turn-helix motif, and IQ calmodulin-binding region were under-represented. In addition, we noted that 338 of the 3020 IPI proteins matched Ensembl genes in the Online Mendelian Inheritance in Man database, including such interesting disease-associated genes as RAG 2 for severe combined immunodeficiency (SCID/Omenn syndrome), polycystin 1 for polycystic kidney disease, and breast cancer BRCA1 and BRCA2, multi-cancer p53, and colon cancer APC for inherited cancer syndromes.

Subsequent analyses include reverse protein-to-DNA matching to identify proteins for previously unidentified ORF [25] and application of stringent adjustments for protein length and for multiple comparisons testing [26]. The latter yielded a data subset of 889 proteins; however, these particular adjustments may be overly stringent, since they assume equivalent random matching to all proteins, whereas proteins occur in various families and have considerable homologies. The plasma findings have been compared with those in liver and brain proteomes [27–29], and have helped to stimulate collective efforts to create protein capture reagent resources [30] and standard peptide, protein, and clinical reference specimens [31].

3 How many proteins can be detected in human plasma and serum?

Counting and comparing numbers of plasma or serum proteins identified is currently a chaotic process, starting with the biological variables of dynamic range, complexity, and physiological influences [32]. For any given MS technology platform, the variables of specimen collection, depletion of abundant proteins, fractionation of intact proteins, fractionation of tryptic peptides, choice of search engine and detailed search engine parameters for declaring peptide matches, choice of protein database and version thereof, consolidation or integration of multiply matched proteins, risk of false-positive identifications, and potential loss of true-positive identifications with stringent criteria to reduce false-positive identifications lead to extreme differences and awkward comparisons. The heterogeneity of results across the HUPO PPP laboratories illustrates the consequences of such variability of the approach. The inherently incomplete sampling of any single MS analysis and lack of data on coefficient of variation for individuals ensure that even a direct replication of the same specimen in the same lab will have limited concordance (typically ranging from 25 to 50%). Though there has been a call for up to 5–10 replications *per* specimen analyzed [33], present discovery-oriented reports often have no replication at all.

Given the exploratory nature of the Pilot Phase, laboratories utilized highly variable sets of reference specimens. In general, those performing extensive depletion and fractionation before tryptic digestion ran few specimens, while those with shotgun or other high throughput methods ran multiple reference specimens [4]. This variation limited the feasibility of cross-comparisons within the collaborative dataset.

There is a growing literature of results from extensive analyses of human plasma and serum specimens, enhanced by the PPP. Here we summarize certain features of previously published studies and compare the overlap of protein identifications with the PPP-3020 Core Dataset (see Table 1 and ref. [12]). The numbers of proteins depend on many factors, including the extent of fractionation and number of MS/MS runs with the sample, the number of peptide ions sequenced in MS/MS, the stringency of criteria for identification of peptides and minimization of false-positives, the restriction to tryptic or semi-tryptic peptides, the exclusion or inclusion of immunoglobulins and keratins, and the tolerance for multiple ambiguous assignments of the peptides to proteins in gene or protein databases.

3.1 Published analyses from others

Anderson *et al.* [34] published a compilation of 1175 non-redundant proteins reported in at least one of four sources (a literature review plus three experimental datasets). Of the 1175, only 195 were reported in any two of the four input datasets; only 46 proteins were reported in all four sources; 284 of the 468 reported in the non-proteomic literature were not found in any of the three experimental datasets; and only three of the 46 were not already known in the literature. Patterson and colleagues [35] suggested that such discordance reflects high false-positive rates from reliance on single-peptide hits. Shen *et al.* [36] used high-efficiency nanoscale RP-LC and strong cation exchange LC in conjunction with IT-MS/MS, Sequest peptide identification criteria (with and without chymotryptic and elastic peptides), and peptide LC normalized elution time constraints. Between 800 and 1682 human proteins were identified, depending on the criteria. They did not deplete albumin or immunoglobulins. Chan *et al.* [37] resolved and analyzed tryptic peptides from a Sigma pooled standard serum into 20 fractions by ampholyte-free liquid phase IEF, followed by strong cation-exchange chromatography, generating 7×20 fractions, which were analyzed by microcapillary RP-LC-MS/MS with an LCQ-DecaXP. In summary, they identified 1444 unique proteins from 2646 unique peptides after searching with Sequest against the Expert Protein Analysis System (www.expasy.org) database. A high percentage of proteins were based on just a single peptide (http://bpp.nci.nih.gov). Zhou *et al.* [38] identified an aggregate of 210 low-MW proteins or peptides after multiple immunoprecipitation steps with antibodies against albumin, IgA, IgG, IgM, transferrin, and apolipoprotein, followed by RP-LC-MS/MS. This aggregate result comprises 9 different experimental methods. Of these proteins, 73 and 67% were not found by the same lab in previous studies of the low-MW or whole-serum proteome [37]. There was no duplicate analysis to ascertain the concordance with the same method, same sample, and same lab. A grand total of only 378 unique peptides (not limited to tryptic peptides) was identified, which matched to 210 proteins; only 1 was identified by Adkins *et al.* [39] in serum depleted of IgG then analyzed by LC-MS/MS, only 4 in plasma by 2-DE with MS, and only 70 of the 1500 claimed by Chan *et al.* [37]. Rose *et al.* [40] reported an industrial-scale fractionation, starting with 6 L of blood/2.5 L of plasma from 53 healthy males, depleted of albumin and IgG with affinity resin and protein G, respectively, to yield 53 g total protein. Proteins of MW $<$ ca 40 kDa, 1.5 g after gel filtration, were separated into 12 960 chromatographic fractions. Fragments of larger proteins could not be excluded. ESI-MS and MALDI-TOF-MS were performed on the small proteins on MALDI plates, then aliquots of tryptic digests were subjected to LC-ESI-MS/MS; 1.5 million MS/MS spectra were analyzed with six different databases to yield 405 different proteins, of which 115 were based on a single peptide. When their criteria were applied to the Adkins *et al.* list of 490 proteins [39], only 164 of the more common proteins were retained.

