Published in final edited form as:
*J Proteome Res*. 2018 December 07; 17(12): 4031–4041. doi:10.1021/acs.jproteome.8b00441.

# Progress on Identifying and Characterizing the Human Proteome: 2018 Metrics from the HUPO Human Proteome Project

**Gilbert S. Omenn**[×,π], **Lydie Lane**[∞], **Christopher M. Overall**[ϕ], **Fernando J. Corrales**[A], **Jochen M. Schwenk**[B], **Young-Ki Paik**[C], **Jennifer E. Van Eyk**[D], **Siqi Liu**[E], **Michael Snyder**[F], **Mark S. Baker**[G], and **Eric W. Deutsch**[π]

[×]Department of Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Avenue, Ann Arbor, Michigan 48109-2218, United States [∞]CALIPHO Group, SIB Swiss Institute of Bioinformatics and Department of Microbiology and Molecular Medicine, Faculty of Medicine, University of Geneva, CMU, Michel-Servet 1, 1211 Geneva 4, Switzerland [ϕ]Life Sciences Institute, Faculty of Dentistry, University of British Columbia, 2350 Health Sciences Mall, Room 4.401, Vancouver, BC Canada V6T 1Z3 [A]Centro Nacional de Biotecnologia (CSIC), Darwin 3, 28049, Madrid [B]Science for Life Laboratory, KTH Royal Institute of Technology, Tomtebodavägen 23A, 17165 Solna, Sweden [C]Yonsei Proteome Research Center, Room 425, Building #114, Yonsei University,50 Yonsei-ro, Seodaemoon-ku, Seoul 120-749, Korea [D]Advanced Clinical BioSystems Research Institute, Cedars Sinai Precision Biomarker Laboratories, Barbra Streisand Women's Heart Center, Cedars-Sinai Medical Center, Los Angeles, CA 90048, United States [E]Department of Molecular Biology, University of Texas Southwestern Medical Center, Dallas, TX 75390-9148, United States [F]Department of Genetics, Stanford University, Alway Building, 300 Pasteur Drive, 3165 Porter Drive, Palo Alto, 94304, United States [G]Department of Biomedical Sciences, Macquarie University, NSW 2109, Australia [π]Institute for Systems Biology, 401 Terry Avenue North, Seattle, Washington 98109-5263, United States

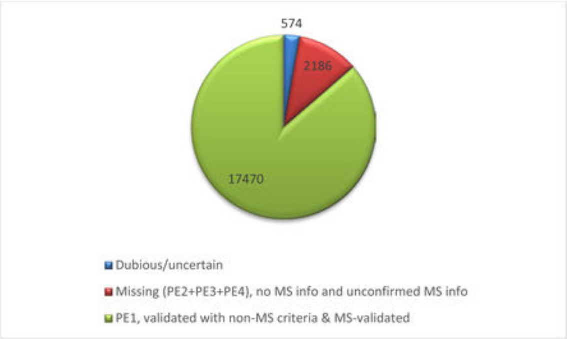
## Abstract

The Human Proteome Project (HPP) annually reports on progress throughout the field in credibly identifying and characterizing the human protein parts list and making proteomics an integral part of multi-omics studies in medicine and the life sciences. neXtProt release 2018–01-17, the baseline for this 6th annual HPP special issue of the *Journal of Proteome Research*, contains 17,470 PE1 proteins, 89% of all neXtProt predicted PE1–4 proteins, up from 17,008 in release 2017–01-23 and 13,975 in release 2012–02-24. Conversely, the number of neXtProt PE2,3,4 missing proteins has been reduced from 2949 to 2579 to 2186 over the past two years. Of the PE1 proteins, 16,092 are based on mass spectrometry results, and 1378 on other kinds of protein studies, notably protein-protein interaction findings. PeptideAtlas has 15,798 canonical proteins, up 625 over the past year, including 269 from SUMOylation studies. The largest reason for missing proteins is low abundance. Meanwhile, the Human Protein Atlas has released its Cell Atlas, Pathology Atlas, and updated Tissue Atlas, and is applying recommendations from the

Corresponding Author: Gilbert S. Omenn, Department of Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Avenue, Ann Arbor, Michigan 48109-2218, United States, gomenn@umich.edu, 734-763-7583.
Author Contributions
The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

International Working Group on Antibody Validation. Finally, there is progress using the quantitative multiplex organ-specific popular proteins targeted proteomics approach in various disease categories.

## Graphical Abstract



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

### Keywords

Metrics; missing proteins; HPP Guidelines; neXtProt; PeptideAtlas; Human Proteome Project (HPP); Chromosome-centric HPP (C-HPP); Biology and Disease-driven HPP (B/D-HPP); Human Proteome Organization (HUPO)

## INTRODUCTION

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 HPP has two over-arching goals: (1) completing the human protein parts list tied to predicted protein-coding genes as curated by neXtProt and updated annually in this HPP Metrics publication; and (2) integrating proteomics with genomics and other omics for use throughout the biomedical/life sciences community, led by the Biology and Disease-driven HPP. 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 omics integration has been facilitated through development of targeted proteomics, including the SRMAtlas and DIA/SWATH-MS, combined with bibliometric analyses that identify popular proteins widely studied in biomedical research.

There are 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, most recently, pathology. This Perspective introduces the sixth annual HPP special issue of the Journal of Proteome

Research[1–6], with emphasis on the identification of neXtProt PE2,3,4 missing proteins (MPs) led by the Chromosome-centric C-HPP.

## PROGRESS ON THE HUMAN PROTEOME PARTS LIST

neXtProt release 2018–01-17 (www.nextprot.org) and PeptideAtlas v2018–01b (www.peptideatlas.org), using HPP Guidelines for Interpretation of MS Data v2.1[7] (hupo.org/Guidelines), provided the baseline for HPP investigators and other scientists around the world to prepare manuscripts on MPs and other topics for this 2018 special issue. Full datasets and metadata are shared with the community through the ProteomeXchange Consortium repositories PRIDE, PASSEL/SRM, MassIVE, jPOST, and iProX as described in 2016[4]. ProteomeXchange[8] as of 2018–05-15 had 5248 publicly-released data sets, of which 2165 are from human samples, up from 3496 and 1478, respectively, one year ago (http://proteomecentral.proteomexchange.org/).

neXtProt release 2018–01-17 updated its validation of PE1 proteins to 17,470, an increase of 462 from release 2017–01-23 (see Table 1 for definitions and data). In 2013–2014 we excluded the “uncertain/dubious” genes of PE5 in our denominator of predicted proteins to be detected[3]. Thus, the number of missing proteins (MP = PE2+3+4) has been reduced to 2186 from 2579 a year ago, 2949 two years ago, and 5511 in 2012. The PE1 proteins now constitute 89% of the total PE 1,2,3,4 predicted proteins.

Meanwhile, the number of canonical proteins in PeptideAtlas increased during 2017 by 625 to reach 15,798, or 98% of the neXtProt figure of 16,092 PE1 proteins based on MS (see Table 1).

Figure 1 provides a detailed depiction of the changes within neXtProt by PE 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 - ASDURF (NX_L0R819), NBDY (NX_A0A0U1RRE5) and SLC35A4 and MIEF1 upstream open reading frame proteins (NX_L0R6Q1 and NX_L0R8F8); the recently characterized LINC00961-encoded SPAR polypeptide (NX_A0A1B0GVQ0); and three chimeric proteins with biological activity – IQCJ-SCHIP1 (NX_B3KU38), CSB-PGBD3 (NX_P0DP91) and BARGIN (NX_Q6ZT62). Among the new PE3 proteins is the sORF Minion (NX_A0A1B0GTQ4), recently characterized in mouse but still lacking transcriptomic or proteomic evidence in human. 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.

UniProtKB/Swiss-Prot and neXtProt integrate manually curated protein-protein interaction (PPI) data from the IntAct database[9]. 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 (labeled as Gold in neXtProt). 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. These criteria ensure that only experimental data are used and there are always at least two experimental datasets to reach the threshold score to trigger a PE1 upgrade. Moreover, the upgrade cannot be triggered if the protein pair only co-occurs in larger complexes as detected by AP/MS, because there must be at least one indication that the proteins are in physical contact. Currently, 530 PE1 protein validations are based on "Gold" 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.

