Published in final edited form as:

J Proteomics. 2014 April 4; 100: 3–7. doi:10.1016/j.jprot.2013.10.012.

THE STRATEGY, ORGANIZATION, AND PROGRESS OF THE HUPO HUMAN PROTEOME PROJECT

Gilbert S. Omenn, MD, PhD

Center for Computational Medicine & Bioinformatics, University of Michigan, Ann Arbor, MI, 48109-2218, USA

Gilbert S. Omenn: gomenn@umich.edu

Abstract

The Human Proteome Project is a major, comprehensive initiative of the Human Proteome Organization. This global collaborative effort aims to identify and characterize at least one protein product and many PTM, SAP, and splice variant isoforms from the 20,300 human protein-coding genes. The deliverables are an extensive parts list and an array of technology platforms, reagents, spectral libraries, and linked knowledge bases that advance the field and facilitate the use of proteomics by a much wider community of life scientists. Such enablement will help address the Grand Challenge of using proteomics to bridge major gaps between evidence of genomic variation and diverse phenotypes.

Keywords

HUPO Human Proteome Project; Chromosome-centric Human Proteome Project; Biology and Disease-driven Human Proteome Project; missing proteins; ProteomeXchange; PeptideAtlas; Human Protein Atlas; neXtProt; GPMDB

1.0 INTRODUCTION

The Human Proteome Organization (HUPO) was founded in 2001 just as the Human Genome Sequencing Project was approaching completion. HUPO's goals are to promote the field of proteomics and its applications through international cooperation and research collaborations and by fostering the development of new technologies, techniques, and training [www.hupo.org]. Its annual World Congress of Proteomics and its array of ten productive collaborative research initiatives have addressed those goals. The initiatives included Plasma, Liver, Brain, Kidney/Urine, Cardiovascular, Model Organism, and Stem Cell proteome projects, the Human Antibody Initiative, the Human Glycoproteomics Initiative, and the Protein Standards Initiative. High quality data resources have been created at the Swiss Institute for Bioinformatics (SwissProt, neXtProt), European Bioinformatics Institute (PRIDE, UniProt), Institute for Systems Biology (PeptideAtlas, SRM Atlas), Karolinska Institute (Human Protein Atlas), and Canadian Global Proteome Machine; they are now linked through ProteomeXchange, based at the European Bioinformatics Institute.

© 2013 Elsevier B.V. All rights reserved.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Over the past several years there were many informal and formal discussions about a major overarching Human Proteome Project. Such a project was announced in Sydney, Australia, at the 9th HUPO World Congress in September 2010 and launched at the 10th World Congress in Geneva in September 2011. Substantial progress was presented at the 11th World Congress in Boston in September 2012, and the 12th World Congress in Yokohama, Japan, in September 2013.

The goal is to identify and characterize at least one protein product and a growing number of post-translational modifications (PTMs), single amino acid polymorphisms (SAPs), and splice variant isoforms from the 20,300 human protein-coding genes [1]. Besides that parts list, the HPP will deliver technology platforms, reagents, spectral libraries, and linked knowledge bases that facilitate the use of proteomics by a much wider community of life scientists. The Grand Challenge is to use proteomics to bridge major gaps between evidence of genomic variation and diverse phenotypes [2]. Much of that progress lies in understanding and visualizing the biochemical and signaling pathways that capture environmental and behavioral interactions with genetic predispositions, the roles of protein isoforms and protein-protein interactions [3] and the roles of non-coding RNAs and regulatory features identified with ENCODE, the Encyclopedia of DNA Elements [4].

2.0 THE VISION AND ORGANIZATION OF THE HUMAN PROTEOME PROJECT

By characterizing the protein products from all 20,300 protein-coding genes of the known genome, the Human Proteome Project will generate a map of the protein-based molecular architecture of the human body. The HPP will be a resource to help elucidate biological and molecular function and advance diagnosis, treatment, and prevention of diseases.

Figure 1 shows the organizational vision of the HPP: the resource pillars from mass spectrometry, antibody-based protein capture, and knowledge bases; the chromosome-centric “adopt-a-chromosome” platform; and the biology and disease-driven platform.

As of mid-2013 there are 25 chromosome-centric teams [5] and 16 biology or disease-driven teams [6]. Figures 2A and 2B show these two consortia.