Table 1. Overlap of HUPO PPP protein identifications with published datasets for plasma or serum

Published data	Total IDs	# IPI proteins	PPP_9504 dataset	PPP_3020 dataset
Anderson [34]	1175	990	471	316
Shen [36]	1682	1842	526	213
Chan [37]	1444	1019	402	257
Zhou [38]	210	148	88	62
Rose [40]	405	287	159	142

3.2 Degree of overlap of HUPO PPP protein ID with published studies of the plasma or serum proteome

Table 1 presents the matches of the five published studies described above with the HUPO PPP protein identifications. Of the 990 proteins which have IPI v2.21 identifiers in the four studies compiled by Anderson *et al.* [34], 316 are found in the PPP 3020 protein Core Dataset. When we relaxed the integration requirement (5102 IPI IDs), this figure rose only to 356 matches. Using the full 9504 dataset, the corresponding matches were 471 with integration and 539 without integration. We re-ran the raw spectra of Shen *et al.* [36] using HUPO PPP Sequest parameters and obtained 1842 IPI protein matches. Of these, 526 and 213 were found in the PPP 9504 and 3020 datasets, respectively.

When we mapped the 1444 proteins reported by Chan *et al.* [37] 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. With the Zhou *et al.* [38] protein-bound proteins, 148 proteins were mapped with IPI identifiers, of which 88 and 62 were found in the 9504 and 3020 PPP protein lists, respectively. Finally, of the 287 low-MW proteins (<40 kDa) from Rose *et al.* [40] which mapped to IPI v2.21identifiers, 159 and 142 are included in our 9504 and 3020 protein datasets, respectively.

These datasets vary remarkably in the protocols for depletion and/or fractionation, the criteria for protein ID, and the inclusion or depletion of immunoglobulins. All claim some relatively low abundance proteins. Nevertheless, abundance remains the single strongest determinant of protein detectability by mass spectrometry, and nearly all of the proteins detected in common across multiple studies are present at relatively high concentrations in blood.

Error rate estimation is a nascent aspect of the literature and a major source of lack of concordance. Methods include use of statistical criteria, as in PeptideProphet/ProteinProphet [41]; matching to non-human protein sequence databases (Archea); matching to reversed sequence [37] or shuffled sequence human databases; Poisson distribution methods [7]; or modeling of random matches to length of protein sequences [26]. The closest to standard usage is the Trans-Proteomic Pipeline based on PeptideProphet/ProteinProphet.

4 Plasma *vs.* serum as the sample of choice

Pre-analytical variables are often ignored or reported casually in the proteomic analysis and comparison of samples. These variables will be critically important in disease marker research. When blood is collected, many changes in proteins occur due to proteolytic enzymes (proteases) and other enzymes that are active in the blood sample during handling and processing. At the HUPO Jamboree concerns were raised that proteins from blood's cellular components may be released ex vivo due to hemolysis (breakdown of red blood cells with release of hemoglobin and other proteins), platelet activation (enhanced at 4°C, at least in some individuals, with release of platelet basic protein, thymosin-beta-4, platelet factor 4, zyxin as platelet markers), or white blood cell degranulation or breakdown with release of proteins. All of these proteins, especially the platelet markers, were found only with low numbers of peptides, signifying low concentration, in the PPP reference specimens.

