Progress on the Draft Human Proteome
2018 Metrics of HUPO's Human Proteome Project

Report to the HUPO Council from the HPP Executive Committee, August 2018

Gilbert S. Omenn (Chair), Mark S. Baker (Chair-elect), Fernando J. Corrales, Eric W. Deutsch, Lydie Lane, Siqi Liu, Chris Overall, Young-Ki Paik, Jochen Schwenk, Sue Weintraub, Jennifer van Eyk, and Michael Snyder

ABSTRACT: The Human Proteome Organization (HUPO) Human Proteome Project (HPP) continues to make progress on its two overall goals: (1) completing the protein parts list, with an annual update of the HUPO Draft Human Proteome, and (2) making proteomics an integrated complement to genomics and transcriptomics throughout biomedical and life sciences research. neXtProt version 2018-01-17 has 17,470 confident protein identifications (Protein Existence [PE] level 1) that are compliant with the HPP Guidelines v2.1 (https://hupo.org/Guidelines), up from 13,975 in 2012-12, 16,518 in 2016-04, and 17,008 in 2017-01. Remaining to be found by mass spectrometry and other methods are 2579 "missing proteins" (PE2+3+4), down from 5511 in 2012-12, 2949 in 2016, and 2579 in 2017. PeptideAtlas 2018-01 has 15,798 canonical proteins, up 625 in the past year and accounting for nearly all of the 16,092 PE1 proteins based on MS data. These resources have extensive data on PTMs, single amino acid variants, and splice isoforms. The Human Protein Atlas has released its Cell Atlas, Pathology Atlas, and updated Tissue Atlas, and is applying recommendations from the International Working Group on Antibody Validation. Organ-specific popular protein lists based on bibliometric analyses are now available for each of the 22 B/D-HPP biology and disease teams. Quantitative targeted proteomics using SRM-MS or DIA-SWATH-MS offers much promise for studies of biology and disease. For a complete report, see Omenn GS, et al, *J Proteome Res* 2018, DOI: 10.1021/acs.jproteome.8b00441. PMID: 30099871.

METRICS OF PROGRESS: The Human Proteome Project (HPP) of the Human Proteome Organization (www.hupo.org) has provided a framework for international communication, collaboration, quality assurance, data sharing, and acceleration of progress in the global proteomics community since its announcement in 2010 and launch in 2011. The parts list starts with at least one HPP Guidelines-compliant identification of a protein product matching the predicted sequences and expands through detection and characterization of the functions of splice variants, sequence variants, post-translational modifications, and protein-protein interactions. The HPP Guidelines for MS Data Interpretation and the HPP Data Workflow involving ProteomeXchange have been adopted widely. The HPP has 50 research teams worldwide organized by chromosome, mitochondria, biological processes, and disease categories plus resource pillar groups for affinity-based protein capture, mass spectrometry, knowledge bases, and, during the past year, pathology. ProteomeXchange had 5248 publicly released datasets, of which 2165 are from human samples (http://proteomecentral.proteomexchange.org/) as of 2018-05-15. In 2016, the C-HPP announced a Missing Proteins Challenge for each Chromosome team to detect and credibly identify ~50 neXtProt PE2,3,4 "missing proteins". In 2017, a second initiative was launched to annotate the functions of 1260 PE1 proteins lacking such information. Table 1 shows the progression of highly confident protein identifications in neXtProt and in PeptideAtlas during the course of the HPP from 2012 to 2018.

Table 1. neXtProt protein existence evidence levels from 2012 to 2018 showing progress in identifying PE1 proteins and PeptideAtlas canonical proteins. More stringent guidelines imposed in 2016 caused adjustments between 2014 and 2016.

