HHS Public Access

Author manuscript

J Proteome Res. Author manuscript; available in PMC 2016 February 16.

Published in final edited form as:

J Proteome Res. 2015 September 4; 14(9): 3452–3460. doi:10.1021/acs.jproteome.5b00499.

Metrics for the Human Proteome Project 2015: Progress on the Human Proteome and Guidelines for High-Confidence Protein Identification

Gilbert S. Omenn[*,†], **Lydie Lane**[‡], **Emma K. Lundberg**[§], **Ronald C. Beavis**[‖], **Alexey I. Nesvizhskii**[⊥], and **Eric W. Deutsch**[#]

[†]Department of Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Avenue, Ann Arbor, Michigan 48109-2218, United States [‡]CALIPHO Group, Swiss Institute of Bioinformatics, Michel-Servet 1, 1211 Geneva 4, Switzerland [§]SciLifeLab Stockholm and School of Biotechnology, KTH, Karolinska Institutet Science Park, Tomtebodavägen 23, SE-171 65 Solna, Sweden [‖]Biochemistry & Medical Genetics, University of Manitoba, Winnipeg, MB, Canada R3T 2N2 [⊥]Pathology Department, University of Michigan, Medical Science Building 1, M4237, Ann Arbor, Michigan 48109-5602, United States [#]Institute for Systems Biology, 401 Terry Avenue North, Seattle, Washington 98109-5263, United States

Abstract

Remarkable progress continues on the annotation of the proteins identified in the Human Proteome and on finding credible proteomic evidence for the expression of “missing proteins”. Missing proteins are those with no previous protein-level evidence or insufficient evidence to make a confident identification upon reanalysis in PeptideAtlas and curation in neXtProt. Enhanced with several major new data sets published in 2014, the human proteome presented as neXtProt, version 2014-09-19, has 16 491 unique confident proteins (PE level 1), up from 13 664 at 2012-12 and 15 646 at 2013-09. That leaves 2948 missing proteins from genes classified having protein existence level PE 2, 3, or 4, as well as 616 dubious proteins at PE 5. Here, we document

[*]Corresponding Author: gomenn@umich.edu.

Notes

The authors declare no competing financial interest.

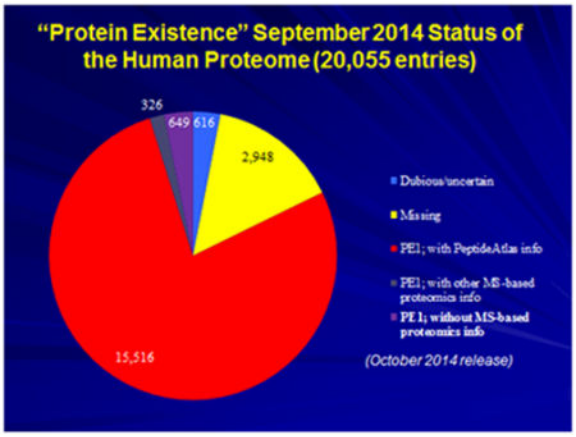
Supporting Information

Table S1: Intersection of canonical proteins in PeptideAtlas with Protein Evidence levels in neXtProt. The intersection of 14 730 canonical proteins in PeptideAtlas (version 2014-08) with PE 1 level evidence in neXtProt (version 2015-04-28) is tabulated here, as of 2015-06-30. Table S2: 20 PE 1 protein entries removed between neXtProt 2014-09-19 and institution of new criterion of two peptides of ≥7 aa OR one peptide of ≥9 aa by neXtProt, as of 2015-05. The difference is due to matches with a single peptide of only 7 or 8 amino acids. Table S3: 432 PE 1 protein entries different between the 2015-05 neXtProt criteria and the criteria for the 2015-03 build of PeptideAtlas, using “two peptides ≥9 amino acids” as criterion for canonical in PeptideAtlas. The stringency results primarily from the requirement for two peptides, whereas neXtProt can validate with only one uniquely mapping peptide of at least 9 aa. Table S4: 2675 PE1 protein entries in neXtProt not found among the 14 012 canonical proteins of PeptideAtlas 2014-08 updated with 2015-03 PeptideAtlas criteria of two peptides ≥9 aa. They may have only one MS peptide and/or be validated by other techniques, as explained in Table 2 and Figure 1. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.5b00499.

Up-to-date findings for each of the PE classes can be observed or downloaded at https://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/GetNextProtChromMapping?atlas_build_id=433&protein_evidence_level_constraint=protein&peptideatlas_constraint=01&redundancy_constraint=on&apply_action=QUERY.

the progress of the HPP and discuss the importance of assessing the quality of evidence, confirming automated findings and considering alternative protein matches for spectra and peptides. We provide guidelines for proteomics investigators to apply in reporting newly identified proteins.

Graphical Abstract



Keywords

Human Proteome Project; HPP metrics; guidelines; high-confidence protein identifications; neXtProt; PeptideAtlas; Human Protein Atlas; Global Proteome Machine database (GPMDB); missing proteins; novel proteins

INTRODUCTION

The HUPO (www.hupo.org) Human Proteome Project (HPP) (www.thehpp.org) has two overall goals: (1) stepwise completion of the Protein Parts List, or the draft human proteome, identifying and characterizing at least one protein product and as many PTM, SAP, and splice variant isoforms as feasible for each of the human protein-coding genes;[1] and (2) making proteomics an integrated counterpart to genomics throughout the biomedical and life sciences community through advances in instruments, assays, reagents, and proteomics knowledgebases for identification, quantitation, and characterization of proteins in network context in diverse biological systems.[2] The 50 HPP research teams are organized in the Chromosome-Centric C-HPP, the Biology and Disease-driven B/D-HPP, and the Antibody, Mass Spectrometry, and Knowledgebase resource pillars. This article is part of the third annual special issue of the *Journal of Proteome Research* organized by the C-HPP.

The aims of this article are to update progress on goal 1 and the core databases for the HPP and to provide new guidance for ensuring highly confident identification of previously missing proteins, as well as claims of novel peptides translated from lncRNAs or pseudogenes.