Plasma is converted to serum by permitting or activating clot formation, usually at room temperature, which involves the protease action of thrombin on fibrinogen and related protein targets and other proteases on other proteins of the coagulation cascade. The forming clot itself provides a physical scaffold for attachment of proteins. Plasma can be protected from clotting by use of sodium citrate, K2-EDTA, or lithium heparin as anti-coagulant.

The HUPO PPP Specimens Committee [42] and the collaborating investigators [4] concluded that EDTA-plasma should be recommended as the preferred specimen from blood. This recommendation was endorsed at the HUPO 5th World Congress of Proteomics in Long Beach, CA, Oct 2006. Although truly systematic studies of the numerous variables involved are not available, scientists from several companies also supported this recommendation, based on unpublished experience. The reasons are (i) less degradation *ex vivo* and (ii) much less variability than arises in the protease-rich process of clotting. In the PPP, Misek *et al.* [43] showed with Cy5, Cy3, Cy2-labeled serum and plasma on DIGE-2D-PAGE after extensive fractionation of intact proteins before tryptic digestion that isoforms of abundant proteins were shifted to lower-than-expected MW more in serum than in plasma, and Tammen *et al.* from BioVisioN [44] reported that 40% of the low-MW peptides detected were serum specific. Clotting is unpredictable, due to influences of temperature, time, and medications, which are hard to standardize. Among anti-coagulants, heparin is a polyanion that activates anti-thrombin III; it may interfere in MS. Both citrate and EDTA inhibit coagulation and other enzymatic processes by chelate formation with ion-dependent enzymes. Citrate, however, introduces a 10–15% dilution effect, since it alone is added in solution. The PPP investigators therefore recommend EDTA-plasma as the preferred specimen from blood.

At present, we have left open the question of whether to include protease inhibitors in the collection tubes or buffers. Among the usual components of these cocktails, the peptide inhibitor aprotinin requires μg/mL concentrations, which may interfere with the analyses, while the small molecule inhibitor ABESF forms covalent bonds with proteins that alter the mobility of the protein [42]. BD has proposed that the PPP also use their new P100 EDTA-plasma tube, with a proprietary protease inhibitor cocktail and a mechanical separator of cells, which has been reported at HUPO and ASMS meetings to give more consistent results (Craft, D., Yi, J., Gelfand, C. A., An in-depth look at plasma peptidome stability in different blood collection tubes using LC-MALDI-MS. ASMS 2006, poster 578). Protocols for the standard EDTA-plasma and for the P100 tube have been prepared for the PPP next phase.

Meanwhile, a surprising finding that disease-related patterns of *ex vivo* proteolysis in plasma undergoing clotting to form serum may be clinically useful has been reported by the Tempst laboratory [45]. There are striking differences in peptide patterns between sera from bladder, breast, and prostate cancer patients compared with normals or with the other groups of cancer patients, presumably due to amino or carboxy peptidases. Protease inhibitors are not used in such experiments.

5 Comparisons of overlap of plasma proteome with proteomes of other biofluids

Reference specimen-quality results have been published from the laboratory of Matthias Mann for urine [46], tear fluid [47], and seminal fluid [48] using the potent new technologies platforms of hybrid linear IT-Fourier transform (LTQ-FT) and linear IT-Orbitrap (LTQ-Orbitrap) after fractionation of the intact proteins. Overlap with the HUPO PPP-3020 proteins is shown in Table 2. Of the 1543 proteins identified in urine, 910 were in the IPI v2.21 database, of which 293 were identified in the HUPO PPP-3020. The corresponding numbers for tears and semen are given in Table 2. We are also collaborating in a study of salivary fluid (Yan, W., Yu, W., Mueller, M., Cole, S. *et al.*, Systematic comparison of two human body fluid proteomes: saliva and plasma. Abst, HUPO 5th World Congress, Long Beach, CA, Oct 31, 2006), in which 1134 protein clusters were found, of which 432 are in the PPP-3020 database. Yamamoto *et al.* [49] have identified 3680 proteins from 2 or more peptides in urine in the HUPO Human Kidney-Urine Proteome Project. Meanwhile, Beretta reported at the January 2007 HUPO Initiatives Workshop over 8000 protein IDs after LTQ analysis of 193 protein fractions, of which 792 matched to the PPP-3020. Other studies of biofluid and organ proteomes being reported in this Issue will be interesting to compare for overlap with the plasma proteome.