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 1378 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 (1798) 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 as noted by Omenn et al[5]. A major advance is the introduction of MSFragger for ultra-fast identification of post-translational and chemical modifications of peptides (Kong et al[10]). An extensive review of human proteoforms has recently been published[11].

## THE FATE OF MISSING PROTEINS NOMINATED FOR neXtProt REVIEW IN THE JPR 2017 PAPERS

The editorial for the JPR 2017 special issue[12] 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 from Li et al[13] using Triton X-100 solubilization plus ProteoMiner hexapeptide-covered beads as an enrichment/equalization strategy for low-abundance proteins with kidney, bladder, liver, and colorectal specimens, and confirmation with PRM; 12 from Carapito et al[14] on the sperm proteome with PRM confirmation; 3 from Wang et al[15] using a multi-protease strategy on testis; 1 from Peng[16] from the kidney phosphoproteome; 1 from Meyfour et al[17] based on biological studies of a Y chromosome protein in cardiac development; plus 41 cautiously forwarded from Elguoshy et al[18] using the stranded peptides approach described below. Twenty MP reported by Li et al[8], including 5 without PRM validation that had not been counted in the editorial, are now classified as canonical in PeptideAtlas and validated as PE1 in neXtProt. Only 4 from Carapito et al[14] were validated as PE1, because the review of PRM data at PeptideAtlas was incomplete, as was the consideration of immunochemistry results using antibodies obtained from HPA. From the list of proteins reported by Wang et al,[15] only P0DMU9 was validated as PE1, and that was based on unrelated data for 5 proteotypic peptides not reported by Wang et al[15]. Beta-defensin 123 (Q8N688) may be a candidate for a policy discussion of implementation of the Guidelines exceptions clause[7], because there are excellent spectra for all of the three potential proteotypic tryptic peptides, but two are only 8 aa in length. SIM1 from Peng et al[16] was validated. TBL1Y proposed by Meyfour et al[17] was validated, but based on other MS data. The Elguoshy et al[18] data, which lacked PXD identifiers, could not be considered by PeptideAtlas and neXtProt, but 16 of their proposed proteins were coincidentally promoted to PE1 from other datasets.

In sum, through various paths, 43 of the 73 MP candidates recommended from the 2017 JPR special issue have to date qualified as PE1 by neXtProt. It takes a village!

With the pace in MP identification seeming to decline as the “more easily detected proteins” had been identified by conventional strategies, the C-HPP and HPP announced at the Sun Moon Lake HPP Workshop after the HUPO World Congress in Taiwan in September 2016 a “MP50 Challenge” to accelerate identification of 50 PE2,3,4 missing proteins per chromosome team over two years. While some teams are still in planning phases, work from the entire proteomics community has raised 952 MPs to PE1 status since neXtProt release 2016–01-11 (see Table 1), including the notable completion of the mitochondrial chromosome with its 15th protein by the Italian team[19]. Barriers include limited temporal and spatial expression of proteins important in fetal development, disease responses, rare cell types, and difficulties to access tissues. Of course, the more stringent guidelines introduced in 2016 appropriately raised the bar for MP identification from MS.

Two new situations have arisen with regard to the standardized reanalysis of datasets by PeptideAtlas. First, as illustrated above, some labs use PRM instead of SRM for comparison of natural and synthetic peptides. PeptideAtlas, which several years ago developed PASSEL for extensive SRM data, has now begun to process PRM data. The second development was the strategy of the Chromosome 7 team (Baker et al[20]) and the Chromosome X team (Elguoshy et al[18]) to identify what we have named "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 (preferably deposited in PRIDE or other accessible repository with a PXD identifier) to compare with the spectra available in SRMAtlas[21] 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. Peptides in PeptideAtlas have been validated. There may be many stranded peptides from different studies in GPMdb, MassIVE, Proteomics DB, or other resources. There is no automated connector between GPMdb and either PeptideAtlas or neXtProt; they use different reference genomes and have other incompatible features. Thus, use of stranded peptides from GPMdb requires search within GPMdb for the original source and links to ProteomeXchange PXD identifiers and to original spectra and metadata. Making the case, including careful scrutiny of the original MS spectra, is the responsibility of authors; requests to original MS data generators seeking their re-submission of original data to ProteomeXchange may be useful in some cases.

## SEEKING PE2,3,4 MISSING PROTEINS IN MAJOR PROTEIN FAMILIES

Hydrophobic proteins are estimated to account for 924 of the 2186 MPs (Zhang et al[22]). Thus, techniques to extract and solubilize these proteins remain a key need. Many of these proteins are members of six major protein families, as shown in Figure 4. NeXtProt releases for 2013, 2016, 2017, and 2018 were downloaded for Chromosomes 1–22, X, Y, and mitochondria; PE2–4 MPs were sorted alphabetically according to protein families or groupings and confirmed through UniProtKB, Pfam, and GeneCards. Miscellaneous "uncharacterized proteins" were excluded. Figure 4 summarizes the progression of findings for the top six PE2,3,4 missing protein families: (i) olfactory receptors (ORs), (ii) non-OR transmembrane proteins, (iii) zinc finger proteins, (iv) homeobox proteins, (v) keratin-associated proteins, and (vi) coiled-coil proteins. Membrane proteins annotated as olfactory receptors and as other transmembrane proteins (including non-OR GPCRs, taste receptors and solute carrier proteins) are now the two most abundant PE2–4 families. Non-OR transmembrane proteins have overtaken zinc-finger proteins, as a significant number of zinc-finger proteins were re-classified as PE1 between 2017 and 2018 (see Adhikari et al[23]). The only top-six protein family where no progress has been made to identify proteins by MS is ORs; taste receptors also show little progress (Adhikari et al[23]). Rare ORs (OR1D2, OR3A4) have reached PE1 status through protein-protein interaction analyses (see Siddiqui et al[24]). There are no credible OR detections in PeptideAtlas.

## RECOGNIZING THE LIMITATIONS OF FINDING PE2,3,4 MISSING PROTEINS

As of neXtProt release 2018–01-17 there were still 2186 PE2,3,4 missing proteins. This Perspective analyzes the progress reflected in the 2018 release: 462 new PE1 proteins, including 158 based on protein-protein interactions via IntAct, and 625 new canonical proteins from mass spectrometry in PeptideAtlas, including a very large contribution of 269 from one study of the post-translational modification SUMOylation (see Figure 3, above). The Editorial that will accompany the published articles from this year will identify additional progress.

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 (3) concentrations of proteins too low to be detected with even the recently greatly-enhanced mass spectrometers plus enrichment with such steps as ProteoMiner hexapeptide beads.

Of the 2186 PE 2,3,4 missing proteins (each lacking any or sufficient MS evidence to meet the HPP Guidelines), only a small number may not be unambiguously detectable with current mass spectrometry techniques, if the transcript is expressed and the protein has sufficient abundance. We performed the following exercise to estimate that number. Using in-silico digestion of the PE1–4 proteome with trypsin without considering missed cleavages, we find that only 141 PE1 proteins (0.8%) and 79 PE2,3,4 proteins (3.6%) cannot potentially generate two non-nested uniquely mapping tryptic peptides of length 9–50 amino acids (after removing 54 sequence-exact duplicates, clipping initiating methionines or signal peptides, and treating isoleucine = leucine, I=L). Trans-membrane region issues are not considered in the computation. Moreover, when a set of five common proteases (trypsin, chymotrypsin, AspN, GluC, LysC) is applied in the same approach, we find that a mere 13 proteins cannot potentially generate the requisite two non-nested uniquely mapping peptides of length 9–50 amino acids. Remarkably, of those 13 proteins, 6 are already PE1 in neXtProt, mostly via other technologies, including Ataxin-8 (Q156A1), whose sequence has all glutamines (79Q). One of these six seemingly MS-unattainable PE=1 proteins (C9JFL3), with only a single lysine and no arginines, is canonical in PeptideAtlas due to several well-detected non-nested peptides that are semi-tryptic. Thus, if a protein is sufficiently abundant, imperfect cleavage may provide suitable peptides to meet the HPP guidelines. Under-digesting or over-digesting with trypsin itself may be a useful tactic. The issues around membrane-bound proteins do complicate the picture, but in PeptideAtlas there are also many canonical detections of proteins that seem unattainable by following strict protease rules around trans-membrane regions. We conclude that nearly all proteins could, in principle, be detectable by mass spectrometry following current HPP MS Guidelines with reasonable additional effort if the proteins are of sufficiently high abundance in an analyzed sample. We recognize that selective enrichment techniques (including affinity capture) will be essential for many of the remaining MPs to achieve the necessary abundance for detection with MS.