SEPTEMBER 2014 UPDATE OF THE HPP METRICS

The annual cycle for the HPP data has been organized around regular updates of PeptideAtlas (www.peptideatlas.org) and neXtProt (www.neXtProt.org), timed to be presented at the annual HUPO World Congress of Proteomics and used by the investigators submitting manuscripts to the C-HPP special issue in *Journal of Proteome Research*.[3–5] These databases, and complementary databases at GPMDB[6] (www.gpmdb.org) and Human Protein Atlas[7] (www.proteinatlas.org), provide chromosome-by-chromosome tabulations and facilitate the work of the C-HPP teams around the world. These databases also perform a critical role in subjecting the output from many laboratories to a standardized reanalysis with well-established statistical methods for high-confidence, corresponding to 1% false discovery rate (FDR) at the protein level for PeptideAtlas (canonical),[8] threshold criteria for highest evidence code EC4 (color-coded green) in GPMDB,[6] and supportive antibody-based evidence at Human Protein Atlas.[7]

Table 1 shows the HPP metrics from these databases for 2012, 2013, and 2014, with major progress each year. neXtProt 2014-09-19 has 16 491 PE1 entries for proteins, with a denominator of 19 439 predicted protein entries from protein existence levels PE1, PE2, PE3, and PE4 genes (85%), excluding the 616 PE5 entries for dubious proteins, as discussed by Lane et al.[4] The remaining 15% are currently “missing proteins”. As shown in Table 2 and Figure 1, neXtProt integrates mass spectrometry and several other types of experimental evidence in curating the protein entries. neXtProt provides extensive resources for protein variants and post-translational modifications plus advanced search and retrieval tools, such as SPARQL query, as updated by Gaudet.[5] The UniProtKB (www.uniprot.org) release of 2014-08[9] contained 13 988 human entries validated at the protein level, so neXtProt 2014-09-19 had integrated experimental evidence for 2503 additional entries.[5] Using mass spectrometry only, with different thresholds, PeptideAtlas 2014-08 designated 14 928 as canonical (FDR 1% at protein level, with Mayu adjustment) and GPMDB designated 15 459 as EC4 (green, peptides identified at least 5 times), whereas Human Protein Atlas v13 in November 2014 characterized 12 007 proteins with immunohistochemistry/immunofluorescence and identified 7024 additional protein entries with transcripts only.

Table 2 shows the protein existence status for all 20 055 entries in neXtProt, based on integrated information. There are 16 491 PE1 entries curated as having confident protein evidence, 2647 PE2 entries with transcript evidence only, 214 PE3 with extensive conservation in nonprimate species, 87 PE4 predicted from gene models often in nonhuman primates, and 616 PE5 genomic sequences predicting dubious or uncertain proteins.

Figure 1 shows that neXtProt utilizes PeptideAtlas spectral evidence of peptides in curation of 15 516 of the PE1 proteins (red wedge); it does not rely upon the PeptideAtlas designation of proteins as canonical. An additional 326 PE1 proteins (gray wedge) have MS-based proteomics information from sources other than PeptideAtlas; this number has decreased from 1071 a year earlier as PeptideAtlas has grown. The purple wedge of 649 scored PE1 by neXtProt without MS evidence includes 65 validated by Edman sequencing, 313 by biochemical characterization papers (directed mutagenesis, enzymatic assays), 83 by papers about PTMs, 76 by protein–protein interactions, 36 by immunohistochemistry, 35 by

experimental 3D structures, 1 by 2D gel electrophoresis, and 40 due to the existence of commercial antibodies used by HPA (prefixed by CAB instead of HPA in the Human Protein Atlas). Of note, this last criterion, initially established by UniProt several years ago, has just been removed from the PE assignment pipeline of UniProtKB and neXtProt. The yellow wedge of 2948 is the combination of PE levels 2, 3, and 4, for which evidence of transcript expression or expression of homologues in other species provides guidance to appropriate tissue specimens for proteomic studies. Finally, the blue wedge has the 616 PE5 dubious protein entries.

Table 3 shows the PE status of proteins chromosome-by-chromosome from neXtProt 2014-09-19 to guide the C-HPP teams' search for missing proteins and from Human Protein Atlas 2014-11-07 to show details of tissue expression and subcellular localization. Corresponding chromosome-by-chromosome tabulations appear in GPMDB and PeptideAtlas. Of course, the numbers of identified and missing proteins continue to be moving targets as new studies are conducted and data sets are shared through ProteomeXchange[10] for systematic review and standardized reanalysis. There are periodic updates of all of the databases, including the UniProt/SwissProt curation process that feeds into neXtProt, with lags between and across the databases.

It is important to recognize that a considerable fraction of the total predicted proteins may not be detectable with mass spectrometry based on tryptic peptides. From some protein sequences, no tryptic peptides suitable for detection in the mass spectrometer can be generated; for peptides from highly homologous proteins, the protein matches may be indistinguishable, leading to choice of just one as a representative protein under rules of parsimony (e.g., for PeptideAtlas, see Deutsch et al.,[8] this issue); for hydrophobic proteins embedded in membranes, special solubilization steps may be required; for genes with no detectable transcripts in human specimens (PE levels 3, 4, 5), it is very unlikely that proteins will be found in those specimens; many proteins may be expressed at levels below the detection limit of the methods, putting a premium on enrichment of sample fractions and sensitivity of analysis, and some proteins may be expressed only in unusual tissues or times of life or upon induction by infection or inflammation or other important biological processes.

MAJOR NEW DATASETS PUBLISHED IN 2014

We highlight here four major data sets subjected to the standard HPP reanalysis by PeptideAtlas and by GPMdb: multiprotease digestion to overcome limitations of trypsin, using HeLa cells[11]; proteomics studies from CPTAC/TCGA/NCI/NIH for colon and breast cancers (http://cancergenome.nih.gov); a Human Proteome Map from Kim et al. (Pandey lab),[12] PXD000561; and a Draft Human Proteome from Wilhelm et al.[13] (Kuster lab), PXD000865. Kim et al.[12] and Wilhelm et al.[13] set a fine example by making their extensive experimental data sets publicly available for reanalysis and focused studies.

As presented in detail by Deutsch et al.,[8] these data sets generated large increases in distinct peptides but much smaller increments in high-confidence canonical protein identification in PeptideAtlas. In PeptideAtlas 2014-08 there are successive increments of 541, 591, 231, and

2 proteins for the CPTAC, Kim et al.,[12] Wilhelm et al.,[13] and Guo et al.[11] data sets, respectively. With the 2015-05 PeptideAtlas revised criteria, these data sets contributed 516, 377, 110, and 4 additional proteins, respectively (Deutsch et al.,[8] Figure 1).

The use of various combinations of 1 to 3 of seven proteolytic enzymes (ArgC, AspN, GluC, Lys-C, chymotrypsin, elastase, trypsin) significantly increased protein sequence coverage at 1% protein FDR. However, as just noted, only very few previously undetected proteins were identified.[11]

The National Cancer Institute Clinical Proteomics Tumor Analysis Consortium (CPTAC) publicly released proteomic data produced from TCGA colorectal tumor samples in 2013[14] and from breast tumor specimens in 2014 [weblink above], with a very substantial increment in protein identifications as already noted. Findings from colon and breast were incorporated into PeptideAtlas 2014-08 and thereby into neXtProt 2014-09-19. Subsequently, results from high-grade serous ovarian cancer samples were announced for release, which have been included in PeptideAtlas 2015.[8]

Kim et al.[12] analyzed 30 normal samples: 17 adult and 7 fetal tissues plus 6 hematopoietic cell types. All mass spectrometry was performed in one lab with Orbitrap Elite and Velos instruments and SEQUEST and MASCOT search engines. Full data were made available via ProteomeXchange as Human Proteome Map (PXD000561). Peptide evidence was matched to 17 294 of 20 687 protein-coding genes (82%), including 2535 of the HPP 3844 then-missing proteins in the 2013 version of the HPP proteome map;[4] 1537 had expression in only one of the 30 specimens; 735 genes were expressed 10-fold in fetal samples compared with adult tissues and cells. Protein coexpression predicted protein–protein interactions better than transcript coexpression. Splice isoform-specific peptides were noted for 2861 protein isoforms from 2450 genes. Western blots using antibodies against 32 proteins gave tissue-specific matches only for 8. Extensive proteogenomic analyses deduced matches to 9 noncoding RNAs, 140 pseudogenes, 44 ORFs, and various N- and C-terminal sequences. Subsets of these results were published as part of the C-HPP 2014 *Journal of Proteome Research* special issue by Pinto et al.[15] for chromosome 22 and by Manda et al.[16] for chromosome 12.

