Advances of the HUPO Human Proteome Project with broad applications for life sciences research

Gilbert S. Omenn

Departments of Computational Medicine & Bioinformatics, Internal Medicine, Human Genetics, and School of Public Health, University of Michigan, Ann Arbor, MI, USA

Keywords

The HUPO Human Proteome Project (HPP) was prominently featured in the 15th annual HUPO World Congress of Proteomics in Taiwan on 18–22 September 2016 as part of a very strong program of keynote and plenary lectures, awards lectures, training workshops, and diverse scientific sessions (hupo.org/hupo2016).

The HPP has two overriding goals: (1) making proteomics broadly complementary to genomics in life sciences and biomedical research and (2) progressively completing the protein parts list with credible evidence for at least one expressed protein product from each predicted protein-coding gene and characterizing the functions of the protein and its structural variants, splice variants, and post-translational modifications. Completion of this global project will enhance understanding of human biology at the cellular level and lay a foundation for development of diagnostic, prognostic, therapeutic, and preventive medical applications.

The Biology and Disease-driven component of the HPP, now led by Jennifer van Eyk and Fernando Corrales, has delivered three important initiatives. First is the dramatic development of Selective Reaction Monitoring (SRM) for targeted proteomics [1,2] under leadership from Ruedi Aebersold at ETH-Zurich and Rob Moritz at the Institute for Systems Biology in Seattle. The latest work, published in July in Cell [3], is a comprehensive SRM Atlas with 166K uniquely mapping (proteotypic) peptides for >99% of the predicted human proteins with matching spectral libraries, expected transitions, and availability of labeled synthetic peptides. Its utility in examining the network response to inhibition of cholesterol synthesis in liver cells and to docetaxel in prostate cancer lines is amply demonstrated.

The second is a strategy of identifying lists or panels of proteins that can be recommended to the relevant research communities for organ-specific or disease-associated studies. The initial approach was to have the proteomics specialists post in the HPP website lists of

'priority proteins' for diabetes, aortic disease, and breast, ovarian, and colon cancers amenable to quantitation with proteomics. The latest approach is bibliometric, capturing evidence for the top-50 proteins published by researchers on various organ systems, beginning with cardiovascular, cerebral, hepatic, intestinal, pulmonary, and renal systems [4]. SRM methods are directly applicable for sensitive and quantitative analysis of these 'popular proteins' well known to researchers of these organ systems; the approach can readily be expanded to other disease categories for studies of normal biology and physiology and pathological processes. This approach to cardiovascular diseases was demonstrated by Jennifer van Eyk in her keynote address in Taipei. There are now 22 Biology and Disease teams who are advised to apply this approach [5].

The third development is a passionate engagement of Early Career Researchers (ECR). For the annual HUPO Congress, the B/D-HPP leadership and ECR leaders organized an all-day Mentoring Program, a manuscript competition, and a travel awards program. For the past year, there has been an ECR representative on the B/D-HPP Executive Committee.

The KnowledgeBase resource pillar of the HPP, led by Eric Deutsch, brings together the HUPO Protein Standards Initiative and the ProteomeXchange created to link data flows from investigators with standardized reanalysis of the raw data and metadata by PeptideAtlas and by GPMdb and curation by neXtProt. A popular feature at the two most recent HUPO Congresses is the Bioinformatics Hub that operated every day to provide drop-in consultations and scheduled deep dives on missing proteins, glycoproteomics, and other nominated topics. A new web resource (missingproteins.org) with structured information from the literature was released at this Congress by the Australian team led by Mark Baker.

Another major component is the Chromosome-centric C-HPP, led now by Young-Ki Paik, Lydie Lane, and Chris Overall, with 25 teams around the world tackling individual chromosomes plus mitochondria to identify and characterize proteins whose coding genes are on the respective chromosome. An analogy with the Human Genome Project, this feature distributed the enormous proteome-wide task and addresses chromosome-specific properties like amplicons and cis-regulatory phenomena. There are 19,467 protein-coding genes according to neXtProt (excluding 588 dubious/uncertain genes, labeled PE5). During the past four years of the HPP, the field has progressed from 13,664 PE1 proteins (based on highly confident protein evidence) in 2013 to 15,646 in 2014 to 16,491 in 2015 to 16,518 in the 2016-02 version of neXtProt. The latest increment was reduced due to the implementation of more stringent Mass Spectrometry (MS) Data Interpretation Guidelines v2 [6]; of 485 proteins that would have been PE1 under the earlier v1 Guidelines, 432 were reclassified as PE2 (transcript evidence only), 46 as PE3 (homology, i.e. identified in non-human species), and 7 as PE4 (based on gene models).

The 2290 PE2 predicted proteins have tissue-based transcript expression data that guide us where to search for evidence of protein expression. This strategy was implemented admirably this year by the Chromosome 2/Chromosome 14 consortium focusing on the large number of genes (879) with testis-specific or testis-enriched expression, as shown by the Human Protein Atlas, the basis of the HPP Antibody Profiling component [7]. From studies of testis and sperm, 253 previously missing proteins were reported as compliant with the

HPP Guidelines v2 [8,9]. Once these (and other) findings have been subjected to reanalysis by PeptideAtlas and incorporated into the 2017-02 version of neXtProt, we project that the number of PE1 proteins will rise to at least 16,873 (87% of the 19,467 predicted proteins) and the number of missing proteins (still PE2,3,4) will decrease from 2949 to 2696 [10,11].