PE Level	Feb 2012	Sept 2013	Oct 2014	April 2016	Jan 2017	Jan 2018
1: Evidence at protein level	13,975	15,646	16,491	16,518	17,008	17,470 [a]
2: Evidence at transcript level	5205	3570	2647	2290	1939	1660
3: Inferred from homology	218	187	214	565	563	452
4: Predicted	88	87	87	94	77	74
5: Uncertain or dubious	*622*	*638*	*616*	*588*	*572*	*574*
Human PeptideAtlas canonical proteins	12,509	13,377	14,928	14,629	15,173	15,798

} 2186 Missing Proteins[b]

[a] Percent of predicted proteins classified as PE1 by neXtProt = PE1/PE1+2+3+4 = 89%.

[b] Missing Proteins PE 2+3+4 = 2186, down from 2579 in neXtProt v2017-01.

DETAILS OF neXtProt MISSING PROTEINS FINDINGS IN PAST YEAR: Figure 1 provides a detailed depiction of the changes within neXtProt by PE (Protein Evidence) categories during the year from release 2017-01-23 to release 2018-01-17. There were 91 proteins added to neXtProt, due to their addition to UniProtKB/Swiss-Prot; 44 were qualified as PE1, while 40 were added as PE2,3, or 4 MPs, and 7 as PE5. Among the new PE1 proteins are four sORFs with strong proteomics evidence; a recently-characterized LINC00961-encoded SPAR polypeptide; and three chimeric proteins with biological activity). Further curation led to deletion of 10 PE1, 3 PE2,3,4 missing proteins, and 7 PE5 dubious entries. The net effect was expansion of neXtProt entries by 71 proteins. The biggest change by far is the movement of 431 PE2,3,4 proteins into PE1, due to additional MS evidence reflected in the large increase of canonical proteins in PeptideAtlas and other high quality data at protein level. ZNF804A, GAGE12G and CEACAM19 were previously validated as PE1 due to PPI and characterization data, but these data have been removed from UniProtKB/Swiss-Prot, resulting in their downgrade to MP status. Finally, four entries were lifted from PE5 status to PE2,3,4, while six MPs were downgraded to PE5 status; none moved from PE5 to PE1.

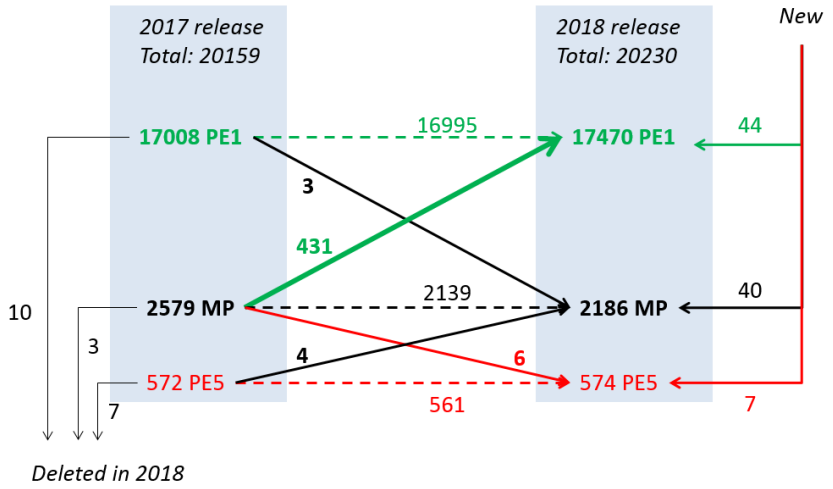

Figure 1. This flow chart depicts the changes in neXtProt PE1-5 categories from release 2017-01-23 to release 2018-01-17. There are 431 missing proteins promoted to PE1 and 44 new SwissProt proteins added as PE1 proteins, while 3 PE1 proteins were demoted to PE2,3,4 MPs and 10 PE1 proteins were deleted altogether.