There are some major surprises in the Kim et al.[12] study. First, quite lax filters were employed to control false-discovery, 1% for 25 million PSM, 1% for 293 000 peptides, and no FDR filter at all for proteins. A match with either search engine was considered to be sufficient. Second, a minimal peptide length of 6 aa was accepted. Many protein matches were based on a single peptide. Third, no comparative analysis using standard HPP metrics was employed, and unlikely identifications were not scrutinized closely for alternative explanations of the spectral and peptide matches. [See below.]

Wilhelm et al.[13] reported evidence for 18 097 of 19 629 SwissProt protein-coding genes (92%). The publicly available data (PXD000865) at ProteomicsDB are from their own analyses of 60 tissues, 13 body fluids, 147 cell types, and 1300 affinity purification assays (40%), while their analysis included 60% from extremely heterogeneous literature and repositories imported into their ProteomicsDB database. They claimed identification of 97%

of 13 378 PE1, 84% of 5531 PE2, 52% of 159 PE3, 64% of 72 PE4, and even 56% of 489 PE5 entries. They used 1% FDR for 1.1 billion PSM, 5% FDR for peptides (minimal length 7 aa), and no FDR for proteins, although they did deploy an early version of their picked target-decoy approach comparing pairs of observed and decoy sequences, now published[17] (see below). They chose to protect against true-positives being removed. They did not compare their 40% of the data with the miscellaneous 60% or with the Kim et al.[12] single-lab results, PeptideAtlas, or neXtProt. Andromeda and MASCOT were used as search engines; a "hit" with either was considered to be sufficient to claim a protein match.

A NEW PHASE FOR THE HUMAN PROTEIN ATLAS

Version 13 of the HPA was released on 7 November 2014, presenting a tissue-based map of the human proteome and a hard copy poster in Science magazine obtainable from www.proteinatlas.org. The Tissue-based Map of the Human Proteome appeared in *Science* on 23 January 2015,[7] with extensive annotations for tissue-elevated, secreted, membrane-spanning, housekeeping, regulatory, druggable, cancer, cell line, isoform, and metabolism-related proteomes based on immunohistochemical studies of 44 tissues and RNA sequencing results for 32 of the tissues. Investigators annotating missing proteins can search HPA for spatial protein expression data on a single-cell level. The tissues with the most mRNAs highly enriched (at least 5× the levels in all other tissues) are testis (999) and brain (318). Several tissue-specific HPA publications have been published, and more are expected during 2015 with tissue-specific immunohistochemical and transcript evidence to guide specimen selection for further MS studies, for example, describing the brain,[18] liver,[19] testis,[20] kidney,[21] pancreas,[22] skin,[23] adipose tissue,[24] gallbladder,[25] lung,[26] gastrointestinal tract,[27] and cardiac and skeletal muscle.[28] HPA, v14, will be released in Fall 2015 along with a novel Rodent Brain Atlas comprising immunofluorescently stained whole-mouse-brain sections. This provides an in-depth view of protein expression in the specialized cells of the mammalian brain and serves as a complement to the limited human brain tissue samples in HPA. HPA also continues to strive toward high-level validation now in terms of colocalization analysis with GFP-tagged human proteins at endogenous levels. At the C-HPP Workshop and EuPA annual meeting in Milan 23–28 June 2015, the Knockdown Initiative was launched under the auspices of HPA, EuPA, and HUPO/HPP. It is a private–public partnership for systematic exploration of antibodies utilizing siRNA and other gene-editing methods to validate the antibody specificity.

PrEST antigens are now being utilized in mass spectrometry studies;[29] further collaborations across antibody profiling and MS would enhance the confidence of findings from both approaches. Such a study has been initiated with Yamamoto and colleagues (unpublished, personal communication) on the human kidney glomerulus.

MEANS OF ENHANCING THE CONFIDENCE AND CREDIBILITY OF PROTEIN IDENTIFICATIONS, ESPECIALLY FOR PREVIOUSLY MISSING PROTEINS

The HPP investigators at a series of workshops and at the HUPO2014 World Congress have given major consideration to the reasons why some gene products have no credibly detected transcripts or proteins. There are many reasons why the gene may not be transcribed, the transcript may not be translated, the protein may not be abundant enough or solubilized sufficiently, or the protein may not be detectable by trypsin-based mass spectrometry or with antibody-based immunohistochemistry or immunofluorescence.

With regard to missing proteins, when routine search engine matches identify peptide sequences from spectra and protein matches from peptide sequences that no other investigator has observed, there should be stringent criteria for confirmation and acceptance of those matches and reports of the findings. Scrutinizing the detailed features of the spectra and the match to one or more peptide sequences comes first. Missing peaks or additional peaks must be accounted for. Alternative explanations for a first-ever-observed peptide match, especially for relatively short peptides, must be considered. If a single amino acid difference due to a SNP would permit the match to a common peptide from an abundant protein, then that should be considered a more probable match than a once-ever match. Sometimes only a single reference proteome is used for searches, as in the case of RefSeq[30] by Kim et al.,[12] when multiple novel matches already existed in UniProt/SwissProt/neXtProt. In some cases, the reference protein is itself an SAAV or splice variant, and the common form of the protein is proposed to be a novel finding. Another major alternative explanation involves PTMs, which may produce exact or nearly exact *m*/*z* values for a common peptide otherwise attributed to a novel peptide. See examples below from the proteogenomics analyses of Nesvizhskii[31] and from the deep dives in PeptideAtlas by Deutsch et al.[8] involving transferrins, actins, keratins, and even porcine trypsin.

A matter of general interest is the analysis of families of proteins with high sequence similarity; these account for many of the indistinguishable representative and marginally distinguished matches in PeptideAtlas and all other databases and experimental data sets. For many years, the common practice has been to select one representative protein from a protein family for which available peptides do not distinguish the matches. Different researchers and different search engines make different choices without sufficiently considering this category of overcounting. This biological feature may make such missing proteins un-identifiable, unless truly uniquely mapping or proteotypic peptides can be found to differentiate the family members.