Of course, it is unlikely that proteins will be found in specimens that have undetectable or very low levels (<1 FPKM) of the corresponding transcript. An initial assessment of transcript data from Human Protein Atlas, GTEx, FANTOM5, and TCGA suggests that as many as 800–1000 PE2,3,4 predicted protein-coding genes may be lacking detectable transcripts in all tissues tested (J. Schwenk, preliminary analyses). There are 400 predicted olfactory receptors that are such missing proteins (Figure 4); Hwang et al[25] were unable to detect expression of even one olfactory receptor protein in human olfactory epithelium.

There remain many under-investigated types of specimens, including unusual tissue types, embryonic and fetal stages of life, and responses to oxidative or inflammatory stress. The Chr 2/14 French/Swiss consortium has exploited the knowledge of exclusive expression of hundreds of transcripts in male reproductive tract by performing deep studies of sperm[14, 26–27], complemented by studies of testis by the Chinese team[15, 28, 29–30]. Other examples are dental pulp[31], male fetal cardiac development[17], kidney and bladder[16], and beta-defensins[32]. The Chromosome 17 team has analyzed how 43 previously MPs have been identified as PE1 since the announcement of the MP50 Challenge and identified 35 of the remaining 105 MPs as amenable to identification by MS or protein-protein interactions[24].

## 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 interactions, but deep knowledge of their functions in health and disease. Notably, according to neXtProt release 2018–01-17, 1937 PE1,2,3,4 proteins lack specific functional annotation, including 1260 uncharacterized PE1 proteins (uPE1) (https://tinyurl.com/upe1proteins) and 677 uncharacterized missing proteins (PE2+PE3+E4). C-HPP investigators agreed in September 2017 at the HPP meeting in Dublin to launch a project focused on characterization of the functions of proteins and proteoforms, not just stringent identification of their expression[33]. Deep-dive biological studies are strongly encouraged; an example is Na et al[34]. According to Paik et al[33], 14 C-HPP teams have committed to begin work on selected uPE1 proteins. The Chromosome 17 team is exploiting a computational approach using I-TASSER and COFACTOR algorithms for prediction of protein functions[35].

## 2018 UPDATE OF THE HUMAN PROTEIN ATLAS

At the end of 2017, the Human Protein Atlas (HPA) released version 18 (based on Ensembl version: 88.38), which included 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[36], the Cell Atlas[37], and a Pathology Atlas[38]. The Tissue Atlas added data for caudate nucleus and thymus. The Cell Atlas was expanded by data from RNA sequencing of 8 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 8000 patients hosted by The Cancer Genome Atlas[39], 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 analysis, elevated relative mRNA expression of 6800 genes correlated with poor prognosis in at least one of the analyzed cancer types, while elevated relative mRNA expression of about 6100 genes was linked to good prognosis.

The HPA portal also integrated the latest guidelines regarding the validation of antibodies, using "enhanced validation" criteria as defined by its International Working Group[40]. A total of 10,540 antibodies in the HPA v18 (40%), targeting 6,787 human proteins, now have enhanced validation data from analyses of cells or tissues with immunocytochemistry, immunohistochemistry, and Western blots. The five validation procedures are: (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 determined protein size with a capture MS method (capture MS validation). As illustrated with a set of 197 antibodies, this validation process is complex and painstaking, including recognition of significant batch-to-batch variation[41].

## NOTABLE THRUSTS IN THE USE OF PROTEOMICS FOR BIOLOGICAL AND DISEASE STUDIES, AN UPDATE FROM THE B/D-HPP

One of the main goals of the Biology and Disease-driven Human Proteome Project (B/D-HPP) is to reveal the molecular basis of physiological/pathological processes by identifying the driver proteins involved. To guide organ and biofluid studies, B/D HPP initiatives have been encouraged to utilize lists of "popular proteins" (highly cited proteins associated with the organ or other topic of interest) to generate functional hypotheses and to pave the way for new clinical applications of targeted proteomics. Two web tools have been developed to perform systematic bibliographic searches and rank the most cited proteins under specific topics (Lam et al[42]; Yu KS et al[43]). The popular protein approach has been used in two studies demonstrating the principal role of reconfiguration of one carbon metabolism in the liver during hepatocarcinogenesis[44] and creating a targeted assay to monitor B-type natriuretic peptidoforms that might be biomarkers for diagnosing and monitoring heart failure[45].

Characterization of proteoforms and PTMs is an unmet need to understand the dynamics of pathogenic processes. A novel mass spectrometry-based whole protein assay enabled quantitation of the percentage of mutant KRAS4b present in colorectal cancer tissue, and the differences on C-terminal carboxymethylation, which is critical for KRAS function[46]. Understanding PTM status of drug targets and the functional implications is key to next generation therapies. Van Eyk et al[47] have shown that S-nitrosylation of GSK3B at specific residues can send the protein to the nucleus, away from its cytoplasmic location, resulting in a different repertoire of phosphorylated substrates and altering responses to drugs[47].

The detailed description of the array of peptides associated with human HLA phenotypes is of paramount importance to understand the immune system and to guide the development of next-generation vaccines and immunotherapies against autoimmunity, infectious diseases, and cancers. Mass spectrometry is the only available technology to interrogate the immunopeptidome in an accurate, systematic, unbiased manner. The Human Immunopeptidome Proteome Project (HIPP)[48] has developed the first public database of quality-controlled immunopeptidomic data generated by mass spectrometry[49]. The combination of MHC isolation, peptide analysis, and exome sequencing facilitates identification of immunoglobulin neoantigens as targets for lymphoma[50] and ovarian cancer[51] immunotherapy and opens new avenues for individualized immunotherapies.

One of the principal aims of the B/D-HPP is to better understand human organ physiology and pathology through comprehensive proteomic insights. To this end, eye and plasma proteomes have been updated recently. A total of 9782 non-redundant proteins (not necessarily compliant with HPP guidelines) are now in the human eye proteome database of 11 tissue compartments plus biofluids, with the highest number (6538) from vitreous humor and the lowest (827) from aqueous humor[52]. More than 122,000 peptide sequences matching to 3509 protein identifications compliant with the HPP guidelines are described in the 2017 Plasma Peptide Atlas[53].

The B/D HPP community has sought to identify protein biomarkers in multiple clinical fields, including Pediatrics[54] and Rheumatoid and Autoimmune Disorders such as knee osteoarthritis[55] and osteoarticular pathologies with MS[56] or protein arrays[57, 58], and rheumatoid arthritis with anti-citrullinated protein antibodies[58]. In-depth analyses of the synaptosomal proteome led to association of specific protein expression patterns with social behavior in patients with schizophrenias[59], through a joint C-HPP (chromosome 15) and B/D HPP (Brazilian Brain initiative) effort. Similar cooperation led to a comprehensive description of the human mitochondrial proteome under standardized protocols[19], to assess the pharmacological prospects of targeting specific mitochondrial proteins to selectively kill cancer cells[60]. Cancer biomarker discovery has experienced significant progress, as shown by the recent studies of breast, ovarian, and colorectal cancers published by members of the Cancer B/D-HPP and NCI CPTAC initiative[61–63]. Novel DIA MS-based and proteogenomic approaches have facilitated discovery of new protein species and splice variants that can be used to improve colorectal cancer screening[64, 65] and point to therapeutic targets in breast cancers[66]. Finally, realizing that food allergy is a global health concern, the pros and cons of current analytical methods for allergenic risk assessment have been reviewed by members of the Food and Nutrition B/D-HPP initiative[67]; this group also evaluated state-of-the-art proteomic and metaproteomic approaches to study host-microbiome interactions[68].

As noted above, productive interaction between HPP groups is shedding light on many relevant aspects of human biology. Such cooperative multi-omics efforts should guide next steps in our endeavor to generate a comprehensive human proteome map with all functional annotations needed to decipher the code of life and support future molecular precision medicine.

## ACKNOWLEDGMENTS

We appreciate the guidance from the HPP Executive Committee and Dr. Amos Bairoch of neXtProt. We thank the UniProt groups at SIB, EBI, and PIR for providing high-quality annotations for the human proteins in UniProtKB/Swiss-Prot. The neXtProt server is hosted by VitalIT in Switzerland. The PeptideAtlas server is hosted at the Institute for Systems Biology in Seattle.