This work will be enhanced by the HUPO/Invitrogen Test Sample Project announced in 2006 to assess MS capabilities of participating laboratories for sensitivity of detection, accuracy of identification, relative quantitation, and recognition of false-positive identifications. The first test sample has a carefully designed mixture of 20 highly purified human proteins produced in *Escherichia coli*. The proteins will be unknowns for the labs. The 20 proteins have equimolar concentrations (5 pmoles in 50 μL) and represent a wide range of p*I*, MW, and hydrophobicity. All labs, both academic and corporate, are encouraged to participate, to improve performance throughout the proteomics community. Subsequent protein mixtures are planned with widely varying concentrations and then larger numbers of proteins (greater complexity).

The National Cancer Institute, the National Institute for Standards and Technology, the European Bioinformatics Institute, and various academic and industry scientists have shown high interest in standardized reference materials of peptides, protein mixtures, and biological specimens [31]. Examples of standard mixtures include the CRM 470 of 15 human plasma proteins, widely used in Europe; the 18 nonhuman commercial proteins combined into a standard mixture by Keller *et al.* [50] at the Institute for Systems Biology; and tryptic digests of 300 commercially available proteins, for each of which at least 25 spectra have been generated with MALDI-TOF/TOF for use as standards in calibrating instruments and spiking samples (Strahler, J. R., Veine, D., Walker, A., Ulintz, P. *et al.*, A publicly available dataset of MALDI-TOF/TOF and LTQ mass spectra of known proteins. ASMS 2005, Bioinformatics poster 398).

The HUPO Protein Standards Initiative has generated several consensus standards [29, 51], including publications pending in Nature Biotechnology (2007) on Molecular Interactions, Sample Processing (MIAPE), Gel Electrophoresis, Mass Spectrometry (merging mzDATA and mzXML), Protein Modifications, and Proteomics Informatics and Controlled Vocabularies (www.psidev.info).

Table 2. Overlap of new biofluid proteome findings with HUPO PPP-3020 protein list

Proteome	Proteins	IPI 2.21	PPP-3020	Ref.
Urine	1543	910	293	[46]
Tears	491	313	117	[47]
Semen	923	560	180	[48]

6 Concept for the next phase of the human Plasma Proteome Project

Present plans for the next phase of the PPP, co-chaired by Ruedi Aebersold, Young-Ki Paik and Gil Omenn, include the following elements: (i) voluntary participation and real-time contribution of large datasets from major laboratories using advanced technology platforms, analyzing plasma and/or serum, often in combination with organ or disease proteome studies; (ii) standard specimen collection, using EDTA-plasma protocols; and (iii) a robust informatics effort with collaborative cross-analyses of plasma findings from multiple HUPO initiatives and from other published work.

The Holy Grail for plasma proteomics is high-resolution, high-sensitivity, and high-throughput analysis. It is certain that most of the protein biomarkers of greatest interest originating from disease processes in specific tissues will be at quite low abundance after dilution into 4L of blood and 17L of extracellular fluid. Thus, targeted approaches linked to proteomics findings of disease relevance in sites of primary disease and proximal biofluids are logical strategies [52]. In order to perform proteomics analyses on dozens, hundreds, or thousands of specimens from participants in clinical trials or epidemiological studies, and do so with replicates to assess intra-individual variation, new strategies with high throughput must complement the presently laborious methods for discovery and even validation of potential protein biomarkers.

Fortunately, new technology platforms for global analysis using LTQ-FT, LC-MS/MS/MS, and LTQ-Orbitrap [53, 54] and new platforms for targeted proteomics using heavy-isotope-labeled N-glycosite-containing proteotypic peptides [17, 55] and/or multiple reaction monitoring (MRM) with or without anti-peptide antibodies [56] offer great promise. Additional mining of high-quality spectra may increase yields of

protein identifications, as well [57]. High-accuracy LC-MS/MS/MS has been applied to intracellular localization and discovery-phase identification of PTMs [54]. As it is likely that most differences between specimens from patients with disease versus specimens from normals may be quantitative, methods of quantitative proteomics, such as isotope-coded affinity tags (ICAT), iTRAQ, and DIGE may be essential. Aebersold is developing a resource that will offer chemically synthesized NxS/T peptides tagged with heavy isotope for each gene and eventually each protein isoform needed for high discrimination as biomarkers [55]. These "proteotypic peptides" would permit spiking of specimens to facilitate identification of mass pairs with the same peptides in the biological specimen and quantitation of the peptide and its protein [17]. A multiplexed multiple reaction monitoring (MRM) approach with 1-D LC-MS/MS can be exploited to identify and quantitate high and medium abundance proteins which may have value as biomarkers. This approach has many potential variations, including use of sensitive anti-peptide antibody-enhanced assays for lower-abundance proteins [56]. These methods will be at the heart of the new phase of the HUPO PPP. Plasma analyses will be the final common pathway as investigators link studies of organ proteomes and disease biomarker discovery in biofluids with findings in plasma specimens collected under standardized conditions from the same patients at the same time.