A biological interest in particular sets of genes and proteins can guide a series of deep dives on newly identified protein matches. For example, Ezkurdia et al.[32] promptly examined the Kim et al.[12] and Wilhelm et al.[13] data sets for the surprising claims of 108 and 200 olfactory receptor genes expressed as proteins, respectively. Ezkurdia et al.[32] examined all of the spectra provided for these proteins and declared none to be sufficient to justify the peptide and protein identifications. For Kim et al.,[12] 40 matches were ambiguous and the other 68

had poor spectra. In Wilhelm et al.,[13] OR6J1 had 8 spectra, none of which survived manual inspection. Deutsch et al.[8] (this issue) concluded that no olfactory receptor proteins have been credibly identified thus far, including two that were in previous versions of PeptideAtlas.

Our search of GPMDB for olfactory receptor turned up 36 pages of results, covering 718 proteins from 425 distinct entries. There are six green (good quality) proteins, five of which are based on single, small, low-complexity peptides that can be assigned to other proteins. For instance, OR1M1's peptide ILVAIMK can be assigned just as well to LIVALMK (from the abundant protein annexin A5). Another finding involves OR51E2, which has multiple observations of a single peptide, GSLFFFPLPLLIK from a single study of seminal plasma.[33] The spectra are high-quality. Further search revealed that an older gene symbol for OR51E2 is PSGR, a prostate cancer-specific G protein-coupled receptor.[34] The most likely cell types to transcribe and translate olfactory receptor genes lie in the olfactory cortex of the brain and possibly the olfactory epithelium of the upper nasal passages. No proteomic analyses of such specimens are yet available.

Beavis reported in GPMDB (2014-07-01) an interesting meta-analysis on the 53 Y chromosome-specific genes, which would be expected to be expressed only in males. Kim et al.[12] claimed to find 32 Y-gene protein products in the ovary specimen. Wilhelm et al.[13] reported 7 such predicted proteins expressed in ovary. In this case, it is a simple process to exclude Y-chromosome proteins from the FASTA file used for female samples or to remove them from the results postanalysis. Many of the identifications were based on single peptides, making them indistinguishable in light of homologous sequences. As of May 2015, GPMDB reports only 45 Y-chromosome genes, of which all but 5 may have homologous X-chromosome genes. Such probes of big data sets give a useful sampling of the likely credibility of automated peptide and protein matches and reveal the need for special scrutiny as we search for the remaining missing proteins.

From the point of view of HPP, all investigators have an opportunity to process their MS findings through the Trans-Proteomic Pipeline (with Mayu adjustment for large data sets[35]) developed for PeptideAtlas and the X!Tandem pipeline developed for GPMDB. It would have been helpful if Kim et al.[12] and Wilhelm et al.[13] had reported findings with the consensus FDR thresholds as well as their own preferred methods and then did a deep dive on the quality of the evidence and the potential alternative matches for the differentially identified proteins.

Independent reanalyses by PeptideAtlas and GPMDB and also Tress, Ezkurdia, and colleagues in Spain (presented at US HUPO March 2015) generated about 13 000 to 14 000 high-quality protein matches from each of these large data sets. Conversely, when Beavis used the filters of Wilhelm et al.[13] on the very large GPMDB database,[6,36] he obtained matches for 97% of the predicted proteome. Similar very high numbers, of course, can be obtained by using the yellow and red classes of findings in GPMDB, rather than just the high-confidence GPMDB green class included in the successive annual metrics papers from the HPP (Table 1), or the ambiguous and redundant entries in PeptideAtlas.

Cox[37] applied a MaxQuant-based two-tier decoy-based filtering method with FDR control of PSMs and protein groups, calculating a posterior error probability (PEP) for each PSM and generating a protein group FDR. Priority is given to matches to validated SwissProt/neXtProt entries over claims of novel and unconventional protein translation. On first pass, Cox obtained 13 242 protein groups from Kim et al.,[12] which was reduced to 11 206 when matched to Ensembl[38] genes. He did not find a single olfactory receptor at 1% protein FDR.

Finally, the Kuster lab[17] published a reassessment of their own Wilhelm et al.[13] analysis. First they claimed that the classical target-decoy method for protein FDR eliminates a very high percentage of true positives in large heterogeneous data sets; however, no such phenomenon has been seen in PeptideAtlas or in GPMDB. More remarkably, their reanalysis of their own data in ProteomicsDB yielded only 14 714 (instead of 18 097 reported by Wilhelm et al.[12]), only modestly more than the classic FDR method (even without Mayu adjustment) at 14 035. When they compared with methods for the combined data sets (Kim et al.,[12] Wilhelm et al.[13]), the results were 15 375 versus 14 638.

A biologically informed assessment of the picked target-decoy results might begin by asking about the credibility of the 200 olfactory receptor proteins or the large number of pseudogene translated products reported in Wilhelm et al.[13]

The HPP decision at HUPO 2013 to put aside the PE5 class of dubious proteins was an explicit action to direct priority to PE2–3–4 genes, especially PE2 genes with evidence of transcript expression in various human tissues suitable for focused proteomic studies by both MS and antibodies. With regard to PE5, many of which are classified as pseudogenes or lncRNAs, we sought to raise the bar for evidence claiming protein existence in the absence of transcript expression and with dubious or uncertain status for the predicted proteins themselves. neXtProt has recently submitted 10 PE5 proteins to SwissProt curators for consideration of reclassification. We recommend careful scrutiny of claims of protein expression from PE5 genes, preferably with orthogonal experimental methods like SRM or SWATH and with examination of alternative database matches.

A PROPOSAL TO STRENGTHEN THE GUIDELINES FOR CLAIMS OF FINDINGS OF MISSING PROTEINS

Kim et al.[12] and Wilhelm et al.[13] leaders Pandey and Kuster have challenged the use of protein FDR for very large heterogeneous data sets. We concluded that more detailed guidance on confirming potential protein matches from mass spectrometry is needed.

On behalf of the HPP Knowledgebase Resource Pillar, Eric Deutsch at PeptideAtlas has proposed the following more demanding guidelines for acceptance of any initial claim of identification corresponding to a previously unobserved protein sequence. See the PeptideAtlas 2015 paper,[8] which documents the effects of applying the new guidelines to the entire 2014-08 PeptideAtlas.

1. Individual data sets must be thresholded at 1% protein FDR, as now. In addition, estimate the peptide and PSM level FDRs and clearly state them in the paper, with the numbers of proteins, peptides, and spectra that passed and failed the thresholds,

compared with number of novel “hits”. As discussed below, large-scale heterogeneous data sets require a modification of the target-decoy FDR.

