HUPO Highlights

Recent Workshops of the HUPO Human Plasma Proteome Project (HPPP): A bridge with the HUPO CardioVascular Initiative and the emergence of SRM targeted proteomics

Gilbert S. Omenn, Mark S. Baker and Ruedi Aebersold

Center for Computational Medicine, University of Michigan, Ann Arbor, MI, USA
Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, Australia
Department of Biology, Institute of Molecular Systems Biology, ETH-Zurich, Switzerland
Faculty of Science, University of Zurich, Switzerland

We hereby provide a two-year update on the HUPO Human Plasma Proteome Project (HPPP) informed by advances presented at the HPPP sessions at the HUPO World Congresses in Toronto in September 2009 and in Sydney in September 2010.

Keywords:
Biomedicine / Human proteome organisation / Plasma / Plasma proteome

The 2009 HPPP+HCVI Workshop at the Toronto HUPO Congress

The 2009 Workshop was an experiment in co-scheduling two HUPO Initiatives, the HPPP and HUPO CardioVascular Initiative. The organizers, chairs Peipei Ping and Gil Omenn, aimed to stimulate interactions among the speakers and the attendees, which was successful.

Terry Farrah of the Institute of Systems Biology gave a 2009 update for the Human Plasma PeptideAtlas, expanding the 2005 HPPP reports [1] and the initial Human Plasma PeptideAtlas of 2007 (www.peptideatlas.org). Shotgun MS/MS data sets were re-analyzed from raw spectra with the Trans Proteomic Pipeline (TPP), searching with SpectraST for non-glyco and with X!Tandem for glycoprotein data. The Human Plasma PeptideAtlas, including N-glyco- and Non-glyco-subatlases, can be used conveniently to check whether proteins identified in tissues, or in new plasma or serum studies, have been identified previously in plasma and what specific peptides and spectra have been observed. Several enhancements to Peptide Atlas were described as under development: more sophisticated statistical modeling with iProphet to drive down protein false discovery rates (FDR); decoy searches to facilitate estimation of protein FDR using Mayu; multi-level protein identifications to reduce redundancy; and estimated abundance from spectral counting and quantitative immunoassay data. iProphet and decoy searches are inserted between PeptideProphet and ProteinProphet in the TPP. The Human Plasma PeptideAtlas was populated first with HPPP-I data sets and expanded with academic and corporate submissions.

HUPO Human Proteome Organisation Highlights

HPPP-II increments from the Smith Lab at Pacific Northwest/Battelle and N-glyco-data sets from others boosted the cumulative protein identifications. At <1% protein FDR, the number of protein matches recognized can be reduced from 8314 sequence specific to 4334 peptide-set-unique to 2462 distinguishable to 2050 covering set to 1986 canonical. In PeptideAtlas these adjectives appear in the column ''presence level.'' This range illustrates how hard it is to compare numbers of protein identifications processed with different rules and different tolerances for potential redundancy. Regardless, the greater the protein sequence coverage, the lesser the ambiguity of peptide-to-protein matches. A detailed publication of these PeptideAtlas enhancements and updates has now been published [2].

Henning Hermjakob of the European Bioinformatics Institute and Eric Deutsch of the Institute for Systems Biology described progress toward the ProteomeXchange that will link primary data repositories at EBI/PRIDE and the then-new NIH/Peptidome via the distributed file-sharing system Tranche at www.proteomecommons.org, developed at the University of Michigan, and then to systematic reanalysis at ISB/PeptideAtlas in Seattle and at GPMDB in Winnipeg. Unfortunately, currently, due to budget pressures, NIH has terminated Peptidome and Tranche is stressed by very high data set load and limited staffing.