The C-HPP is led by Young-Ki Paik (Korea), Bill Hancock (US), and Gyorgy Marko-Varga (Sweden) [4]; the full roster of its executive committee and the 24 chromosome (and mitochondria) teams and their many activities can be found at www.c-hpp.org. At the CNPN April 2013 meeting, Paul McKeown and Christoph Borchers, respectively, announced the fresh mobilization of Chr 6 and Chr 21 teams in Canada [see this issue].

The B/D-HPP is led by Ruedi Aebersold (Switzerland), Jennifer van Eyk (US), and Jun Qin (China), with its executive committee and 16 teams around the world; details are available at www.thehpp.org.

The HPP organizational chart (Figure 1) also shows the Senior Scientific Advisory Board, chaired by Michael Snyder (US), with Cathy Costello (US), Kunliang Guan (US/China), Denis Hochstrasser (Switzerland), Lee Hood (US), Matthias Mann (Germany), Kate Rosenbloom (US), Naoyuki Taniguchi (Japan), Mathias Uhlen (Sweden), John Yates (US), and the HPP-Executive Committee, comprised of Gil Omenn (US, chair), Ruedi Aebersold (Switzerland), Amos Bairoch (Switzerland), Fuchu He (China), Bill Hancock (US), Emma Lundberg (Sweden), Young-Ki Paik (Korea), and, ex-officio, Pierre Legrain (France, HUPO president). The HPP resource pillars are led by Bruno Domon (Luxembourg, MS), Mathias Uhlen (Sweden, AB), and Lydie Lane (Switzerland, KB).

3.0 PROGRESS OF THE HUMAN PROTEOME PROJECT

3.1 C-HPP and B/D-HPP

The *Journal of Proteome Research* published a special issue in January 2013 organized by the leaders of the C-HPP [5]. There are 33 papers from or related to the Chromosome-centric HPP, including papers from chromosome teams 1, 4, 7, 8, 11, 13, 16, 17, 18, 19, 20, X, and Y and multiple database, technology, and cross-cutting articles. An additional 15 papers that did not make the deadline for the January issue appeared in June. Together these two sets of articles constitute the 2013 virtual C-HPP special issue, <http://pubs.acs.org/page/jprobs/vi/c-hhp.html>. The *Journal* plans a 2014 January special issue, timed to capture new work presented at the Yokohama Congress.

In parallel, the Biology and Disease-driven HPP has emerged, as envisioned by Legrain et al [1]. The pre-existing HUPO proteome projects (see section 1.0) joined the B/D-HPP, and six new project teams on diabetes, cancers, infectious diseases, epigenomics, eye, and autoimmune disorders were launched [6]. Additional project teams are in the early stages of formation. A 10-year timeline for the HPP in two phases of 6 and 4 years was laid out in 2012 [4]. As described below in section 4.0, the broad deliverables will be practical technology platforms, reagents, spectral libraries, and linked knowledge bases that enable many life scientists to utilize proteomics in their research and omics-based clinical practices [6]. The HPP-EC convenes monthly; the leaders of the component units of the HPP have regular conference calls; the C-HPP has held 3–4 meetings of investigators per year; and everyone gathers at the annual Congress.

3.2 Metrics and the Baseline Master Table

We created a Master Table as a baseline for the HPP and specifically for the C-HPP for each chromosome using five standard metrics [5]: Ensembl (v69) provides the number of protein-coding genes; neXtProt (gold), PeptideAtlas (canonical), and GPMDB (green) provide numbers for confidently identified proteins from mass spectrometry studies, with special features for each; and the Human Protein Atlas gives the number of proteins for which polyclonal antibodies generated against one or two different epitopes along the protein sequence have been used to characterize protein expression across 46 cell types, intracellular organelles, and selected cancer cells (with evidence scored at the medium or high levels).

As of December 2012, the numbers across those five resources were 20,059 for Ensembl, 13,664 for neXtProt, 12,509 for Human PeptideAtlas, 14,300 for GPMDB, and 10,794 for Human Protein Atlas. The article explains in considerable depth the special features of these complementary resources [5]. Each resource has provided a chromosome-by-chromosome analysis as part of their engagement with the Human Proteome Project. Updates of these metrics are available at www.c-hpp.org/wiki and at the websites of the individual resources.

neXtProt is a quality-filtered corpus of manually-curated annotations from UniProtKB/Swiss-Prot specifically for human proteins [7]. Entries are displayed from the perspectives of the protein, the underlying gene, and the relevant references. Complex mapping of Ensembl protein sequences to genes and transcripts is performed routinely. All but 125 neXtProt entries display precise genomic coordinates for at least one isoform; only 9 are not assigned to any chromosome. neXtProt has put major emphasis on import of variant and PTM data, which may account for many of the unattributed spectra in mass spectrometry studies. There are 312,000 sequence variants from dbSNP and COSMIC and 8135 PTM sites on 3312 entries for N-glycosylation, phosphorylation, S-nitrosylation, ubiquitination, and sumoylation, with arginine methylation to be added. Splice variants are documented and mapped. Immunohistochemistry data from the Human Protein Atlas and subcellular

localization results from DKFZ GFP-cDNA and the Weizmann Institute Kahn Dynamic Proteomics Database are also integrated [7]. Abundance of transcripts serves as a clue for which tissues or cell types are most likely to express the protein.