Funding Sources

G.S.O. acknowledges grant support from National Institutes of Health grants P30ES017885–01A1 and NIH U24CA210967; E.W.D. from NIH grants R01GM087221, R24GM127667, and U54EB020406; L.L. and neXtProt from the SIB Swiss Institute of Bioinformatics; C.M.O. by a Canadian Institutes of Health Research 7-year Foundation Grant and a Canada Research Chair in Protease Proteomics and Systems Biology; M.S.B by NHMRC Project Grant APP1010303; and J.M.S by the Knut and Alice Wallenberg Foundation for the Human Protein Atlas.

## References

1. Legrain P; Aebersold R; Archakov A; Bairoch A; Bala K; Beretta L; Bergeron J; Borchers CH; Corthals GL; Costello CE; Deutsch EW; Domon B; Hancock W; He F; Hochstrasser D; Marko-Varga G; Salekdeh GH; Sechi S; Snyder M; Srivastava S; Uhlen M; Wu CH; Yamamoto T; Paik YK; Omenn GS, The Human Proteome Project: current state and future direction. Mol Cell Proteomics 2011, 10, (7), M111 009993.
2. 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), 1–5. [PubMed: 23256439]
3. Lane L; Bairoch A; Beavis RC; Deutsch EW; Gaudet P; Lundberg E; Omenn GS, Metrics for the human proteome project 2013–2014 and strategies for finding missing proteins. J. Proteome Res 2014, 13, (1), 15–20. [PubMed: 24364385]
4. Omenn GS; Lane L; Lundberg EK; Beavis RC; Nesvizhskii AI; Deutsch EW, Metrics for the Human Proteome Project 2015: progress on the human proteome and guidelines for high-confidence protein identification. J. Proteome Res 2015, 14, (9), 3452–60. [PubMed: 26155816]
5. Omenn GS; Lane L; Lundberg EK; Beavis RC; Overall CM; Deutsch EW, Metrics for the Human Proteome Project 2016: Progress on identifying and characterizing the human proteome, including post-translational modifications. J Proteome Res 2016, 15, (11), 3951–3960. [PubMed: 27487407]
6. Omenn GS; Lane L; Lundberg EK; Overall CM; Deutsch EW, Progress on the HUPO draft human proteome: 2017 metrics of the Human Proteome Project. J Proteome Res 2017, 16, (12), 4281–4287. [PubMed: 28853897]
7. Deutsch EW; Overall CM; Van Eyk JE; Baker MS; Paik YK; Weintraub ST; Lane L; Martens L; Vandenbrouck Y; Kusebauch U; Hancock WS; Hermjakob H; Aebersold R; Moritz RL; Omenn GS, Human Proteome Project mass spectrometry data interpretation Guidelines 2.1. J Proteome Res 2016, 15, (11), 3961–3970. [PubMed: 27490519]
8. Deutsch EW; Csordas A; Sun Z; Jarnuczak A; Perez-Riverol Y; Ternent T; Campbell DS; Bernal-Llinares M; Okuda S; Kawano S; Moritz RL; Carver JJ; Wang M; Ishihama Y; Bandeira N; Hermjakob H; Vizcaino JA, The ProteomeXchange consortium in 2017: Supporting the cultural change in proteomics public data deposition. Nucleic Acids Res 2017, 45, (D1), D1100–d1106. [PubMed: 27924013]
9. Kerrien S; Aranda B; Breuza L; Bridge A; Broackes-Carter F; Chen C; Duesbury M; Dumousseau M; Feuermann M; Hinz U; Jandrasits C; Jimenez RC; Khadake J; Mahadevan U; Masson P; Pedruzzi I; Pfeiffenberger E; Porras P; Raghunath A; Roechert B; Orchard S; Hermjakob H, The IntAct molecular interaction database in 2012. Nucleic Acids Res 2012, 40, (Database issue), D841–6. [PubMed: 22121220]
10. Kong AT; Leprevost FV; Avtonomov DM; Mellacheruvu D; Nesvizhskii AI, MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. Nat Methods 2017, 14, (5), 513–520. [PubMed: 28394336]
11. Aebersold R; Agar JN; Amster IJ; Baker MS; Bertozzi CR; Boja ES; Costello CE; Cravatt BF; Fenselau C; Garcia BA; Ge Y; Gunawardena J; Hendrickson RC; Hergenrother PJ; Huber CG; Ivanov AR; Jensen ON; Jewett MC; Kelleher NL; Kiessling LL; Krogan NJ; Larsen MR; Loo JA; Ogorzalek Loo RR; Lundberg E; MacCoss MJ; Mallick P; Mootha VK; Mrksich M; Muir TW; Patrie SM; Pesavento JJ; Pitteri SJ; Rodriguez H; Saghatelian A; Sandoval W; Schluter H; Sechi S; Slavoff SA; Smith LM; Snyder MP; Thomas PM; Uhlen M; Van Eyk JE; Vidal M; Walt DR; White FM; Williams ER; Wohlschlager T; Wysocki VH; Yates NA; Young NL; Zhang B, How many human proteoforms are there? Nat Chem Biol 2018, 14, (3), 206–214. [PubMed: 29443976]
12. Paik YK; Overall CM; Deutsch EW; Van Eyk JE; Omenn GS, Progress and future direction of Chromosome-Centric Human Proteome Project. J Proteome Res 2017, 16, (12), 4253–4258. [PubMed: 29191025]

13. Li S; He Y; Lin Z; Xu S; Zhou R; Liang F; Wang J; Yang H; Liu S; Ren Y, Digging more missing proteins using an enrichment approach with ProteoMiner. J Proteome Res 2017, 16, (12), 4330–4339. [PubMed: 28960076]

14. Carapito C; Duek P; Macron C; Seffals M; Rondel K; Delalande F; Lindskog C; Freour T; Vandenbrouck Y; Lane L; Pineau C, Validating missing proteins in human sperm cells by targeted mass-spectrometry- and antibody-based methods. J Proteome Res 2017, 16, (12), 4340–4351. [PubMed: 28891297]

15. Wang Y; Chen Y; Zhang Y; Wei W; Li Y; Zhang T; He F; Gao Y; Xu P, Multi-protease strategy identifies three PE2 missing proteins in human testis tissue. J Proteome Res 2017, 16, (12), 4352–4363. [PubMed: 28959888]

16. Peng X; Xu F; Liu S; Li S; Huang Q; Chang L; Wang L; Ma X; He F; Xu P, Identification of missing proteins in the phosphoproteome of kidney cancer. J Proteome Res 2017, 16, (12), 4364–4373. [PubMed: 28857561]

17. Meyfour A; Ansari H; Pahlavan S; Mirshahvaladi S; Rezaei-Tavirani M; Gourabi H; Baharvand H; Salekdeh GH, Y Chromosome missing protein, TBL1Y, may play an important role in cardiac differentiation. J Proteome Res 2017, 16, (12), 4391–4402. [PubMed: 28853286]

18. Elguoshy A; Hirao Y; Xu B; Saito S; Quadery AF; Yamamoto K; Mitsui T; Yamamoto T, Identification and validation of human missing proteins and peptides in public proteome databases: Data mining strategy. J Proteome Res 2017, 16, (12), 4403–4414. [PubMed: 28980472]

19. Alberio T; Pieroni L; Ronci M; Banfi C; Bongarzone I; Bottoni P; Brioschi M; Caterino M; Chinello C; Cormio A; Cozzolino F; Cunsolo V; Fontana S; Garavaglia B; Giusti L; Greco V; Lucacchini A; Maffioli E; Magni F; Monteleone F; Monti M; Monti V; Musicco C; Petrosillo G; Porcelli V; Saletti R; Scatena R; Soggiu A; Tedeschi G; Zilocchi M; Roncada P; Urbani A; Fasano M, Toward the standardization of mitochondrial proteomics: The Italian Mitochondrial Human Proteome Project Initiative. J Proteome Res 2017, 16, (12), 4319–4329. [PubMed: 28828861]

20. Baker MS; Ahn SB; Mohamedali A; Islam MT; Cantor D; Verhaert PD; Fanayan S; Sharma S; Nice EC; Connor M; Ranganathan S, Accelerating the search for the missing proteins in the human proteome. Nat Commun 2017, 8, 14271. [PubMed: 28117396]

21. Kusebauch U; Campbell DS; Deutsch EW; Chu CS; Spicer DA; Brusniak MY; Slagel J; Sun Z; Stevens J; Grimes B; Shteynberg D; Hoopmann MR; Blattmann P; Ratushny AV; Rinner O; Picotti P; Carapito C; Huang CY; Kapousouz M; Lam H; Tran T; Demir E; Aitchison JD; Sander C; Hood L; Aebersold R; Moritz RL, Human SRMAtlas: A resource of targeted assays to quantify the complete human proteome. Cell 2016, 166, (3), 766–78. [PubMed: 27453469]

22. Zhang Y; Lin Z; Hou K; Sui Y; Zhang K; Li H; Liu S, Improvement of peptide separation for exploring the missing proteins localized on membranes. J. Proteome Res 2018, submitted.