2. Raise the minimum threshold to two uniquely mapping peptides of nine or more amino acids and make the spectra publicly available. There could be exceptions for proteins that simply do not generate two uniquely mapping peptides upon tryptic digestion for mass spectrometry. In PeptideAtlas, there are many short peptides of 7 or 8 aa length that yield a high-quality Mascot score, often without many peaks, which may be wrong or may be explained better by other protein matches. Kim et al.[12] and Wilhelm et al.[13] accepted peptides of 6 or 7 aa, respectively.

   neXtProt, led by Lydie Lane and Amos Bairoch, participated in discussions and analyses of the proposed new threshold. neXtProt has chosen for its 2015-03 update to utilize a threshold of two uniquely mapping peptides of 7 or more amino acids in length or one such peptide with 9 or more amino acids in length. Only 20 neXtProt 2014-09-19 proteins were excluded with these new neXtProt criteria; basically, there were 20 with only one proteotypic peptide of 7 or 8 aa (listed in Supporting Information Table S2). The revised PeptideAtlas criteria are much more stringent; 432 proteins that are validated with neXtProt’s new criteria would not be validated if neXtProt used the PeptideAtlas criteria of two peptides ≥9 aa (see Supporting Information Table S3). Most of the decrease in canonical proteins in the PeptideAtlas update of 2015-03 is due to downgrading a significant number of canonical matches to “weak” in the new scheme.[8] Directly comparing numbers of PE1 proteins in neXtProt with the number of canonical proteins in PeptideAtlas is complicated by the additional categories of indistinguishable representative (ambiguous) and indistinguishable or marginally distinguishable (redundant) in the new PeptideAtlas as well as inclusion of immunoglobulins from SwissProt and a few highly rated proteins from IPI in PeptideAtlas. From this vantage point, there are 14 012 entries that are PE1 in the neXtProt 2015-05 and have at least two proteotypic peptides of ≥9 aa; that means there are 2675 proteins that are PE1 but would not comply with this criterion (Supporting Information Table S4). They may have only one MS peptide and/or be validated by other techniques, as explained in Table 2 and Figure 1.

   Discussions by the HPP Executive Committee were supportive of such thresholds with higher quality and fewer false-positives. Having a more stringent threshold at PeptideAtlas than at neXtProt is understandable and useful, as raw data files submitted through ProteomeXchange are reanalyzed at PeptideAtlas and available for scrutiny. Then, neXtProt uses those identified peptide sequences, together with other sources of curated information about the proteins, to rank evidence of protein existence in the PE1–5 system shown in Tables 2 and 3. neXtProt does not rely upon the protein classification in PeptideAtlas. Investigators seeking information about their own peptide findings can search PeptideAtlas by peptide and then see whether the claimed uniquely mapping peptides were of sufficient quality and uniqueness to generate a canonical or indistinguishable protein match in PeptideAtlas and then PE1 status in neXtProt.

The HPP-EC also supports use of orthogonal methods to confirm peptide identifications, especially the proteome-wide resources of SRM peptides, spectral library, and SRM/ PASSEL knowledgebase developed under the aegis of the B/D-HPP and the related methods of SWATH-MS. Several authors in this *Journal of Proteome Research* special issue have utilized SRM and/or synthetic peptide spectral matching to confirm findings of missing proteins.

3. During data analysis, it is not acceptable to exclude protein sequences expected to be present in the sample, thereby reducing the search database to be searched to those missing. Matching only to missing protein sequences means that more credible matches to known proteins will be missed.

The Spanish chromosome 16 team provided a valuable example for all the chromosome teams by annotating the 108 missing proteins coded by genes on Chr 16, using standard sources for prediction of transmembrane motifs, signal peptide, protein domains, and protein families. They compared findings in neXtProt, PA, and GPMDB, noting 16 proteins ranked high in GPMDB that were missing in PA and neXtProt. They also estimated the probability of finding a missing protein in a given cell type/tissue from analyses of >3000 transcriptomic experiments in cells, normal tissues, and cancers. They shared a work-flow for MS data analysis. They offered to give other chromosome teams a head start with a preliminary similar tabulation, as will be discussed at HUPO2015 in Vancouver.

PERSPECTIVES FROM PROTEOGENOMICS ON QUALITY OF PROTEIN MATCHES

Nesvizhskii[31] discussed the challenge of false positives in proteogenomics studies that seek to identify novel peptides, i.e., peptides not present in the major reference protein sequence databases: RefSeq,[17] UniProt,[7] and Ensembl.[2] He also provided guidelines for analyzing the data and reporting the results of proteogenomics studies in the literature. Most importantly, when searching custom protein databases that include predicted protein sequences, data filtering and FDR estimation (e.g., using the target-decoy strategy) should be done separately for novel and known peptides (referred to as *class-specific* FDR). An alternative solution that also helps to ensure more accurate FDR estimates for novel peptides is to search custom protein databases as a second step in the analysis using spectra that remain unidentified after the initial search against a reference protein sequence database.

A second issue is specifically related to the identification of novel peptides that are highly homologous to peptides in a reference protein sequence data set. Note that this is the case for the majority of novel peptides in the pseudogene category. The rates of false identifications of homologous peptides are likely to be underestimated by all decoy database methods. All identifications of novel peptides with a high sequence homology (e.g., less than three amino acid difference) should be removed or at least scrutinized for alternative, more likely explanations.

A common problem is false identification of a novel peptide based on a spectrum for a chemically modified highly abundant peptide ion with a mass shift equaling the mass difference between the novel and unmodified known peptide. For example, Wilhelm et al.[13]

reported the identification of a peptide LATQLTGPVMPIR from a pseudogene sequence. However, this peptide is highly similar to peptide LATQLTGPVMPVR (V-to-I substitution; mass shift of ~14 Da) from a highly abundant ribosomal protein, RPL13. In this case, methylation on the neighboring arginine residue introduces a similar mass shift, and this modification is a more likely event than the identification of a pseudogene peptide. Of course, isoleucine/leucine substitutions cannot be distinguished using mass spectrometry and thus should not be reported as identifications of single amino acid variants, as was done in Zhang et al.[14]

When estimating protein-level FDR, it is important to keep in mind that the conventional target-decoy estimation strategy may not provide accurate estimates when the number of identified proteins is a high proportion of the target protein sequences in the searched protein sequence database. Instead, FDR can be estimated using the adjusted decoy counts with R-factor correction,[39] with a picked FDR strategy,[17] or with the MAYU method, as in the Trans-Proteomic Pipeline of PeptideAtlas. Each of these methods will be less stringent than the 1% FDR by target-decoy.

We have concluded that missing and novel proteins require higher stringency. Following the Nesvizhskii advice to utilize class-specific FDR, that approach could be applied to the different PE classes from neXtProt (PE 2+3+4; PE 5) for missing proteins analyses, after removing the PSMs and peptides that match to PE1 proteins. For novel peptides, possibly arising from PTMs or splice variants or lncRNAs, a similar separate analysis is desirable. We expect these matters to be a major topic of discussion throughout this year, including the HPP Workshop at the Vancouver HUPO2015 Congress.