UniProtKB/Swiss-Prot and neXtProt integrate manually curated protein-protein interaction (PPI) data from the IntAct database. These data are primarily based on Yeast Two Hybrid (Y2H) methods, supplemented by affinity purification/mass spectrometry (AP/MS), phage display, and co-immunoprecipitation. To validate a protein as PE1 based on PPI data, UniProtKB/Swiss-Prot and neXtProt use a subset of PPI data from IntAct. This subset is built by IntAct using a scoring system detailed at https://www.ebi.ac.uk/intact/pages/faq/faq.xhtml#4 with weighting of different kinds of experimental evidence. Currently, 530 PE1 protein validations are based on PPI, up from 372 a year ago. Y2H methods utilize artificially-expressed bait proteins to detect protein interaction partners. Generally, this approach does not identify the tissues of expression or guide researchers to a choice of biological specimens to study; however, preys may be selected from a library of transcripts expressed in a particular tissue.

Using the neXtProt "Interactions" view, https://www.nextprot.org/entry/NX_O60479/interactions, users can look for the Gold Protein-Protein interactions for each of the proteins of interest to see the number of experiments behind each interaction. Clicking on the "evidence" button will link to the IntAct page with details about the experimental datasets used.

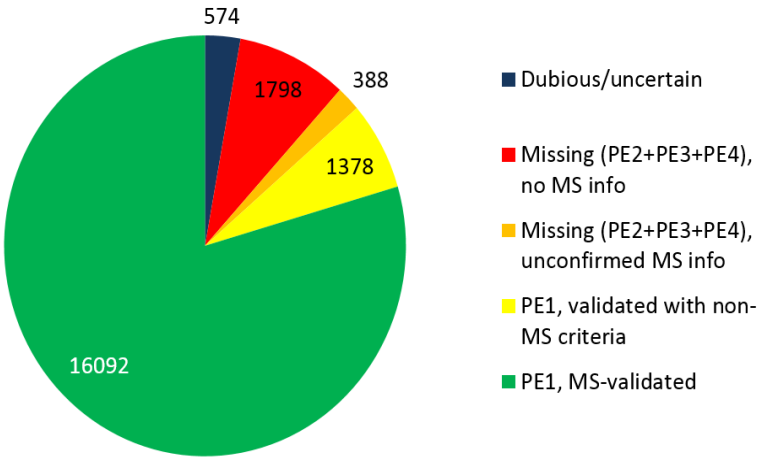

Figure 2. Identified and predicted proteins by PE level in neXtProt release 2018-01-17.

The pie chart (Figure 2) shows the nature of the evidence data for PE1 proteins, as well as the numbers in other categories as of neXtProt release 2018-01-17. There are 16,092 PE1 proteins identified with MS data compliant with HPP Guidelines, of which 98% are canonical in PeptideAtlas. There are 1,378 additional PE1 proteins identified with other kinds of protein evidence: 99 by Edman sequencing, 176 by disease mutations, 75 from 3D structures, 530 by protein-protein interactions, 58 with Ab-based techniques, 170 from PTMs or proteolytic processing, and 270 from other biochemical studies. PE2,3,4 missing proteins are divided into those with no MS data (1,798) and those with insufficient or unconfirmed MS data (388, down from 453 in 2017 and 485 in 2016), primarily due to the application in 2016 of the more stringent HPP Guidelines[7] for accepting MS-based identifications. Nearly 100 of those excluded in 2016 have been restored as additional data have been reported and reviewed.

EXPANSION OF PEPTIDEATLAS FROM 2017 to 2018: The canonical proteins in PeptideAtlas increased from 15,173 to 15,798 during the year from v2017-01 to v2018-01. Of the 40 MS datasets that were added to the PeptideAtlas Human Build during 2017, eight provided 528 of the 625 proteins that were

newly validated as canonical (Figure 3). Often these datasets and publications provided a second uniquely mapping non-nested peptide with length ≥9 amino acids, raising the corresponding protein to canonical status. In some cases, the datasets provided both peptides. The greatest contributions came from studies that enriched for proteins that had not been well-represented previously, including SUMOylated proteins, membrane proteins, and zinc finger proteins. PeptideAtlas would contain 273 fewer canonical proteins if PTM-containing peptides were excluded. PTMs are the primary focus of the HPP MS resource pillar, including an ongoing community project with a specially-prepared sample of 96 phosphopeptides. Both neXtProt and PeptideAtlas have growing content of PTMs. A major advance is the introduction of MSFragger for ultra-fast identification of post-translational and chemical modifications of peptides.