7 Pitfalls in biomarker discovery and validation studies

Since the next phase of the PPP will be coupled with the umbrella HUPO Biomarker Initiative [3], a few comments on such studies are in order. Ransohoff [58, 59] outlined many sources of bias that may invalidate biomarker studies, regardless of technology platform. Bias occurs when cancer and non-cancer groups or their specimens are handled in systematically different ways, which occurs often. Problems arise in study design, selection of individuals, collection and processing of specimens, and complicated gene expression and proteomic laboratory analyses. Randomized clinical trials are designed to provide a fair and unbiased comparison by the creation of duplicate sets of circumstances in which only one factor that is relevant to the outcome is permitted to vary; other sources of variation can be evaluated against the independent variable. This principle of "keeping all variables equal between groups" is difficult to achieve in observational studies. Bias is more difficult to address in study design, conduct, and interpretation than such other problems as the generalizability of results and the influence of "chance". Bias is unconscious and unintentional. For example, if samples from elderly hospitalized cancer patients are compared with samples from young lab workers, any variables associated with age or hospitalization could dominate the differences found. Similarly, specimens collected at different times by different staff or in different locations, are likely to carry critical differences unrelated to the disease under study. If cancer patient specimens have been stored for many years longer than non-cancer control individuals' specimens, unevaluated variables may accumulate. In better designs, individuals are selected for sampling and study before they receive the test that is being assessed and before it is known whether they have the disease, as in prospective population studies. Every participant would receive the same evaluation for the test and the disease and, hopefully, have specimen collection under identical protocols in the same laboratories.

Regrettably, the amount of detail on such variables typically provided in publications is meager to perfunctory, whether one depends upon the published text or the supplementary material. Detailed guidelines have been published for Minimum Information about Microarray Experiments (MIAME) and Minimum Information about Proteomics Experiments (MIAPE) (www.psidev.info), and for research publications [60]. A multi-author paper with guidance for clinical proteomics has appeared in this Journal [61], giving attention to selection of donors of specimens, detailed diagnostic criteria, and preanalytical aspects, including specimen collection and handling. A scheme called Standards for Reporting of Diagnostic Accuracy (STARD) [62] guides investigators and readers to sources of bias and details to be considered and be prevented. It is crucial to recognize that certain important aspects of study design do not overcome bias: (i) large sample size, which reduces the statistical confidence interval around a result; (ii) reproducibility of findings, as in use of training and testing sets of samples; and (iii) sophisticated statistical analyses of collected data. None of these methods and no guidelines for reporting can replace the investigator's insight and careful attention to details in seeking to ensure equality of characteristics of individuals studied, specimen collection, handling, and storage, and time and place of analyses.

8 Concluding remarks

Biomarker discovery is now a main theme in each of the many HUPO initiatives. There is increased awareness of the challenges ahead of us in biomarker discovery. In particular, the need for cross-laboratory examination of datasets from various organs, careful validation of targets, and establishment of regularly updated databases, is better appreciated. The path toward collaborative biomarker discovery at HUPO was initiated with the HPPP, and the work on plasma proteome functional annotation has inspired and facilitated many follow-up analyses. Organ-specific cross-analyses between the HPPP and the Liver and Brain Proteome Projects are being conducted, and cross-analyses of the Kidney-Urine, Salivary, and other projects will be organized.

In summary, work on all aspects of proteomics has generated renewed confidence that such approaches will reveal important features of normal biology and physiology and assist in the discovery, validation, and application of protein

biomarker panels in early diagnosis of disease and monitoring responses to therapies. Overall, the technical advances of proteomics and the intellectual power of international and inter-sectoral collaboration place HUPO in a key position to lay a strong foundation for clinical proteomics, through credible application of protein biomarkers in population screening and eventually in patient care.

The HUPO Plasma Proteome Project was funded under a trans-NIH grant 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, and Kidney Diseases, Neurological Diseases & Stroke, and Environmental Health Sciences. The Michigan Core has had support from Michigan Life Sciences Corridor grants MEDC-238 and MTTC-687, and NIH grants 1 U54DA021519 and 23XS110A. 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, Dade-Behring, GenWay, Molecular Staging, Sigma-Aldrich, and BioVisioN. I am grateful to the many collaborating scientists throughout the world (see Ref. [4]), including my colleagues in the University of Michigan Proteomics Alliance M. Adamski, P. Andrews, T. Blackwell, D. Fermin, B. Haab, S. Hanash, R. Kuick, R. Menon, D. Misek, M. Pisano, and D. States.