23. Adhikari S; Sharma S; Ahn S; Baker M, How much of the human olfactory receptor proteome is findable using high-stringency mass spectrometry? J. Proteome Res 2018, submitted.

24. Siddiqui O; Zhang H; Guan Y; Omenn GS, Chromosome 17 missing proteins: Recent progress and future directions as part of the next-50MP challenge. J. Proteome Res 2018, submitted.

25. Hwang H; Jeong JE; Lee HK; Yun KN; An HJ; Lee B; Paik YK; Jeong TS; Yee GT; Kim JY; Yoo JS, Identification of missing proteins in human olfactory epithelial tissue by liquid chromatography-tandem mass spectrometry. J. Proteome Res 2018, submitted.

26. Vandenbrouck Y; Lane L; Carapito C; Duek P; Rondel K; Bruley C; Macron C; Gonzalez de Peredo A; Coute Y; Chaoui K; Com E; Gateau A; Hesse AM; Marcellin M; Mear L; Mouton-Barbosa E; Robin T; Burlet-Schiltz O; Cianferani S; Ferro M; Freour T; Lindskog C; Garin J; Pineau C, Looking for missing proteins in the proteome of human spermatozoa: An update. J Proteome Res 2016, 15, (11), 3998–4019. [PubMed: 27444420]

27. Melaine N; Com E; Bellaud P; Guillot L; Lagarrigue M; Morrice N; Guevel B; Lavigne R; Velez de la Calle J-F; Pineau C, Deciphering the dark proteome: use of the testis and characterization of two dark proteins. J. Proteome Res 2018, submitted.

28. Zhang Y; Li Q; Wu F; Zhou R; Qi Y; Su N; Chen L; Xu S; Jiang T; Zhang C; Cheng G; Chen X; Kong D; Wang Y; Zhang T; Zi J; Wei W; Gao Y; Zhen B; Xiong Z; Wu S; Yang P; Wang Q; Wen B; He F; Xu P; Liu S, Tissue-based proteogenomics reveals that human testis endows plentiful missing proteins. J Proteome Res 2015, 14, (9), 3583–94. [PubMed: 26282447]

29. Wei W; Luo W; Wu F; Peng X; Zhang Y; Zhang M; Zhao Y; Su N; Qi Y; Chen L; Zhang Y; Wen B; He F; Xu P, Deep coverage proteomics identifies more low-abundance missing proteins in human testis tissue with Q-exactive HF mass spectrometer. J Proteome Res 2016, 15, (11), 3988–3997. [PubMed: 27535590]

30. Sun J; Shi J; Wang Y; Chen Y; Kong D; Chang L; Liu F; Lv Z; Zhou Y; He F; Zhang Y; Xu P, Multi-proteases combining high-pH reverse-phase separation strategy verifies fourteen missing proteins in human testis tissue. J. Proteome Res 2018, submitted.

31. Eckhard U; Marino G; Abbey SR; Tharmarajah G; Matthew I; Overall CM, The human dental pulp proteome and N-Terminome: Levering the unexplored potential of semitryptic peptides enriched by TAILS to identify missing proteins in the Human Proteome Project in underexplored tissues. J. Proteome Res 2015, 14, (9), 3568–82. [PubMed: 26258467]

32. Fan Y; Zhang Y; Xu S; Kong N; Zhou Y; Ren Z; Deng Y; Lin L; Ren Y; Wang Q; Zi J; Wen B; Liu S, Insights from ENCODE on missing proteins: Why beta-defensin expression is scarcely detected. J Proteome Res 2015, 14, (9), 3635–44. [PubMed: 26258396]

33. Paik YK; Lane L; Overall CM, neXt-CP50, the C-HPP pilot project for functional characterization of identified proteins with no known function. J. Proteome Res 2018.

34. Na K; Shin H; Cho JY; Jung SH; Lim J; Lim JS; Kim EA; Kim HS; Kang AR; Kim JH; Shin JM; Jeong SK; Kim CY; Park JY; Chung HM; Omenn GS; Hancock WS; Paik YK, Systematic proteogenomic approach to exploring a novel function for NHERF1 in human reproductive disorder: Lessons for exploring missing proteins. J Proteome Res 2017, 16, (12), 4455–4467. [PubMed: 28960081]

35. Zhang C; Omenn GS; Zhang Y, Structure and protein interaction-based gene ontology annotations reveal likely functinos of uncharacterized proteins of human chromosome 17. J. Proteome Res 2018, submitted.

36. Uhlen M; Fagerberg L; Hallstrom BM; Lindskog C; Oksvold P; Mardinoglu A; Sivertsson A; Kampf C; Sjostedt E; Asplund A; Olsson I; Edlund K; Lundberg E; Navani S; Szigyarto CA; Odeberg J; Djureinovic D; Takanen JO; Hober S; Alm T; Edqvist PH; Berling H; Tegel H; Mulder J; Rockberg J; Nilsson P; Schwenk JM; Hamsten M; von Feilitzen K; Forsberg M; Persson L; Johansson F; Zwahlen M; von Heijne G; Nielsen J; Ponten F, Proteomics. Tissue-based map of the human proteome. Science 2015, 347, (6220), 1260419. [PubMed: 25613900]

37. Thul PJ; Akesson L; Wiking M; Mahdessian D; Geladaki A; Ait Blal H; Alm T; Asplund A; Bjork L; Breckels LM; Backstrom A; Danielsson F; Fagerberg L; Fall J; Gatto L; Gnann C; Hober S; Hjelmare M; Johansson F; Lee S; Lindskog C; Mulder J; Mulvey CM; Nilsson P; Oksvold P; Rockberg J; Schutten R; Schwenk JM; Sivertsson A; Sjostedt E; Skogs M; Stadler C; Sullivan DP; Tegel H; Winsnes C; Zhang C; Zwahlen M; Mardinoglu A; Ponten F; von Feilitzen K; Lilley KS; Uhlen M; Lundberg E, A subcellular map of the human proteome. Science 2017, 356, (6340).

38. Uhlen M; Zhang C; Lee S; Sjostedt E; Fagerberg L; Bidkhori G; Benfeitas R; Arif M; Liu Z; Edfors F; Sanli K; von Feilitzen K; Oksvold P; Lundberg E; Hober S; Nilsson P; Mattsson J; Schwenk JM; Brunnstrom H; Glimelius B; Sjoblom T; Edqvist PH; Djureinovic D; Micke P; Lindskog C; Mardinoglu A; Ponten F, A pathology atlas of the human cancer transcriptome. Science 2017, 357, (6352).

39. Weinstein JN; Collisson EA; Mills GB; Shaw KR; Ozenberger BA; Ellrott K; Shmulevich I; Sander C; Stuart JM, The cancer genome atlas pan-cancer analysis project. Nat Genet 2013, 45, (10), 1113–20. [PubMed: 24071849]

40. Uhlen M; Bandrowski A; Carr S; Edwards A; Ellenberg J; Lundberg E; Rimm DL; Rodriguez H; Hiltke T; Snyder M; Yamamoto T, A proposal for validation of antibodies. Nat Methods 2016, 13, (10), 823–7. [PubMed: 27595404]

41. Skogs M; Stadler C; Schutten R; Hjelmare M; Gnann C; Bjork L; Poser I; Hyman A; Uhlen M; Lundberg E, Antibody validation in bioimaging applications based on endogenous expression of tagged proteins. J Proteome Res 2017, 16, (1), 147–155. [PubMed: 27723985]

42. Lam MP; Venkatraman V; Xing Y; Lau E; Cao Q; Ng DC; Su AI; Ge J; Van Eyk JE; Ping P, Data-driven approach to determine popular proteins for targeted proteomics translation of six organ systems. J Proteome Res 2016, 15, (11), 4126–4134. [PubMed: 27356587]

43. Yu KH; Lee TL; Wang CS; Chen YJ; Re C; Kou SC; Chiang JH; Kohane IS; Snyder M, Systematic protein prioritization for targeted proteomics studies through literature mining. J Proteome Res 2018.