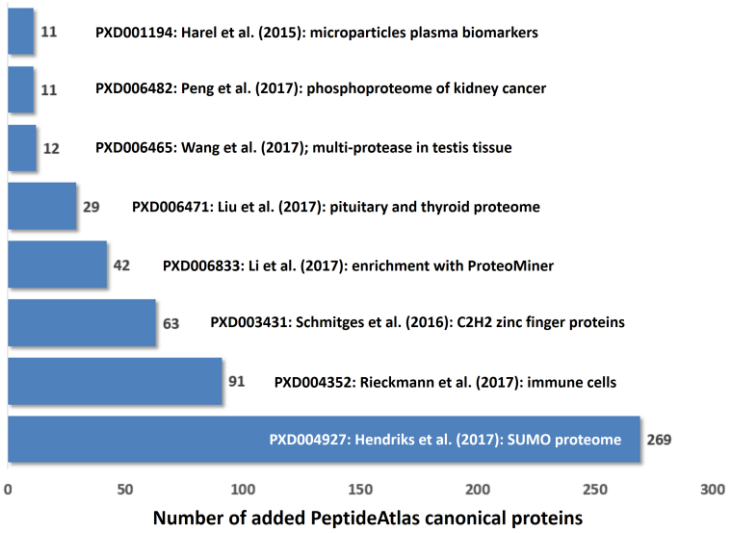

Figure 3. These eight datasets added to PeptideAtlas in 2017 provided the evidence needed to raise the PeptideAtlas protein category to "canonical" for more than 10 proteins each. Canonical status requires two or more uniquely-mapping non-nested peptides with length ≥9 residues with high-quality spectra, not accounted for by sequence variants or isobaric PTMs in other proteins. PXD identifiers refer to ProteomeXchange and are required any the deposition of new MS data by many journals, including the 2018 HPP Special issue in JPR.

THE FATE OF MISSING PROTEINS NOMINATED FOR neXtProt REVIEW IN THE JPR 2017 PAPERS: The editorial for the JPR 2017 HPP special issue highlighted six papers that used a variety of promising methods to find Missing Proteins, with a total of 32 identified for validation as PE1: 15 using Triton X-100 solubilization plus ProteoMiner hexapeptide-covered beads as an enrichment/equalization strategy for low-abundance proteins and confirmation with PRM; 12 from the sperm proteome; 3 using a multi-protease strategy on testis; 1 from the kidney phosphoproteome; and 1 based on biological studies of a Y chromosome protein in cardiac development; plus 41 proposed from the Yamamoto lab using a "stranded peptides" approach. It is feasible to identify stranded peptides in major databases that could be combined to make a pair of proteotypic peptides for individual missing proteins and then use the reported spectra from the original work (if raw data are available) to compare with the spectra available in SRMAtlas for the corresponding synthetic peptide. This work will be facilitated by development of a Universal Spectrum Identifier, soon to be released by PeptideAtlas and PSI. In brief, through various paths, 43 of the 73 MP candidates recommended from the 2017 JPR special issue had by 2018-01 qualified as PE1 by neXtProt. Meanwhile, work from the entire proteomics community has raised 952 MPs to PE1 status since neXtProt release 2016-01-11 (see Table 1). It takes a village!

RECOGNIZING THE LIMITATIONS OF FINDING PE2,3,4 MISSING PROTEINS: As of neXtProt release 2018-01 there were still 2186 PE2,3,4 missing proteins. The major limitations in finding more PE2,3,4 missing proteins remain (1) protein sequences that cannot yield two proteotypic tryptic peptides, (2) lack of detectable expression of transcripts in tissues studied, and especially (3) concentrations of proteins too low to be detected even with highly sensitive mass spectrometers plus enrichment. See the complete manuscript for a detailed analysis.