9 References

[1] Hanash, S. M., Celis, J. E., The Human Proteome Organization: a mission to advance proteome knowledge. *Mol. Cell. Proteomics* 2002, *1*, 413–414.

[2] Omenn, G. S., The Human Proteome Organization Plasma Proteome Project pilot phase: reference specimens, technology platform comparisons, and standardized data submissions and analyses. *Proteomics* 2004, *4*, 1235–1240.

[3] Omenn, G. S., Ping, P., *The future: Translation from discovery to the clinic—roles of HUPO and industry in biomarker discovery*. In: Van Eyk, J., Dunn, M. (Eds.), *Clinical Proteomics*, Wiley-VCH, Weinheim 2007 (in press).

[4] Omenn, G. S., States, D. J., Adamski, M., Blackwell, T. W. *et al.*, Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of proteins and a publicly-available database. *Proteomics* 2005, *5*, 3226–3245. *Proteomics* HPPP Special Issue, pp. 3223–3519.

[5] Omenn, G. S. (Ed.), *Exploring the Human Plasma Proteome*. Wiley-VCH, Weinheim 2006.

[6] Kersey, P. J., Duarte, J., Williams, A., Karavidopoulou, Y. *et al.*, The International Protein Index: an integrated database for proteomics experiments. *Proteomics* 2004, *4*, 1985–1988 [http://www.ebi.ac.uk/IPI/IPIhelp.html].

[7] Adamski, M., Blackwell, T., Menon, R., Martens, L. *et al.*, Data management and preliminary data analysis in the pilot phase of the HUPO Plasma Proteome Project. *Proteomics* 2005, *5*, 3246–3261.

[8] Kapp, E. A., Schutz, F., Connolly, L. M., Chakel, J. A. *et al.*, An evaluation, comparison and accurate benchmarking of several publicly-available MS/MS search algorithms: sensitivity and specificity analysis. *Proteomics* 2005, *5*, 3475–3490.

[9] Carr, S., Aebersold, R., Baldwin, M., The need for guidelines in publication of peptide and protein identification data. *Mol. Cell. Proteomics* 2004, *3*, 351–353.

[10] Tang, H. Y., Ali-Khan, N., Echan, L. A., Levenkova, N. *et al.*, A novel 4-dimensional strategy combining protein and peptide separation methods enables detection of low abundance proteins in human plasma and serum proteomes. *Proteomics* 2005, *5*, 3329–3342.

[11] Echan, L. A., Tang, H-Y., Ali-Khan, N., Lee, K., Speicher, D. W., Depletion of multiple high-abundance proteins improves protein profiling capacities of human serum and plasma. *Proteomics* 2005, *5*, 3292–3303.

[12] Omenn, G. S., Menon, R., Adamski, M., Blackwell T. *et al.*, The human plasma and serum proteome. In: Thongboonkerd, V., (Ed.), *Proteomics of Human Body Fluids: Principles, Methods, and Applications*. Humana Press, Totowa, NJ, USA 2007, pp. 195–224.

[13] Beer, I., Barnea, E., Admon, A., Centralized data analysis of a large interlaboratory proteomics project: a feasibility study. *Proteomics* 2005, *5*, 3491–3496.

[14] Deutsch, E. W., Eng, J. K., Zhang, H., King, N. L. *et al.*, Human plasma peptide atlas. *Proteomics* 2005, *5*, 3497–3500.

[15] Zhang, H., Li, X. J., Martin, D. B., Aebersold, R., Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry. *Nat. Biotechnol.* 2003, *21*, 660–666.

[16] Yang, Z., Hancock, W. S., Richmond-Chew, T., Bonilla, L., A study of glycoproteins in human serum and plasma reference standards (HUPO) using multi-lectin affinity chromatography coupled with RPLC-MS/MS. *Proteomics* 2005, *5*, 3353–3366.

[17] Kuster, B., Schirle, M., Mallick, P., Aebersold, R., Scoring proteomes with proteotypic peptide probes. *Nat. Rev. Mol. Cell. Biol.* 2005, *6*, 577–583.

[18] Tanaguchi, N., Miyoshi, E., Jianguo, G., Honke, K., Matsumoto, A., Decoding sugar functions by identifying target glycoproteins. *Curr. Opin. Struct. Biol.* 2006, *16*, 561–566.

[19] Wada, Y., Azadi, P., Costello, C.E., Dell, A. *et al.*, Comparison of the methods for profiling glycoprotein glycans: HUPO HGPI (Human Proteome Organization Human Disease Glycomics/Proteome Initiative) multi-institutional study. *Glycobiology* 2007, *17*, 411–422.