44. Mora MI; Molina M; Odriozola L; Elortza F; Mato JM; Sitek B; Zhang P; He F; Latasa MU; Avila MA; Corrales FJ, Prioritizing popular proteins in liver cancer: remodelling one-carbon metabolism. J Proteome Res 2017, 16, (12), 4506–4514. [PubMed: 28944671]

45. Zhang S; Raedschelders K; Santos M; Van Eyk JE, Profiling B-Type natriuretic peptide cleavage peptidoforms in human plasma by capillary electrophoresis with electrospray ionization mass spectrometry. J Proteome Res 2017, 16, (12), 4515–4522. [PubMed: 28861997]

46. Ntai I; Fornelli L; DeHart CJ; Hutton JE; Doubleday PF; LeDuc RD; van Nispen AJ; Fellers RT; Whiteley G; Boja ES; Rodriguez H; Kelleher NL, Precise characterization of KRAS4b proteoforms in human colorectal cells and tumors reveals mutation/modification cross-talk. Proc Natl Acad Sci U S A 2018, 115, (16), 4140–4145. [PubMed: 29610327]

47. Wang S; Venkatraman V; Crowgey EL; Liu T; Fu Z; Holewinski RJ; Ranek MJJ; Kass DA; O'Rourke B; Van Eyk JE, Protein s-nitrosylation controls glycogen synthase kinase 3β function independent of its phosphorylation state. Circulation Research 2018.

48. Caron E; Aebersold R; Banaei-Esfahani A; Chong C; Bassani-Sternberg M, A case for a human immuno-peptidome project consortium. Immunity 2017, 47, (2), 203–208. [PubMed: 28813649]

49. Shao W; Pedrioli PGA; Wolski W; Scurtescu C; Schmid E; Vizcaino JA; Courcelles M; Schuster H; Kowalewski D; Marino F; Arlehamn CSL; Vaughan K; Peters B; Sette A; Ottenhoff THM; Meijgaarden KE; Nieuwenhuizen N; Kaufmann SHE; Schlapbach R; Castle JC; Nesvizhskii AI; Nielsen M; Deutsch EW; Campbell DS; Moritz RL; Zubarev RA; Ytterberg AJ; Purcell AW; Marcilla M; Paradela A; Wang Q; Costello CE; Ternette N; van Veelen PA; van Els C; Heck AJR; de Souza GA; Sollid LM; Admon A; Stevanovic S; Rammensee HG; Thibault P; Perreault C; Bassani-Sternberg M; Aebersold R; Caron E, The SysteMHC atlas project. Nucleic Acids Res 2018, 46, (D1), D1237–d1247. [PubMed: 28985418]

50. Khodadoust MS; Olsson N; Wagar LE; Haabeth OA; Chen B; Swaminathan K; Rawson K; Liu CL; Steiner D; Lund P; Rao S; Zhang L; Marceau C; Stehr H; Newman AM; Czerwinski DK; Carlton VE; Moorhead M; Faham M; Kohrt HE; Carette J; Green MR; Davis MM; Levy R; Elias JE; Alizadeh AA, Antigen presentation profiling reveals recognition of lymphoma immunoglobulin neoantigens. Nature 2017, 543, (7647), 723–727. [PubMed: 28329770]

51. Schuster H; Peper JK; Bosmuller HC; Rohle K; Backert L; Bilich T; Ney B; Loffler MW; Kowalewski DJ; Trautwein N; Rabsteyn A; Engler T; Braun S; Haen SP; Walz JS; Schmid-Horch B; Brucker SY; Wallwiener D; Kohlbacher O; Fend F; Rammensee HG; Stevanovic S; Staebler A; Wagner P, The immunopeptidomic landscape of ovarian carcinomas. Proc Natl Acad Sci U S A 2017, 114, (46), E9942–e9951. [PubMed: 29093164]

52. Ahmad MT; Zhang P; Dufresne C; Ferrucci L; Semba RD, The human eye proteome project: Updates on an emerging proteome. Proteomics 2018, 18, (5–6), e1700394. [PubMed: 29356342]

53. Schwenk JM; Omenn GS; Sun Z; Campbell DS; Baker MS; Overall CM; Aebersold R; Moritz RL; Deutsch EW, The human plasma proteome draft of 2017: Building on the human plasma PeptideAtlas from mass spectrometry and complementary assays. J Proteome Res 2017, 16, (12), 4299–4310. [PubMed: 28938075]

54. Shores DR; Everett AD, Children as biomarker orphans: Progress in the field of pediatric biomarkers. J Pediatr 2018, 193, 14–20.e31. [PubMed: 29031860]

55. Lourido L; Ayoglu B; Fernandez-Tajes J; Oreiro N; Henjes F; Hellstrom C; Schwenk JM; Ruiz-Romero C; Nilsson P; Blanco FJ, Discovery of circulating proteins associated to knee radiographic osteoarthritis. Sci Rep 2017, 7, (1), 137. [PubMed: 28273936]

56. Fernandez-Puente P; Calamia V; Gonzalez-Rodriguez L; Lourido L; Camacho-Encina M; Oreiro N; Ruiz-Romero C; Blanco FJ, Multiplexed mass spectrometry monitoring of biomarker candidates for osteoarthritis. J Proteomics 2017, 152, 216–225. [PubMed: 27865793]

57. Geraldino-Pardilla L; Russo C; Sokolove J; Robinson WH; Zartoshti A; Van Eyk J; Fert-Bober J; Lima J; Giles JT; Bathon JM, Association of anti-citrullinated protein or peptide antibodies with left ventricular structure and function in rheumatoid arthritis. Rheumatology (Oxford) 2017, 56, (4), 534–540. [PubMed: 27994093]

58. Sierra-Sanchez A; Garrido-Martin D; Lourido L; Gonzalez-Gonzalez M; Diez P; Ruiz-Romero C; Sjober R; Droste C; De Las Rivas J; Nilsson P; Blanco F; Fuentes M, Screening and validation of novel biomarkers in osteoarticular pathologies by comprehensive combination of protein array technologies. J Proteome Res 2017, 16, (5), 1890–1899. [PubMed: 28379711]

59. Velasquez E; Nogueira FCS; Velasquez I; Schmitt A; Falkai P; Domont GB; Martins-de-Souza D, Synaptosomal proteome of the orbitofrontal cortex from schizophrenia patients using quantitative label-free and iTRAQ-based shotgun proteomics. J Proteome Res 2017, 16, (12), 4481–4494. [PubMed: 28949146]

60. Leanza L; Romio M; Becker KA; Azzolini M; Trentin L; Manago A; Venturini E; Zaccagnino A; Mattarei A; Carraretto L; Urbani A; Kadow S; Biasutto L; Martini V; Severin F; Peruzzo R; Trimarco V; Egberts JH; Hauser C; Visentin A; Semenzato G; Kalthoff H; Zoratti M; Gulbins E; Paradisi C; Szabo I, Direct pharmacological targeting of a mitochondrial ion channel selectively kills tumor cells in vivo. Cancer Cell 2017, 31, (4), 516–531.e10. [PubMed: 28399409]

61. Wang J; Ma Z; Carr SA; Mertins P; Zhang H; Zhang Z; Chan DW; Ellis MJ; Townsend RR; Smith RD; McDermott JE; Chen X; Paulovich AG; Boja ES; Mesri M; Kinsinger CR; Rodriguez H; Rodland KD; Liebler DC; Zhang B, Proteome profiling outperforms transcriptome profiling for coexpression based gene function prediction. Mol Cell Proteomics 2017, 16, (1), 121–134. [PubMed: 27836980]