FINDING EVIDENCE FOR FUNCTIONAL ANNOTATION OF UNCHARACTERIZED neXtProt PE1 PROTEINS: A comprehensive understanding of the human proteome requires not just the "parts" list and their variants and interactions, but deep knowledge of their functions in health and disease. Notably, according to neXtProt release 2018-01, 1937 PE1,2,3,4 proteins lack specific functional annotation, including 1260 uncharacterized PE1 proteins (referred to forthwith as **uPE1** proteins) (https://tinyurl.com/upe1proteins). C-HPP investigators agreed in September 2017 at the HUPO Congress in Dublin to launch a project focused on characterization of the functions of proteins and proteoforms, in addition to stringent identification of their expression. Deep-dive biological studies are strongly encouraged and 14 C-HPP teams have committed to begin work on selected uPE1 proteins. Meanwhile, the Chromosome 17 team has initiated and exploited a computational approach using I-TASSER and COFACTOR algorithms for prediction of protein function/s.

2018 UPDATE OF THE HUMAN PROTEIN ATLAS: At the end of 2017, the Human Protein Atlas (HPA) released version 18, which included data derived from the use of 26,009 antibodies, targeting proteins from almost 17,000 human genes (~87% of the human protein-coding genes). The HPA now presents three major atlases: The Tissue Atlas, the Cell Atlas, and a Pathology Atlas. The Tissue Atlas added data for caudate nucleus and thymus. The Cell Atlas was expanded by data from RNA sequencing of 8 different cell lines and increased the panel for immunofluorescence staining to 26 cell lines, as well as introducing cleavage furrow as an annotated structure. The Pathology Atlas integrates mRNA expression levels from 17 cancer types and 8,000 patients hosted by The Cancer Genome Atlas, links the expression of protein-encoding genes to the overall survival time for each patient, and complements these insights with protein level data from immunohistochemistry. Based on Kaplan-Meier analyses, elevated relative mRNA expression of 6,800 genes correlated with poor prognosis in at least one of the analyzed cancer types, while elevated relative mRNA expression of about 6,100 genes was linked to good prognosis in at least one cancer type.

The HPA portal integrated the latest guidelines regarding the validation of antibodies, using "enhanced validation" criteria from five procedures: (i) before and after knock-down of target genes (denoted genetic validation), (ii) induced overexpression or fluorescent tagging of proteins (recombinant expression validation), (iii) comparison of staining pattern with two antibodies targeting different epitopes (independent antibody validation), (iv) antibody-free methods (orthogonal validation), and (v) relating the staining pattern and protein size with an MS method (capture MS validation). As illustrated with 197 antibodies, the validation process is complex and painstaking and includes significant batch-to-batch variation.

NOTABLE THRUSTS IN THE USE OF PROTEOMICS FOR BIOLOGICAL AND DISEASE STUDIES, AN UPDATE FROM THE B/D-HPP: see additional reports from the B/D-HPP, as well as the C-HPP.

NEWS FROM THE HPP DURING THE PAST YEAR:

1. The C-HPP announced its uPE1 Initiative to annotate PE1 proteins lacking functions in neXtProt.
2. The HPP continued its highly-valued relationship with the ACS *Journal of Proteome Research* for another annual special issue.
3. The MS resource pillar launched a community analysis of a 96-phosphopeptide standard.

4. HPP Scientific Terms, Definitions & Abbreviations document has been prepared and will be posted on the HUPO website and circulated through HUPO's social media channels
5. The HUPO Council and Executive Committee, led by President Mike Snyder, identified several nominees and elected Mark Baker as successor to Gil Omenn as Chair of the HPP, effective January 2019. Gil will moderate a discussion where Mark will present some plans to the SSAB and HPP community during the HPP Thursday Workshop after the Orlando Congress.