[20] Rai, A. J., Stemmer, P. M., Zhang, Z., Adam, B. L. *et al.*, Analysis of HUPO PPP reference specimens using SELDI-TOF mass spectrometry: multi-institution correlation of spectra and identification of biomarkers. *Proteomics* 2005, *5*, 3467–3474.

[21] Haab, B. B., Geierstanger, B. H., Michailidis, G., Vitzthum, F. *et al.*, Immunoassay and antibody microarray analysis of the HUPO PPP reference specimens: systematic variation between sample types and calibration of mass spectrometry data. *Proteomics* 2005, *5*, 3278–3291.

[22] Ping, P., Vondriska, T. M., Creighton, C. J., Gandhi, T. K. *et al.*, A functional annotation of subproteomes in human plasma. *Proteomics* 2005, *5*, 3506–3519.

[23] Berhane, B. T., Zong, C., Liem, D. A., Huang, A. *et al.*, Cardiovascular-related proteins identified in human plasma by the HUPO Plasma Proteome Project pilot phase. *Proteomics* 2005, *5*, 3520–3530.

[24] Muthasamy, B., Hanumanthu, G., Suresh, S., Rekha, B. *et al.*, Plasma proteome database as a resource for proteomics research. *Proteomics* 2005, *5*, 3531–3536.

[25] Fermin, D., Allen, B. B., Blackwell, T. W., Menon, R. *et al.*, Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. *Genome Biol.* 2006, *7*, R35.

[26] States, D. J., Omenn, G. S., Blackwell, T. W., Fermin, D. *et al.*, Deriving high confidence protein identifications from a HUPO collaborative study of human serum and plasma. *Nat. Biotechnol.* 2006, *24*, 333–338.

[27] Zheng, J., Gao, X., Beretta, L., He, F. The Human Liver Proteome Project (HLPP) Workshop during the 4th HUPO World Congress. *Proteomics* 2006, *6*, 1716–1718.

[28] Hamacher, M., Apweiler R., Arnold, G., Becker, A. *et al.*, HUPO Brain Proteome Project: summary of the pilot phase and introduction of a comprehensive data reprocessing strategy. *Proteomics* 2006, *6*, 4890–4898.

[29] Mueller, M., Martens, L., Apweiler, R. Annotating the human proteome: beyond establishing a parts list. *Biochim. Biophys. Acta* 2007, *1774*, 175–191.

[30] Haab, B. B., Paulovich, A. G., Anderson, N. L., Clark, A. M. *et al.*, A reagent resource to identify proteins and peptides of interest for the cancer community. *Mol. Cell. Proteomics* 2006, *5*, 1996–2007.

[31] Barker, P. E., Wagner, P. D., Stein, S. E., Bunk, D. M. *et al.*, Standards for plasma and serum proteomics in early cancer detection: a needs assessment from the NIST-NCI SMART workshop. *Clin. Chem.* 2006, *52*, 1669–1674.

[32] Anderson, N. L., Anderson, N. G., The human plasma proteome: history, character, and diagnostic prospects. *Mol. Cell. Proteomics* 2002, *1*, 845–867.

[33] Hamacher, M., Marcus, K., Stuhler, K., van Hall, A. *et al.*, (Eds.), *Proteomics in Drug Research*, Wiley-VCH, Weinheim 2007, p. xiii.

[34] Anderson, N. L., Polanski, M, Pieper, R., Gatlin, T. *et al.*, The human plasma proteome: a nonredundant list developed by combination of four separate sources. *Mol. Cell. Proteomics* 2004, *3*, 311–316.

[35] Johnson, R. S., Davis, M. T., Taylor, J. A., Patterson, S. D., Informatics for protein identification by mass spectrometry. *Methods* 2005, *35*, 223–236.

[36] Shen, Y., Jacobs, J. M., Camp, D. G. II, Fang, R. *et al.*, Ultra-high-efficiency strong cation exchange LC/RPLC/MS/MS for high dynamic range characterization of the human plasma proteome. *Anal. Chem.* 2004, *76*, 1134–1144.

[37] Chan, K. C., Lucas, D. A., Hise, D. Schaefer, C. F. *et al.*, Analysis of the human serum proteome. *Clin. Proteomics* 2004, *1*, 101–226.

[38] Zhou, M., Lucas, D. A., Chan, K. C., Issaq, H. J. *et al.*, An investigation in the human serum interactome. *Electrophoresis* 2004, *25*, 1289–1298

[39] Adkins, J. N., Varnum, S. M., Auberry, K. J., Moor, R. J. *et al.*, Toward a human blood serum proteome: analysis by multidimensional separation coupled with mass spectrometry. *Mol. Cell. Proteomics* 2002, *1*, 947–952.