From the ETH-Zurich Aebersold Laboratory, Bernd Wollscheid highlighted progress on proteotypic peptides and selected reaction monitoring (SRM) for plasma. This approach provides a strategy and N-glycosite peptide reagents for quantitative identification of targeted peptides. Glycoproteins represent an especially informative subproteome, enriched for cell surface and secreted proteins. Because albumin is not a glycoprotein, this strategy reduces complexity and enhances dynamic range. Meanwhile, more than 80% of FDA-approved clinical marker assays are glycoprotein analytes. Since the early adaptation of MRM by Leigh Anderson and Christie Hunter (Applied Bio), there has been rapid emergence and availability of QQQ mass spectrometers and software to guide the choice of peptides with favorable ionization and generation of distinctive transitions in the third quadrupole at scheduled retention times. Wollscheid discussed how to select the proteins and their peptides for SRM assay development and how to generate and optimize the SRM assays. PeptideAtlas and UniPep (human N-linked glycosites) are key databases. Synthetic peptides can be produced on solid-stage substrates from which they are released by chemical cleavage for use in SRM assays (e.g. JPT, Berlin). ETHZ has produced a library of SRM spectra for the entire yeast proteome [3], and is rapidly generating 5000 peptides for 2300 proteins from the human N-glycosite proteome.

Rajasree Menon of the University of Michigan presented an analytical framework for the detection and annotation of alternative splice variant peptides, a new class of isoform-specific protein cancer biomarkers in plasma and tumors. Alternative splicing generates protein diversity without increasing genome size. The analytical pipeline identifies and quantifies known and novel splice isoforms from peptide sequences determined by LC-MS/MS. MS/MS spectra are interrogated against a non-redundant database of exhaustive three-frame translation of Ensembl transcripts and gene models from ECgene using the X!Tandem software. The search results are processed for peptide-to-protein integration by TPP and Michigan Peptide-to-Protein Integration (MPPI) [4]. Over-expressed novel variants are validated by qRT-PCR. The exact splicing findings and the biological annotations for differentially expressed splice variant proteins in pancreatic, breast, and colon cancers are quite interesting.

Jennifer Van Eyk of Johns Hopkins University illustrated the application of proteomics in cardiovascular diseases with a focus on aortic aneurysms. Aortic dissection, resulting from a tear in the aorta, is associated with a mortality rate of 1% per hour without surgical intervention. Aortic dissection can be caused by

multiple mechanisms, including inherited connective tissue disorders, like Marfan syndrome (MFS, a result of genetic defects in fibrillin 1). Proteomics have been applied to investigate both the natural history and MFS disease. 2-DE and MS-based iTRAQ quantification of aortic tissue obtained from young and old rats indicated numerous secreted proteins were increased, including the glycoprotein MFG-E8, which was shown to alter smooth muscle cells via a novel interplay with Ang II [5]. Quantification of MFG-E8 along with other age-related secreted proteins (like TGF-β1) has potential to determine ''biological age'' and associated risk for cardiovascular disease also in humans. In fact, the quantitation of circulating levels in the mouse model of MFS showed that TGF-β1 may allow assessment of the activity of aortic aneurysms and was sensitive to therapy with losartan, an AT1-inhibitor that significantly decreases aortic diameter in MFS patients. TGF-β1 was verified as a potential marker in the GenTac patient registry that enrolls patients with aortic aneurysms from a broad spectrum of heritable cardiac diseases and syndromes [6]. Mingning Ning of Massachusetts General Hospital and Harvard Medical School in Boston discussed how the brain–heart connection can lead to various neurovascular disease states. An extreme version of the influence of the brain on heart functions occurs when people are literally ''scared to death" resulting, most likely, from a massive catecholamine surge. Conversely, the heart can influence and induce neurovascular injury. Dr. Ning's laboratory concentrates on the underlying mechanisms in cardiac structural abnormalities, such as patent foramen ovale (PFO), which can lead to embolic strokes. Using direct bedside intra-cardiac plasma sampling and quantitative temporal protein profiling of patients undergoing PFO closure, her team has been able to discern some of the underlying physiological mechanisms. Analyses of the plasma microparticle sub-proteome has been useful in mapping the important brain–heart interactions in such neurovascular injuries.