CONCLUDING REMARKS

neXtProt, version 2014-09-19, was chosen as the baseline for the 2015 cycle of papers from the C-HPP teams and other authors. The HPP strongly encourages full sharing through ProteomeXchange of all data sets and accompanying metadata and highly recommends community-wide use of standardized reanalysis pipelines, attention to the enhanced guidelines proposed here, and confirmation of novel findings with SRM and SWATH-MS methods. When investigators prefer other thresholds, we strongly encourage comparison with the thresholds in the HPP guidelines.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We appreciate the guidance and comments from Amos Bairoch of neXtProt and the HPP Executive Committee. G.S.O. acknowledges grant support from National Institutes of Health grant U54ES017885 to the University of Michigan; E.W.D., NIH grants RO1GM087221, 2P50GM076547, and U54EB020406 to the Institute for Systems Biology and EU FP7 ProteomeXchange grant 260558 to the European Bioinformatics Institute; L.L., support from the Swiss Federation Commission for Technology and Innovation grant CTI 10214; and E.K.L., support from the Knut and Alice Wallenberg Foundation and EU 7th Framework.

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. [PubMed: 21742803]
2. Gaudet P, Argoud-Puy G, Cusin I, Duek P, Evalet O, Gateau A, Gleizes A, Pereira M, Zahn-Zabal M, Zwahlen C, Bairoch A, Lane L. neXtProt: organizing protein knowledge in the context of human proteome projects. J Proteome Res. 2013; 12(1):293–8. [PubMed: 23205526]
3. 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]
4. 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]
5. Gaudet P, Michel PA, Zahn-Zabal M, Cusin I, Duek PD, Evalet O, Gateau A, Gleizes A, Pereira M, Teixeira D, Zhang Y, Lane L, Bairoch A. The neXtProt knowledgebase on human proteins: current status. Nucleic Acids Res. 2015; 43(D1):D764–70. [PubMed: 25593349]
6. Fenyo D, Beavis RC. The GPMDB REST interface. Bioinformatics. 2015; 31:2056. [PubMed: 25697819]
7. 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]
8. Deutsch EW, Sun Z, Campbell D, Kusebauch U, Chu CS, Mendoza L, Shteynberg D, Omenn GS, Moritz RL. The state of the Human Proteome in 2015 as viewed through PeptideAtlas: enhacing accuracy and coverage. J Proteome Res. 201510.1021/acs.jproteome.5b00500
9. The UniProt Consortium. Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res. 2014; 42(Database issue):D191–8. [PubMed: 24253303]
10. Vizcaino JA, Deutsch EW, Wang R, Csordas A, Reisinger F, Rios D, Dianes JA, Sun Z, Farrah T, Bandeira N, Binz PA, Xenarios I, Eisenacher M, Mayer G, Gatto L, Campos A, Chalkley RJ, Kraus HJ, Albar JP, Martinez-Bartolome S, Apweiler R, Omenn GS, Martens L, Jones AR, Hermjakob H. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nat Biotechnol. 2014; 32(3):223–6. [PubMed: 24727771]
11. Guo X, Trudgian DC, Lemoff A, Yadavalli S, Mirzaei H. Confetti: a multiprotease map of the HeLa proteome for comprehensive proteomics. Mol Cell Proteomics. 2014; 13(6):1573–84. [PubMed: 24696503]
12. Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S, Thomas JK, Muthusamy B, Leal-Rojas P, Kumar P, Sahasrabuddhe NA, Balakrishnan L, Advani J, George B, Renuse S, Selvan LD, Patil AH, Nanjappa V, Radhakrishnan A, Prasad S, Subbannayya T, Raju R, Kumar M, Sreenivasamurthy SK, Marimuthu A, Sathe GJ, Chavan S, Datta KK, Subbannayya Y, Sahu A, Yelamanchi SD, Jayaram S, Rajagopalan P, Sharma J, Murthy KR, Syed N, Goel R, Khan AA, Ahmad S, Dey G, Mudgal K, Chatterjee A, Huang TC, Zhong J, Wu X, Shaw PG, Freed D, Zahari MS, Mukherjee KK, Shankar S, Mahadevan A, Lam H, Mitchell CJ, Shankar SK, Satishchandra P, Schroeder JT, Sirdeshmukh R, Maitra A, Leach SD, Drake CG, Halushka MK, Prasad TS, Hruban RH, Kerr CL, Bader GD, Iacobuzio-Donahue CA, Gowda H, Pandey A. A draft map of the human proteome. Nature. 2014; 509(7502):575–81. [PubMed: 24870542]
13. Wilhelm M, Schlegl J, Hahne H, Gholami AM, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H, Mathieson T, Lemeer S, Schnatbaum K, Reimer U, Wenschuh H, Mollenhauer M, Slotta-Huspenina J, Boese JH, Bantscheff M, Gerstmair A, Faerber F, Kuster B.

Mass-spectrometry-based draft of the human proteome. Nature. 2014; 509(7502):582–7. [PubMed: 24870543]

14. Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z, Chambers MC, Zimmerman LJ, Shaddox KF, Kim S, Davies SR, Wang S, Wang P, Kinsinger CR, Rivers RC, Rodriguez H, Townsend RR, Ellis MJ, Carr SA, Tabb DL, Coffey RJ, Slebos RJ, Liebler DC. Proteogenomic characterization of human colon and rectal cancer. Nature. 2014; 513(7518):382–7. [PubMed: 25043054]
15. Pinto SM, Manda SS, Kim MS, Taylor K, Selvan LD, Balakrishnan L, Subbannayya T, Yan F, Prasad TS, Gowda H, Lee C, Hancock WS, Pandey A. Functional annotation of proteome encoded by human chromosome 22. J Proteome Res. 2014; 13(6):2749–60. [PubMed: 24669763]
16. Manda SS, Nirujogi RS, Pinto SM, Kim MS, Datta KK, Sirdeshmukh R, Prasad TS, Thongboonkerd V, Pandey A, Gowda H. Identification and characterization of proteins encoded by chromosome 12 as part of chromosome-centric human proteome project. J Proteome Res. 2014; 13(7):3166–77. [PubMed: 24960282]
17. Savitski MM, Wilhelm M, Hahne H, Kuster B, Bantscheff M. A scalable approach for protein false discovery rate estimation in large proteomic data sets. Mol Cell Proteomics. 2015; M114:046995.
18. Sjostedt E, Fagerberg L, Hallstrom BM, Haggmark A, Mitsios N, Nilsson P, Ponten F, Hokfelt T, Uhlen M, Mulder J. Defining the Human Brain Proteome Using Transcriptomics and Antibody-Based Profiling with a Focus on the Cerebral Cortex. PLoS One. 2015; 10(6):e0130028. [PubMed: 26076492]
19. Kampf C, Mardinoglu A, Fagerberg L, Hallstrom BM, Edlund K, Lundberg E, Ponten F, Nielsen J, Uhlen M. The human liver-specific proteome defined by transcriptomics and antibody-based profiling. FASEB J. 2014; 28(7):2901–14. [PubMed: 24648543]
20. Djureinovic D, Fagerberg L, Hallstrom B, Danielsson A, Lindskog C, Uhlen M, Ponten F. The human testis-specific proteome defined by transcriptomics and antibody-based profiling. Mol Hum Reprod. 2014; 20(6):476–88. [PubMed: 24598113]
21. Habuka M, Fagerberg L, Hallstrom BM, Kampf C, Edlund K, Sivertsson A, Yamamoto T, Ponten F, Uhlen M, Odeberg J. The kidney transcriptome and proteome defined by transcriptomics and antibody-based profiling. PLoS One. 2014; 9(12):e116125. [PubMed: 25551756]
22. Danielsson A, Ponten F, Fagerberg L, Hallstrom BM, Schwenk JM, Uhlen M, Korsgren O, Lindskog C. The human pancreas proteome defined by transcriptomics and antibody-based profiling. PLoS One. 2014; 9(12):e115421. [PubMed: 25546435]
23. Edqvist PH, Fagerberg L, Hallstrom BM, Danielsson A, Edlund K, Uhlen M, Ponten F. Expression of human skin-specific genes defined by transcriptomics and antibody-based profiling. J Histochem Cytochem. 2015; 63(2):129–41. [PubMed: 25411189]
24. Mardinoglu A, Kampf C, Asplund A, Fagerberg L, Hallstrom BM, Edlund K, Bluher M, Ponten F, Uhlen M, Nielsen J. Defining the human adipose tissue proteome to reveal metabolic alterations in obesity. J Proteome Res. 2014; 13(11):5106–19. [PubMed: 25219818]
25. Kampf C, Mardinoglu A, Fagerberg L, Hallstrom BM, Danielsson A, Nielsen J, Ponten F, Uhlen M. Defining the human gallbladder proteome by transcriptomics and affinity proteomics. Proteomics. 2014; 14(21–22):2498–507. [PubMed: 25175928]
26. Lindskog C, Fagerberg L, Hallstrom B, Edlund K, Hellwig B, Rahnenfuhrer J, Kampf C, Uhlen M, Ponten F, Micke P. The lung-specific proteome defined by integration of transcriptomics and antibody-based profiling. FASEB J. 2014; 28(12):5184–96. [PubMed: 25169055]
27. Gremel G, Wanders A, Cedernaes J, Fagerberg L, Hallstrom B, Edlund K, Sjostedt E, Uhlen M, Ponten F. The human gastrointestinal tract-specific transcriptome and proteome as defined by RNA sequencing and antibody-based profiling. J Gastroenterol. 2015; 50(1):46–57. [PubMed: 24789573]
28. Lindskog C, Linne J, Fagerberg L, Hallstrom BM, Sundberg CJ, Lindholm M, Huss M, Kampf C, Choi H, Liem DA, Ping P, Varemo L, Mardinoglu A, Nielsen J, Larsson E, Ponten F, Uhlen M. The human cardiac and skeletal muscle proteomes defined by transcriptomics and antibody-based profiling. BMC Genomics. 2015; 16(1):475. [PubMed: 26109061]
29. Zeiler M, Straube WL, Lundberg E, Uhlen M, Mann M. A Protein Epitope Signature Tag (PrEST) library allows SILAC-based absolute quantification and multiplexed determination of protein copy numbers in cell lines. Mol Cell Proteomics. 2012; 11(3):O111.009613. [PubMed: 21964433]

30. Frankish A, Uszczynska B, Ritchie GR, Gonzalez JM, Pervouchine D, Petryszak R, Mudge JM, Fonseca N, Brazma A, Guigo R, Harrow J. Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction. BMC Genomics. 2015; 16(Suppl 8):S2. [PubMed: 26110515]

31. Nesvizhskii AI. Proteogenomics: concepts, applications and computational strategies. Nat Methods. 2014; 11(11):1114–25. [PubMed: 25357241]

32. Ezkurdia I, Vazquez J, Valencia A, Tress M. Analyzing the first drafts of the human proteome. J Proteome Res. 2014; 13:3854. [PubMed: 25014353]

33. Batruch I, Smith CR, Mullen BJ, Grober E, Lo KC, Diamandis EP, Jarvi KA. Analysis of seminal plasma from patients with non-obstructive azoospermia and identification of candidate biomarkers of male infertility. J Proteome Res. 2012; 11(3):1503–11. [PubMed: 22188163]

34. Xu LL, Stackhouse BG, Florence K, Zhang W, Shanmugam N, Sesterhenn IA, Zou Z, Srikantan V, Augustus M, Roschke V, Carter K, McLeod DG, Moul JW, Soppett D, Srivastava S. PSGR, a novel prostate-specific gene with homology to a G protein-coupled receptor, is overexpressed in prostate cancer. Cancer Res. 2000; 60(23):6568–72. [PubMed: 11118034]

35. Reiter L, Claassen M, Schrimpf SP, Jovanovic M, Schmidt A, Buhmann JM, Hengartner MO, Aebersold R. Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. Mol Cell Proteomics. 2009; 8(11):2405–17. [PubMed: 19608599]

36. Perez-Riverol Y, Alpi E, Wang R, Hermjakob H, Vizcaino JA. Making proteomics data accessible and reusable: current state of proteomics databases and repositories. Proteomics. 2015; 15(5–6): 930–950. [PubMed: 25158685]

37. Cox, J. Controlling false discovery rates (FDRS) in genome-wide proteomics datasets. Abstract O-9, EuPA Annual Meeting; Milan, Italy. June 24, 2015;

38. Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, Gil L, Giron CG, Gordon L, Hourlier T, Hunt S, Johnson N, Juettemann T, Kahari AK, Keenan S, Kulesha E, Martin FJ, Maurel T, McLaren WM, Murphy DN, Nag R, Overduin B, Pignatelli M, Pritchard B, Pritchard E, Riat HS, Ruffier M, Sheppard D, Taylor K, Thormann A, Trevanion SJ, Vullo A, Wilder SP, Wilson M, Zadissa A, Aken BL, Birney E, Cunningham F, Harrow J, Herrero J, Hubbard TJ, Kinsella R, Muffato M, Parker A, Spudich G, Yates A, Zerbino DR, Searle SM. Ensembl 2014. Nucleic Acids Res. 2014; 42(Database issue):D749–D755. [PubMed: 24316576]

39. Shanmugam AK, Yocum AK, Nesvizhskii AI. Utility of RNA-seq and GPMDB protein observation frequency for improving the sensitivity of protein identification by tandem MS. J Proteome Res. 2014; 13(9):4113–9. [PubMed: 25026199]