[40] Rose, K., Bougueleret, L., Baussant, T., Bohm, G. *et al.*, Industrial-scale proteomics: from liters of plasma to chemically synthesized proteins. *Proteomics* 2004, *4*, 2125–2150.

[41] Nesvizhskii, A. I., Keller, A., Kolker, E., Aebersold, R., A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* 2003, *75*, 4646–4658.

[42] Rai, A. J., Gelfand, C. A., Haywood, B. C., Warunek, D. J. *et al.*, Human Proteome Organization—Plasma Proteome Project specimen collection and handling: towards the standardization of parameters for plasma proteome samples. *Proteomics* 2005, *5*, 3262–3277.

[43] Misek, D. E., Kuick, R., Wang, H., Galchev, V. *et al.*, A wide range of protein isoforms in serum and plasma uncovered by a quantitative intact protein analysis system. *Proteomics* 2005, *5*, 3343–3352.

[44] Tammen, H., Schulte, I., Hess, R., Menzel, C. *et al.*, Peptidomic analysis of human blood specimens: comparison between plasma specimens and serum by differential peptide display. *Proteomics* 2005, *5*, 3414–3422.

[45] Villanueva, J., Shaffer, D.R., Philip, J., Chaparro, C. A. *et al.*, Differential exoprotease activities confer tumor-specific serum peptidome patterns. *J. Clin. Invest.* 2006, *116*, 271–284.

[46] Adachi, J., Kumar, C., Zhang, Y., Olsen, J. V., Mann, M., The human urinary proteome contains more than 1500 proteins, including a large proportion of membrane proteins. *Genome Biol.* 2006, *7*, R80.

[47] de Souza, G., Godoy, L. M. F., Mann, M., Identification of 491 proteins in the tear fluid proteome reveals a large number of proteases and protease inhibitors. *Genome Biol.* 2006, *7*, R72.

[48] Pilch, B., Mann, M., Large-scale and high-confidence proteomic analysis of human seminal plasma. *Genome Biol.* 2006, *7*, R40.

[49] Yoshida, Y., Miyazaki, K., Kamiie, J., Sato, M. *et al.*, Two-dimensional electrophoresis profiling of normal human kidney glomerulus proteome and construction of an extensible markup language (XML)-based database. *Proteomics* 2005, *5*, 1083–1096.

[50] Keller, A., Purvine, S., Nesvizhskii, A. I., Stolyar, S. *et al.*, Experimental protein mixture for validating tandem mass spectral analysis. *OMICS: A Journal of Integrative Biology* 2002, *6*, 207–212.

[51] Orchard, S., Hermjakob, H., Apweiler, R., The proteomics standards initiative. *Proteomics* 2003, *3*, 1374–1376.

[52] Omenn, G. S., Strategies for proteomic profiling of cancers. *Proteomics* 2006, *6*, 5662–5673.

[53] Domon, B., Aebersold, R., Mass spectrometry and protein analysis. *Science* 2006, *312*, 212–217.

[54] Olsen, J. V., Mann, M., Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. *Proc. Natl. Acad. Sci. USA* 2004, *101*, 13417–13422.

[55] Aebersold, R., Constellations in a cellular universe. *Nature* 2003, *422*, 115–116.

[56] Anderson, L., Hunter, C. L., Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. *Mol. Cell. Proteomics* 2006, *5*, 573–588.

[57] Nesvizhskii, A. I., Roos, F. F., Grossman, J., Vogelzang, M. *et al.*, Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data. *Mol. Cell. Proteomics* 2006, *5*, 652–670.

[58] Ransohoff, D. F., Lessons from controversy: ovarian cancer screening and serum proteomics. *J. Natl. Cancer Inst.* 2005, *97*, 315–319.

[59] Ransohoff, D. F., Bias as a threat to the validity of cancer molecular-marker research. *Nat. Rev. Cancer* 2005, *5*, 142–149.

[60] Bradshaw, R. A., Burlingame, A. L., Carr, S., Aebersold, R., Reporting protein identification data: the next generation of guidelines. *Mol. Cell. Proteomics* 2006, *5*, 787–788.

[61] Mischak, H., Apweiler, R., Banks, R. E., Conaway, M. *et al.*, Clinical proteomics: a need to define the field and to begin to set adequate standards. *Proteomics Clin. Appl.* 2007, *1*, 148–156.

[62] Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A. *et al.*, The STARD Statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Ann. Intern. Med.* 2003, *138*, W1–W12.