Michael Kuzyk of the University of Victoria in British Columbia, Canada presented data on plasma-based SRM assays for 45 analytes, which were selected as potential biomarkers specifically to be able to distinguish patients with coronary artery disease from normal control subjects. Peptide standards were created for each SRM assay allowing absolute quantitation of the analytes from tryptic digests of human EDTA-plasma without prior affinity depletion or enrichment. To achieve maximal sensitivity and specificity, instrument parameters were adjusted to generate the most abundant precursor ions and y ion fragments. Excellent linear responses ($r > 0.99$) were obtained for 43 of the 45 proteins, with attomole level limits of quantitation and $< 20\%$ coefficient of variation for 27 of the 45 proteins and with analytical precision for 44 of the 45 assays, varying by $< 10\%$ [7]. Concentrations for 39 of the 45 proteins were within a factor of 2 of the reported literature values based on ELISAs.

The 2010 HPPP Workshop Session in the Sydney HUPO Congress

Eric Deutsch of the Institute for Systems Biology provided an update on the Human Plasma PeptideAtlas. Currently, there are approximately 2000 distinguishable, canonical proteins with a protein FDR of 1%, the largest high-confidence proteome compiled to date (note 2009 Human Plasma PeptideAtlas update above). Spectra came from 69 conventional LC-MS/MS-based plasma proteome studies and from 22 studies in which tryptic peptides isolated via solid-phase glycocapture were searched against a target-decoy sequence database using the X!Tandem [1]. All peptide-spectrum matches (PSMs) were assigned probabilities using the TPP statistical validation tools. Spectra for each experiment below a given PSM FDR (0.0002 for non-glycocapture, 0.00003 for

glycocapture) were compiled. FDRs were decoy-estimated at 1% for proteins and 0.16% for peptides, lower than previously reported human plasma peptide lists. He showed how these data can be leveraged by the PeptideAtlas query tools to enable targeted proteomics via SRM mass spectrometry.

At the Human Proteome Project (HPP) launch session of the 2010 HUPO Congress, Deutsch presented data from Laura Beretta, Tadashi Yamamoto, and Gil Omenn from preliminary cross-proteome analyses for the HUPO liver, kidney, urine, and plasma initiatives. These and other HUPO initiatives are one of the foundations for the biology and disease-driven component of the Human Proteome Project (B/D-HPP) [8].

Richard Lipscombe of Proteomics International and the Centre for Food and Genomic Medicine in Perth, Western Australia, spoke about ''*diabesity*,'' a term that combines the twin epidemics of diabetes and obesity, and is considered to be a ''metabolic time bomb.'' In the last decade, numbers of diabetics worldwide have doubled and are projected to grow from 250 to 380 million by 2015. Plasma biomarkers are being sought for prediction of early-onset of diabesity and its complications, and ultimately to facilitate discovery of new therapeutic targets. Ten cohorts representing diverse Western Australian populations have been studied, including both children and adults with either type 1 or type 2 diabetes or who are obese. This group also has gained access to the Busselton Health Study (16 000 participants) and the Fremantle Diabetes Study (1426 participants). Plasma samples from selected subgroups of various cohorts have been quantified using iTRAQ following immunodepletion, analysis by LC-MALDI-MS/MS, and then validation. Candidate biomarkers using an orthogonal MRM approach. Between 200 and 300 proteins per cohort were identified; across all studies 60 proteins showed significant differences in expression when compared to controls, of which two-third were new associations.

Richard Smith of the Pacific Northwest National Laboratory (PNNL) in Richland, Washington, USA, described several emerging proteomics technologies that are highly promising. These discovery-based tools are able to measure thousands of proteins in plasma and other complex biological samples and represent a technical breakthrough. Success to date in establishing new clinical biomarkers has been limited, in general, despite such significant technological advances and large investments in biomarker discovery efforts. One explanation is that current proteomic technologies have not provided sufficient sensitivity for broad coverage of the proteome in combination with the throughput required for confident biomarker discovery applications. This is further complicated by the extent of human biological variation, the large dynamic range of protein abundances in blood, and the sensitivity needed to detect many clinically relevant analytes that are expected to be present in low ng/mL to pg/mL concentrations. As part of the NIH-supported Proteomics Research Center at PNNL, Dr. Smith's group has developed a nano-liquid chromatography-ion mobility MS proteomics platform that provides significantly improved sensitivity and a throughput of >50 samples per day. Examples were provided that illustrated the best currently available platforms with a variety of discovery and verification studies.