PeptideAtlas is a core resource for the Human Proteome Project; its builds contain all relevant datasets for the entire Human Proteome [8] and for the Human Plasma Proteome [9] and other organs and biofluids. The raw spectra are subjected to uniform re-analysis with the Trans Proteomic Pipeline. Farrah et al (2013) extensively compared the human proteins identified with <1% false-positive rate for 12,629 protein-coding genes with the ~7500 proteins not yet seen in the PeptideAtlas [8]; [10]].

The GPMDB has compiled comprehensive lists of all human protein phosphorylation sites, lysine-acetylation sites, and N-terminal-acetylation sites represented by good quality data in GPMDB. This list has been subdivided on a chromosome-by-chromosome basis.

3.3 “Missing proteins”

There are many reasons why protein entries in SwissProt lack evidence of protein expression in our knowledge bases and, therefore, the baseline Metrics Table: (1) proteins expressed significantly only in unusual organs or cell types, perhaps particular brain regions, nasal epithelium/olfactory cortex, testis, and placenta, among ~230 cell types [7]; (2) proteins expressed only in early developmental stages, especially embryonic and fetal; (3) given the criteria of PeptideAtlas, neXtProt, and GPMDB for selecting only one “representative protein” among highly homologous members of protein families, large numbers of proteins must be missed or not counted from families of cytokeratins, immunoglobulins, histocompatibility antigens, and olfactory receptors, as well as broader families of kinases, phosphatases, and membrane-embedded G-protein coupled receptors (GPCR); (4) the abundance/ concentration of many proteins is surely below our present limits of detection, due to low rates of synthesis or rapid degradation, or both; and (5) neXtProt evidence category 5 contains 638 “uncertain” or “dubious” genes, which may need to be removed from the denominator (Amos Bairoch, Lydie Lane, Gil Omenn, personal communication, Yokohama Congress).

As summarized elsewhere [Omenn GS, in press], evidence from the Human Protein Atlas, human embryonic and induced pluripotent stem cells, and 11 cell lines of different lineages shows only quite limited numbers of tissue-specific or developmental-specific proteins [11].

In the most recent update of the Human Protein Atlas, Fagerberg et al [11] reported antibody-profiling and RNA-Seq data for a total of 11 cell lines; 13,971 proteins were identified with high or medium confidence by immunochemistry (69% of protein-coding genes) and an additional 22% of protein-coding genes had measurable transcripts. There were relatively few cell or tissue-specific transcripts (a total of 928 or <100 per cell line); 10,078 were identified in all 11 cell lines versus 2798 not detected in any of the 11. Some examples of quite specific expression were noted, including TEX1010 and ARGHGAP28 in spermatocytes and spermatids in the testis.

Mann et al provided a comprehensive perspective on progress toward “complete, accurate, and ubiquitous proteomes” [12]. Progress with chromatography and online peptide analysis, combined with far faster, more sensitive, and more accurate mass spectrometers that can handle a much greater dynamic range yields deep, quantitative proteomes. Sophisticated algorithms now process much larger datasets in a completely automated manner, with quantitative and statistical rigor. Single 4h gradient runs and advanced MS instruments can reliably identify about 10,000 different human proteins [13] in cultured cells. The filtered transcriptome (RK PM ≥ 1) and the proteomes of these cell lines both identify about 10,000

to 12,000 protein-coding loci [12], [14]. The character of individual tissues seems to depend on the quantitative levels and interactions of these proteins, rather than presence or absence of the protein. Studies of primary tissues are beginning to appear, such as 7500 proteins from colon cancer specimens [15]. Body fluid proteomes remain more complex, due to higher proportions of proteins with very low abundance [12] secreted or released from many different tissues.

3.4 Dataset Submission

The HPP investigators have committed to open and timely sharing of datasets and metadata. Fortunately, a repository coordination system has emerged from efforts of the past five years to create the ProteomeXchange, based at the European Bioinformatics Institute in England [Vizcaino JA, et al, submitted].