62. Zhang H; Liu T; Zhang Z; Payne SH; Zhang B; McDermott JE; Zhou JY; Petyuk VA; Chen L; Ray D; Sun S; Yang F; Chen L; Wang J; Shah P; Cha SW; Aiyetan P; Woo S; Tian Y; Gritsenko MA; Clauss TR; Choi C; Monroe ME; Thomas S; Nie S; Wu C; Moore RJ; Yu KH; Tabb DL; Fenyo D; Bafna V; Wang Y; Rodriguez H; Boja ES; Hiltke T; Rivers RC; Sokoll L; Zhu H; Shih IM; Cope L; Pandey A; Zhang B; Snyder MP; Levine DA; Smith RD; Chan DW; Rodland KD, Integrated proteogenomic characterization of human high-grade serous ovarian cancer. Cell 2016, 166, (3), 755–765. [PubMed: 27372738]

63. Yu KH; Levine DA; Zhang H; Chan DW; Zhang Z; Snyder M, Predicting ovarian cancer patients' clinical response to platinum-based chemotherapy by their tumor proteomic signatures. J Proteome Res 2016, 15, (8), 2455–65. [PubMed: 27312948]

64. Bosch LJW; de Wit M; Pham TV; Coupe VMH; Hiemstra AC; Piersma SR; Oudgenoeg G; Scheffer GL; Mongera S; Sive Droste JT; Oort FA; van Turenhout ST; Larbi IB; Louwagie J; van Criekinge W; van der Hulst RWM; Mulder CJJ; Carvalho B; Fijneman RJA; Jimenez CR; Meijer GA, Novel stool-based protein biomarkers for improved colorectal cancer screening: A case-control study. Ann Intern Med 2017, 167, (12), 855–866. [PubMed: 29159365]

65. Komor MA; Pham TV; Hiemstra AC; Piersma SR; Bolijn AS; Schelfhorst T; Delis-van Diemen PM; Tijssen M; Sebra RP; Ashby M; Meijer GA; Jimenez CR; Fijneman RJA, Identification of differentially expressed splice variants by the proteogenomic pipeline splicify. Mol Cell Proteomics 2017, 16, (10), 1850–1863. [PubMed: 28747380]

66. Huang KL; Li S; Mertins P; Cao S; Gunawardena HP; Ruggles KV; Mani DR; Clauser KR; Tanioka M; Usary J; Kavuri SM; Xie L; Yoon C; Qiao JW; Wrobel J; Wyczalkowski MA; Erdmann-Gilmore P; Snider JE; Hoog J; Singh P; Niu B; Guo Z; Sun SQ; Sanati S; Kawaler E; Wang X; Scott A; Ye K; McLellan MD; Wendl MC; Malovannaya A; Held JM; Gillette MA; Fenyo D; Kinsinger CR; Mesri M; Rodriguez H; Davies SR; Perou CM; Ma C; Townsend RR; Chen X; Carr SA; Ellis MJ; Ding L, Corrigendum: Proteogenomic integration reveals therapeutic targets in breast cancer xenografts. Nat Commun 2017, 8, 15479. [PubMed: 28440318]

67. Mazzucchelli G; Holzhauser T; Cirkovic Velickovic T; Diaz-Perales A; Molina E; Roncada P; Rodrigues P; Verhoeckx K; Hoffmann-Sommergruber K, Current (food) allergenic risk assessment: Is it fit for novel foods? Status quo and identification of gaps. Mol Nutr Food Res 2018, 62, (1).

68. Starr AE; Deeke SA; Li L; Zhang X; Daoud R; Ryan J; Ning Z; Cheng K; Nguyen LVH; Abou-Samra E; Lavallee-Adam M; Figeys D, Proteomic and metaproteomic approaches to understand host-microbe interactions. Anal Chem 2018, 90, (1), 86–109. [PubMed: 29061041]

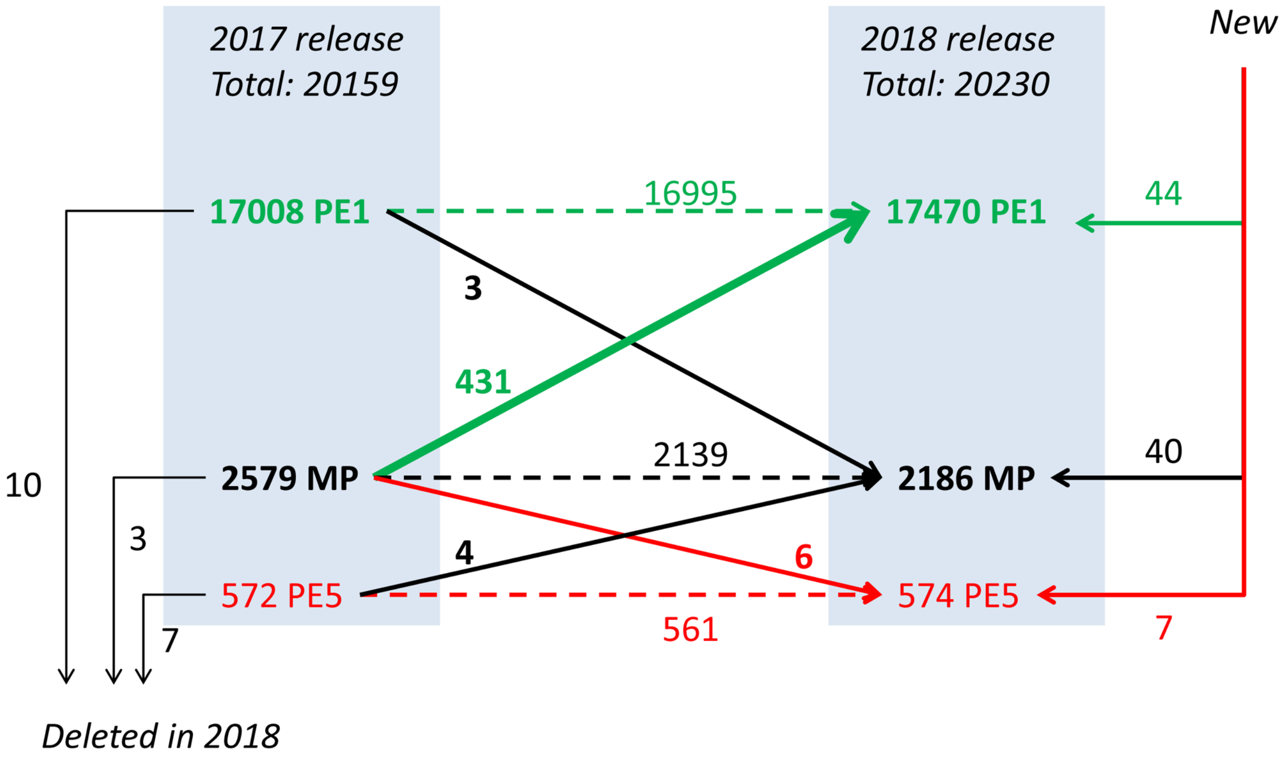


**Figure 1.**

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. See text for further discussion.