Bernd Wollscheid of the Institute for Molecular Systems Biology at ETH-Zurich presented platforms for sensitive, reproducible MS-based measurements, primarily an update on SRM platforms. N-Glycosylated proteins represent a clinically interesting subproteome which by SRM can achieve low ng/mL limits of detection without prior sample fractionation. Dr. Wollscheid presented data on a new library of verified SRM assays and associated tools for 7000 N-glycosites that are mapped to 3298 human proteins [9]. The parameters for these SRM assays are available through the new public resource freely accessible through the SRMAtlas web-repository. SRMAtlas will enable faster design and implementation of quantitative SRM-based experiments by providing tested SRM peptide measurement coordinates for a wide range of clinical research questions. Examples were provided on the

quantification of 53 selected glycoproteins spanning almost seven orders of magnitude in concentration in sera from 32 individual donors.

We anticipate that the progress from such studies of the human plasma proteome will be a good foundation for the HPP and many other basic and clinical applications.

The authors thank all the speakers and participants in these workshops, Peipei Ping and Jennifer Van Eyk for providing summaries of the Ning and Van Eyk presentations, and Sylvie Ouellette for coordinating communications with all the HUPO initiatives from HUPO HQ.

The authors have declared no conflict of interest.

References

[1] Omenn, G. S., States, D. J., Adamski, M. R., 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 3020 proteins and a publicly-available database. *Proteomics* 2005, *5*, 3226–3245. [Also, Omenn G. S. (Ed.), *Exploring the Human Plasma Proteome*, Wiley-Liss, Weinheim, Germany 2006, p. 372.]

[2] Farrah, T., Deutsch, E. W., Omenn, G. S., Campbell, D. S. et al., A high-confidence human plasma proteome reference set with estimated concentrations in PeptideAtlas. *Mol. Cell Proteomics* 2011. DOI: 10.1074/mcp.M110.006353.

[3] Picotti, P., Lam, H., Campbell, D., Deutsch, E. et al., Database of validated assays for the targeted mass spectrometric analysis of the *S.cerevisiae* proteome. *Nat. Methods* 2008, *5*, 913–914.

[4] Menon, R., Omenn, G. S., Proteomic characterization of novel alternative splice variant proteins in human epidermal growth factor receptor 2/neu-induced breast cancers. *Cancer Res.* 2010, *70*, 3440–3449.

[5] Fu, Z., Wang, M., Gucek, M., Zhang, J. et al., Milk fat globule protein-epidermal growth factor-8: a pivotal relay element within the angiotensin II and monocyte chemoattractant protein-1 signaling cascade mediating vascular smooth muscle cells invasion. *Circ. Res.* 2009, *104*, 1337–1346. DOI: 10.1161/CIRCRESAHA.108.187088.

[6] Matt, P., Schoenhoff, F., Habashi, J., Holm, T. et al., The GenTAC consortium. *Circulation* 2009, *120*, 526–532. DOI: 10.1161/CIRCULATIONAHA.108.841981.

[7] Kuzyk, M. A., Smith, D., Yang, J., Cross, T. J. et al. Multiple reaction monitoring-based, multiplexed, absolute quantitation of 45 proteins in human plasma. *Mol. Cell Proteomics* 2009, *8*, 1860–1877.

[8] Legrain, P., Aebersold, R., Archakov, A., Bairoch, A. et al., The human proteome project: current state and future direction. *Mol. Cell Proteomics* 2011. DOI: 10.1074/mcp.M111.009993-1.

[9] Picotti, P., Rinner, O., Stallmach, R., Dautel, F. et al., High-throughput generation of selected reaction-monitoring assays for proteins and proteomes. *Nat. Methods* 2010, *7*, 43–46.