40. Farrah T, Deutsch EW, Omenn GS, Sun Z, Watts JD, Yamamoto T, Shteynberg D, Harris MM, Moritz RL. State of the human proteome in 2013 as viewed through PeptideAtlas: comparing the kidney, urine, and plasma proteomes for the biology- and disease-driven Human Proteome Project. J Proteome Res. 2014; 13(1):60–75. [PubMed: 24261998]

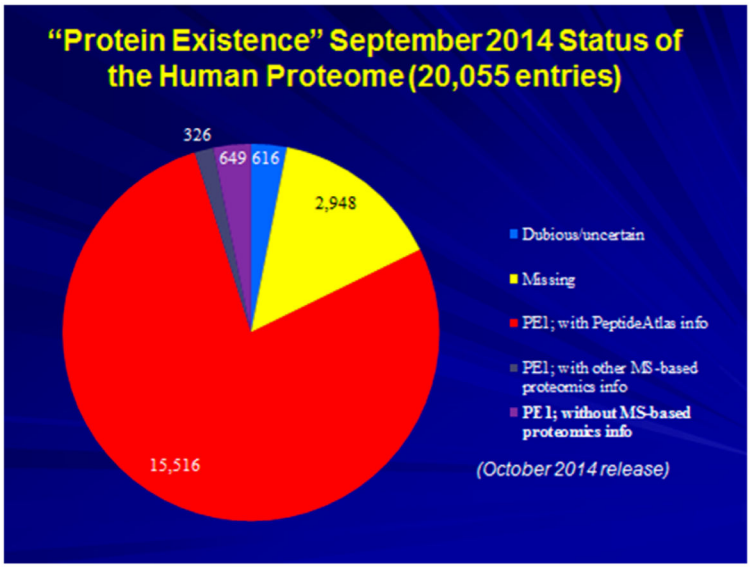


Figure 1.

Baseline classification of protein existence evidence in neXtProt as of version 2014-09-19, with the red (PE1 entries with PeptideAtlas MS data), black (PE1 entries with other MS-proteomics data), and purple (PE1 entries with non-MS protein data together comprising the PE1 total of 16 491 (see Tables 1 and 2). The yellow wedge comprises PE levels 2 + 3 + 4, the "missing proteins", and the blue wedge represents the PE5 "dubious proteins".

Table 1

Numbers of Highly Confident Protein Identifications in Each of the Major Data Resources as of December 2012,[3] September 2013,[4] and October 2014[a]

	Chr	neXtProt Protein Entries	neXtProt PE1 Proteins	Human PeptideAtlas (1% FDR)	GPMDB EC4 Proteins (green)	Human Protein Atlas Evidence (high/medium; now supportive)
Dec 2012	All	20 059	13 664	12 509	14 300	10 794
Sept 2013	All	20 123	15 646	13 377	14 869	10 976
Oct 2014	All	20 055	16 491	14 928	15 459	12 007

[a]Metrics specified in the Call for Papers for this special issue: neXtProt 2014-09-19 protein evidence level PE1 and PeptideAtlas 2014-08 canonical proteins (1% FDR protein level). Additional data from GPMDB 2014-07-01 and from Human Protein Atlas 2014-11-07. The intersection of 14 730 canonical proteins in PeptideAtlas (version 2014-08) with PE1 level evidence in neXtProt (version 2015-04-28) can be observed or downloaded at https://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/GetNextProtChromMapping?atlas_build_id=433&protein_evidence_level_constraint=protein&peptideatlas_constraint=01&redundancy_constraint=on&apply_action=QUERY (see Supporting Information Table S1). The highest evidence class EC4/color code green in GPMDB can be accessed at http://wiki.thegpm.org/wiki/GPMDB_evidence_codes. GPMDB currently utilizes ENSEMBL, v76, while PeptideAtlas uses SwissProt as its reference. This top category is based on the set of common peptides that contains at least one peptide with a higher than minimal scoring distribution in the test for nonrandomness, corresponding to ≤5 observations with skew and excess kurtosis for the peptide both ≤1.5 or the weighted mean ≤ −5.5. GPMDB also differs from PeptideAtlas in not having a rule of parsimony for multiple protein matches. An extensive comparison of GPMDB and PeptideAtlas was reported by Farrah et al.[40]

Table 2

neXtProt Computes a Protein Existence Status Based on Experimental Information from Multiple Types of Studies[a]

PE Level	September 2013	October 2014	
1: Evidence at protein level*	15,646 (77.7%)	16,491 (82%)	
2: Evidence at transcript level	3,570	2,647	} Missing
3: Inferred from homology	187	214	
4: Predicted	87	87	
5: Uncertain or dubious	638	616	

[a]Clear experimental evidence for the existence of the protein. The criteria include partial or complete Edman sequencing, clear identification by mass spectrometry, X-ray or NMR structure, good quality protein–protein interaction, or detection of the protein by antibodies (see text).

Table 3

Chromosome-by-Chromosome Status of the Human Proteome in by Protein Evidence Level in neXtProt 2014-09-19 and Companion Data on Tissue Expression and Subcellular Localization from Human Protein Atlas 2014-11-07

		neXtProt					Human ProteinAtlas		
Chr	**neXtProt Totals per Chr**	**PE1**	**PE2**	**PE3**	**PE4**	**PE5**	**Protein evidence**	**Transcript evidence only**	**No evidence**
1	2056	1683	284	31	7	51	1229	709	138
2	1232	1076	127	4	6	19	760	471	50
3	1071	910	132	10	2	17	685	346	47
4	759	647	70	19	1	22	442	271	54
5	868	740	105	7	5	11	545	317	31
6	1110	949	115	10	4	32	644	357	51
7	936	743	131	8	5	49	567	300	50
8	701	577	68	13	4	39	405	252	44
9	810	636	124	9	6	35	444	299	62
10	760	627	108	4	4	17	495	248	28
11	1319	962	291	17	7	42	713	389	215
12	1031	886	116	4	3	22	670	328	72
13	326	278	31	2	6	9	205	106	18
14	624	516	81	7	3	17	397	202	53
15	602	434	67	12	1	38	363	230	24
16	836	701	91	16	1	27	538	307	30
17	1163	1011	117	7	5	23	710	420	78
18	276	237	26	2	1	10	186	91	11
19	1426	1125	248	10	7	36	800	604	77
20	550	457	78	0	2	13	322	201	37
21	252	180	42	4	0	26	129	64	49
22	487	378	57	3	1	21	279	162	9
X	1420	655	125	9	6	31	444	349	37
Y	47	25	11	4	0	7	22	22	10
MT	14	13	0	0	0	1	4	9	0
Unmapped	9	3	3	2	0	1	9	0	20

Chr	neXtProt Totals per Chr	neXtProt					Human ProteinAtlas		
		PE1	PE2	PE3	PE4	PE5	Protein evidence	Transcript evidence only	No evidence
All	20 064	16 499	2648	214	87	616	12 007	7054	1295