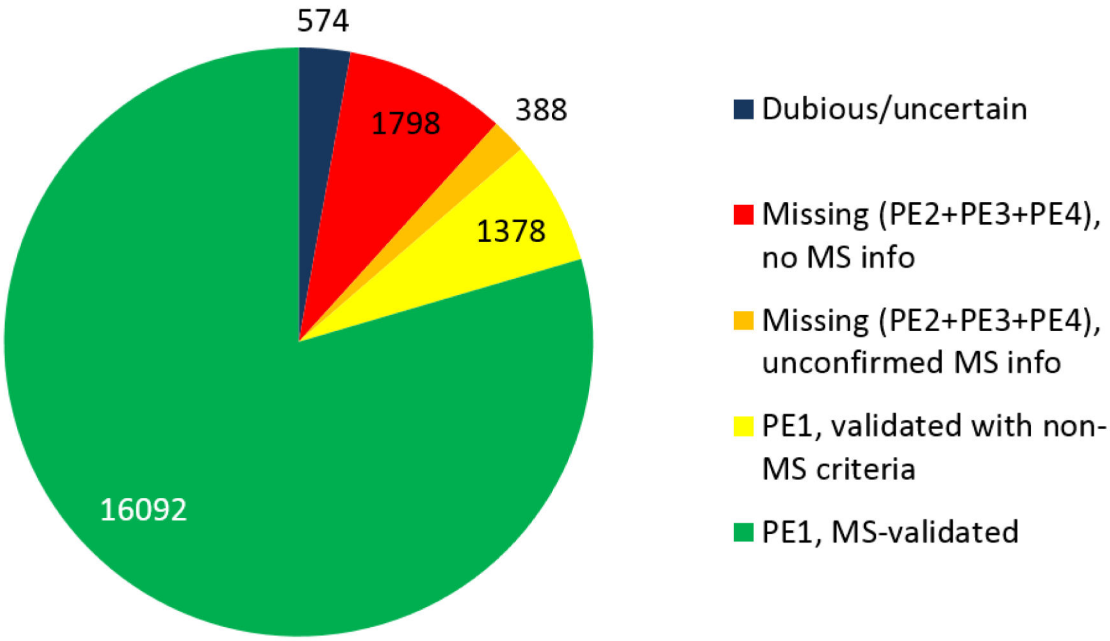


**Figure 2.**

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

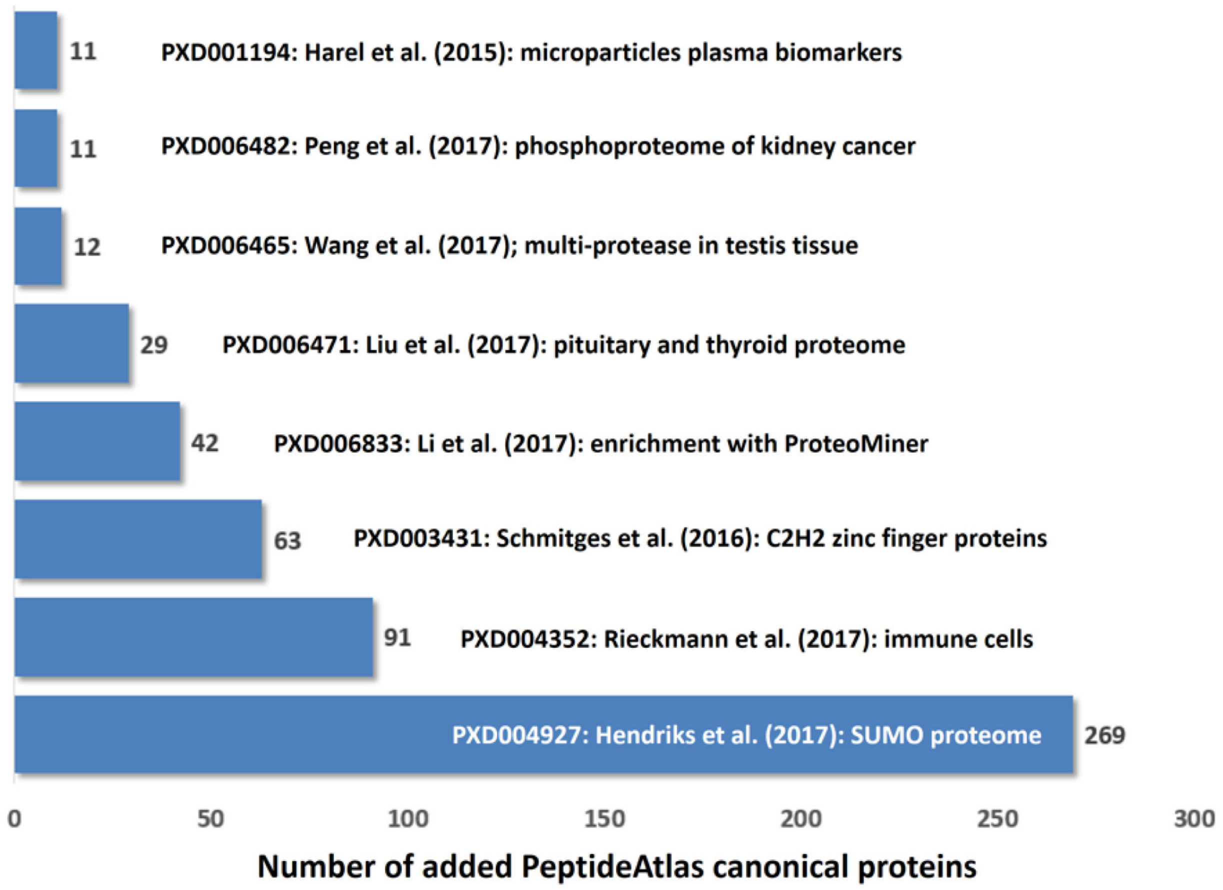


**Figure 3.**

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.[8]

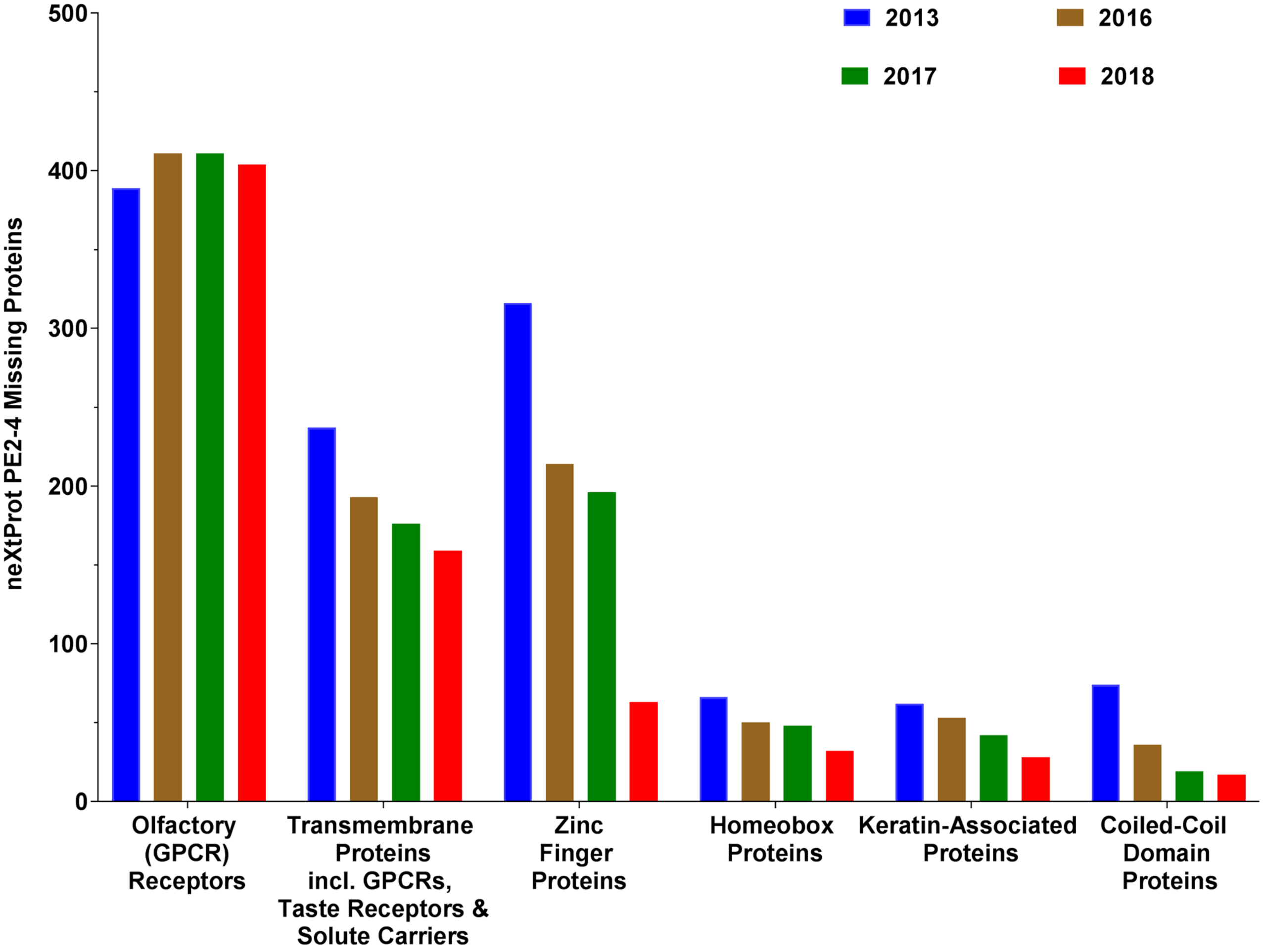


**Figure 4.**

Progress on identification of members of the six most numerous protein families in neXtProt Missing Protein categories PE2,3,4 from 2013 to 2018 (updated from Baker et al[20]).

**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.

| 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 | 2186 Missing Proteins[b] |
| 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 | |

[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.