At the post-Congress HPP Workshop at beautiful Sun Moon Lake, there was extensive strategic discussion about accelerating progress on identifying and validating the remaining missing proteins, enhanced by HPP Scientific Advisory Board members Cathy Costello, John Yates, and Naoyuki Taniguchi and by comments from a HUPO-wide survey that had stimulated 162 responses. Besides the focus on known sites of significant transcript expression, enhanced methods for sample preparation and protein solubilization will increase the yield for membrane-embedded proteins that tend to be highly hydrophobic with few tryptic sites, and more sensitive mass spectrometry instruments will help detect proteins with low abundance. Informed choice of the tissue specimen, consideration of life stages, search for induction of protein expression (such as beta-defensins) by infection or inflammation or by pharmacological stimulation may also help. The C-HPP announced a Top-50 Challenge for a coordinated effort across every chromosome to identify the most tractable ~50 predicted proteins for detection, combined with use of SRM or other methods for confirmation.

The strategic discussion also recognized the many reasons why proteins may be undetectable by mass spectrometry. Of the 16,518 PE1 proteins as of neXtProt v2016-02, 14,629 were based on PeptideAtlas for mass spectrometry, while 1860 were based on a variety of non-MS methods, including three-dimensional structures, sequencing results, biochemical studies of many kinds, and mutational analyses [10]. Present MS methods will not detect proteins of very low abundance (which include those with inaccessible chromatin or very low transcript expression), proteins lacking lysine and arginine residues such that tryptic digestion cannot yield peptides of 9–50 aa in length, proteins in highly homologous protein families such that the detected peptides cannot be distinguished, or proteins not solubilized from membranes. Such proteins may justify an adjustment in the denominator for estimating the percentage of expected proteins.

There was also strong sentiment for accelerating the work, long-planned, and now incorporated into the neXtProt, PeptideAtlas, and GPMdb databases, on splice variants, PTMs, N-termini, and sequence variants and their functional and pathological features. This strategy reinforces the decisions to link the B/D-HPP teams with the C-HPP and its clusters focused on cancers, neurodegenerative disorders, reproductive biology, membrane proteins, and the *in vitro* transcription/translation platform.

It was agreed that there will be a 5th annual HPP Special Issue of the *Journal of Proteome Research* in 2017; the Call for Papers was announced by *JPR* on 4 November 2016, with a deadline of 31 May 2017. The *JPR* confirmed that the HPP Guidelines v2.1 and Checklist will apply. We hope that investigators throughout the world will find these guidelines helpful for routine mass spectrometry reporting, especially respecting the protein-level FDR <1%, and for 'extraordinary claims' of detection of previously undetected proteins and/or translation products from long non-coding RNAs (lncRNAs) or pseudogenes or small open

heading frames (ORFs) [6]; (hupo.org/hpp/guidelines). The latter include careful scrutiny of the spectra, use of thresholds of 9 aa in length and 2 uniquely mapping proteins for peptide-to-protein matches, and careful consideration of alternative protein matches, especially to sequence variants or isobaric PTMs of abundant proteins. These more probable matches have been shown to explain many claims of previously unreported proteins [6,12].

Independent researchers from around the world are warmly welcomed to submit their papers that match the broad thematic priorities specified.

Acknowledgments

References

1. Anderson NL, Anderson NG, Pearson TW, et al. A human proteome detection and quantitation project. Mol Cell Proteomics. 2009; 8:883–886. [PubMed: 19131327]

2. Keshishian H, Addona T, Burgess M, et al. Quantitative, multiplexed assays for low abundance proteins in plasma by targeted mass spectrometry and stable isotope dilution. Mol Cell Proteomics. 2007; 6:2212–2229. [PubMed: 17939991]

3. Kusebauch U, Campbell DS, Deutsch EW, et al. Human SRMAtlas: a resource of targeted assays to quantify the complete human proteome. Cell. 2016; 166:766–778. [PubMed: 27453469]

4. Lam M, Xing Y, Cao Q, et al. Data-driven approach to determine popular proteins for targeted proteomics translation. J Proteome Res. 2016 Jul 19.15:4126–4134. [PubMed: 27356587]

5. Van Eyk JE, Corrales FJ, Aebersold R, et al. Highlights of the biology and disease-driven Human Proteome Project, 2015-2016. J Proteome Res. 2016 Sep 20.15:3979–3987. [PubMed: 27573249]

6. Deutsch EW, Overall CM, Van Eyk JE, et al. Human Proteome Project mass spectrometry data interpretation guidelines 2.1. J Proteome Res. 2016 Aug 24.15:3961–3970. [PubMed: 27490519]

7. Uhlen M, Fagerberg L, Hallstrom BM, et al. Proteomics. Tissue-based map of the human proteome. Science. 2015; 347:1260419. [PubMed: 25613900]

8. Vandenbrouck Y, Lane L, Carapito C, et al. Looking for missing proteins in the proteome of human spermatozoa: an update. J Proteome Res. 2016 Aug 23.15:3998–4019. [PubMed: 27444420]

9. Wei W, Luo W, Wu F, et al. Deep coverage proteomics identifies more low-abundance missing proteins in human testis tissue withQ-exactive HF mass spectrometer. J Proteome Res. 2016 Aug 29.15:3988–3997. [PubMed: 27535590]

10. Omenn GS, Lane L, Lundberg EK, et al. Metrics for the Human Proteome Project 2016: progress on identifying and characterizing the human proteome, including post-translational modifications. J Proteome Res. 2016 Sep 20.15:3951–3960. [PubMed: 27487407]

11. Paik YK, Overall CM, Deutsch EW, et al. Progress in chromosome-centric Human Proteome Project as highlighted in the annual special issue IV. J Proteome Res. 2016; 15:3945–3950. [PubMed: 27809547]

12. Omenn GS, Lane L, Lundberg EK, et al. Metrics for the Human Proteome Project 2015: progress on the human proteome and guidelines for high-confidence protein identification. J Proteome Res. 2015; 14:3452–3460. [PubMed: 26155816]