Datasets registered with ProteomeXchange (PX) are made available through curation at EBI/PRIDE and Swiss Institute for Bioinformatics (SIB)/neXtProt and are downloaded and re-analyzed by Peptide Atlas and by the Global Proteome Machine Database (GPMDB). The basic workflow is as follows: a) a dataset is submitted by a member of the community to PRIDE at EBI for MS/MS or SRM Atlas/PASSEL at Institute for Systems Biology for SRM data as the receiving data repositories; b) the receiving repository requests a unique ProteomeXchange identifier from the ProteomeCentral service; c) the receiving repository tracks the dataset through the journal review process; d) upon acceptance by the journal, the receiving journal prepares a dataset announcement in the form of an XML document and submits it to ProteomeCentral; and e) ProteomeCentral archives the document and transmits an announcement to the ProteomeXchange RSS feed, which is open to all interested parties. ProteomeXchange has assured export of datasets from PRIDE using mzXML and import of datasets from PX into PeptideAtlas and by GPMDB.

The shotgun component of PeptideAtlas (as contrasted with the SRM component) starts from spectra as raw data from shotgun MS/MS experiments and reprocesses them using the PeptideAtlas pipeline, which includes validation of new search results with the Trans-Proteomic Pipeline. The Human PeptideAtlas was updated through 2012 [8]. At present, only the PRIDE database is designated to accept MS/MS experiments for the ProteomeXchange consortium. The intended workflow is then that the datasets deposited into PRIDE are announced via ProteomeXchange, and then PeptideAtlas downloads suitable ones, reprocesses them according to its pipeline, and then adds them to its builds. Other repositories, like GPMDB and other repositories or community consumers may do the same.

Several recent human proteomics datasets (PXDs 230, 239, 284, and 290) have been submitted, given a PDX registration number, been downloaded by Peptide Atlas, been reprocessed with TransProteomicPipeline, and been made available to the public [www.ProteomeXchange.org].

4.0 Enablement of the Broader Life Sciences Research Community

The overriding goal for the HPP, beyond building the parts list of proteins and their isoforms, interactions, and functions, is to enable the much broader scientific community, including clinical investigators, to use proteomics methods as part of an integrated omics strategy to link genome and phenotype through pathways, modules, networks, and regulatory mechanisms [6]. Quantification of protein abundances and of the myriad products of protein enzymes is necessary to drive that strategy. Too many scientists still rely on Western blot and ELISA for their very limited studies of proteins. The B/D-HPP will provide all researchers priority lists of proteins and the attendant reagents and data for studies of functional modules in particular organs and diseases. This approach should help

overcome the well-recognized problem that the majority of publications remain focused on a relatively small set of human proteins [16].

Results from targeted proteomics with selected reaction monitoring/multiple reaction monitoring (SRM/MRM) now are shared through the SRMAtlas at the Institute for Systems Biology and PASSEL, the PeptideAtlas SRM Experiment Library; the SRM peptide reagents, spectral libraries, and most informative transitions from data-independent acquisition (DIA) mass spectrometry are now public for proteins from nearly all human protein-coding genes. Polyclonal antibodies and immunohistochemical findings are available for 13,985 proteins from the Human Protein Atlas [11]. The B/D-HPP and the MS pillar committee are working with manufacturers to encourage robust, simple-to-operate mass spectrometers for use by the non-expert community in clinical laboratories, as was highlighted at the 2013 HUPO World Congress in Yokohama alongside many other sessions presented by the HPP.

References

1. Legrain P, Aebersold R, Archakov A, Bairoch A, Bala K, Beretta L, et al. The Human Proteome Project: current state and future direction. *Mol Cell Proteomics*. 2011; 10:M111 009993. [PubMed: 21742803]
2. Hood LE, Omenn GS, Moritz RL, Aebersold R, Yamamoto KR, Amos M, et al. New and improved proteomics technologies for understanding complex biological systems: addressing a grand challenge in the life sciences. *Proteomics*. 2012; 12:2773–83. [PubMed: 22807061]
3. Vidal M, Chan DW, Gerstein M, Mann M, Omenn GS, Tagle D, et al. The human proteome - a scientific opportunity for transforming diagnostics, therapeutics, and healthcare. *Clinical Proteomics*. 2012; 9:6. [PubMed: 22583803]
4. Paik YK, Jeong SK, Omenn GS, Uhlen M, Hanash S, Cho SY, et al. The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. *Nature Biotechnology*. 2012; 30:221–3.
5. Marko-Varga G, Omenn GS, Paik YK, Hancock WS. A first step toward completion of a genome-wide characterization of the human proteome. *J Proteome Res*. 2013; 12:1–5. [PubMed: 23256439]
6. Aebersold R, Bader GD, Edwards AM, van Eyk JE, Kussmann M, Qin J, et al. The biology/disease-driven human proteome project (B/D-HPP): enabling protein research for the life sciences community. *J Proteome Res*. 2013; 12:23–7. [PubMed: 23259511]
7. Gaudet P, Argoud-Puy G, Cusin I, Duek P, Evalet O, Gateau A, et al. neXtProt: organizing protein knowledge in the context of human proteome projects. *J Proteome Res*. 2013; 12:293–8. [PubMed: 23205526]
8. Farrah T, Deutsch EW, Hoopmann MR, Hallows JL, Sun Z, Huang CY, et al. The state of the human proteome in 2012 as viewed through PeptideAtlas. *J Proteome Res*. 2013; 12:162–71. [PubMed: 23215161]
9. Farrah T, Deutsch EW, Omenn GS, Campbell DS, Sun Z, Bletz JA, et al. A high-confidence human plasma proteome reference set with estimated concentrations in PeptideAtlas. *Mol Cell Proteomics*. 2011; 10:M110 006353. [PubMed: 21632744]
10. Omenn GS, Menon R, Zhang Y. Innovations in proteomic profiling of cancers: Alternative splice variants as a new class of cancer biomarker candidates and bridging of proteomics with structural biology. *Journal of Proteomics*. 2013; 90:28–37. [PubMed: 23603631]
11. Fagerberg L, Oksvold P, Skogs M, Algenas C, Lundberg E, Ponten F, et al. Contribution of antibody-based protein profiling to the Chromosome-centric Human Proteome Project (C-HPP). *J Proteome Res*. 2013; 12:2439–48. [PubMed: 23276153]
12. Mann M, Kulak NA, Nagaraj N, Cox J. The coming age of complete, accurate, and ubiquitous proteomes. *Molecular Cell*. 2013; 49:583–90. [PubMed: 23438854]
13. Beck M, Schmidt A, Malmstroem J, Claassen M, Ori A, Szymborska A, et al. The quantitative proteome of a human cell line. *Mol Syst Biol*. 2011; 7:549. [PubMed: 22068332]

14. Hebenstreit D, Fang M, Gu M, Charoensawan V, van Oudenaarden A, Teichmann SA. RNA sequencing reveals two major classes of gene expression levels in metazoan cells. *Mol Syst Biol.* 2011; 7:497. [PubMed: 21654674]
15. Wisniewski JR, Ostasiewicz P, Dus K, Zielinska DF, Gnad F, Mann M. Extensive quantitative remodeling of the proteome between normal colon tissue and adenocarcinoma. *Mol Syst Biol.* 2012; 8:611. [PubMed: 22968445]
16. Edwards AM, Isserlin R, Bader GD, Frye SV, Willson TM, Yu FH. Too many roads not taken. *Nature.* 2011; 470:163–5. [PubMed: 21307913]

Highlights

- The global Human Proteome Project (HPP) aims to characterize the proteins from all protein-coding genes.
- The Chromosome-centric HPP component (C-HPP) has 24 chromosome teams (plus one for mitochondria).
- The Biology and Disease-driven HPP (B/D-HPP) now has 16 project teams, enabling a broad range of research.
- The HPP baseline Master Table has ~13,000 confidently-identified proteins.
- The Journal of Proteome Research published a 2013 C-HPP special issue with a total of 48 articles.

Figure 1.

Organizational chart for the Human Proteome Project showing the Chromosome-centric HPP, Biology and Disease-driven B/D-HPP, the resource pillars from mass spectrometry (MS), antibody-based protein capture (AB), and knowledge bases (KB), the HPP Executive Committee (EC), HPP Senior Scientific Advisory Board (SSAB), and the Principal Investigator Councils (PIC).

Figure 2a

Figure 2b

Figure 2.

The teams for the Chromosome-centric C-HPP (2A) and the Biology and Disease-driven B/D-HPP (2B) are shown here. Details about the members of the teams and the activities of each can be found at www.thehpp.org and www.c-hpp.org. [From Young-Ki Paik on behalf of C-HPP for Figure 2A; from reference 6, Aebersold et al, Journal of Proteome Research, with permission, for Figure 2B.]