Published in final edited form as:
J Proteome Res. 2014 January 3; 13(1): 15–20. doi:10.1021/pr401144x.

Metrics for the Human Proteome Project---2013-2014 and Strategies for Finding Missing Proteins

Lydie Lane[1], **Amos Bairoch**[1], **Ronald C. Beavis**[2], **Eric W. Deutsch**[3], **Pascale Gaudet**[1], **Emma Lundberg**[4], and **Gilbert S. Omenn**[3,5]

[1]SIB-Swiss Institute of Bioinformatics, Geneva, Switzerland [2]University of British Columbia, Biomedical Research Centre, Vancouver, Briitsh Columbia, Canada [3]Institute for Systems Biology, Seattle, WA, 98101, USA [4]KTHRoyal Institute of Technology, Department of Proteomics, School of Biotechnology, Stockholm, Sweden [5]University of Michigan, Ann Arbor, MI, 48109-2218, USA

Abstract

One year ago the Human Proteome Project (HPP) leadership designated the baseline metrics for the Human Proteome Project to be based upon neXtProt with a total of 13 664 proteins validated at protein evidence level 1 (PE1) by mass spectrometry, antibody-capture, Edman sequencing, or 3D structures. Corresponding chromosome-specific data were provided from PeptideAtlas, GPMdb, and Human Protein Atlas. This year the neXtProt total is 15 646 and the other resources, which are inputs to neXtProt, have high quality identifications and additional annotations for 14 012 in PeptideAtlas, 14 869 in GPMdb, and 10 976 in HPA. We propose to remove 638 genes from the denominator that are “uncertain” or “dubious” in Ensembl, UniProt/SwissProt, and neXtProt. That leaves 3844 “missing proteins”, currently having no or inadequate documentation, to be found from a new denominator of 19 490 protein-coding genes. We present those tabulations and weblinks and discuss current strategies to find the missing proteins.

Keywords

Human Proteome Project; neXtProt; PeptideAtlas; GPMdb; Human Protein Atlas; metrics; missing proteins

Introduction

The overall goals for the Human Proteome Project (HPP) are: (1) to complete in stepwise fashion the Protein Parts List—identifying and characterizing at least one protein product and as many PTM, SAP, and splice variant isoforms as possible from each of the full complement of human protein-coding genes; and (2) to make proteomics a more useful counterpart to genomics by enhancing the work of the entire biomedical research community with high-throughput robust instruments, reagents, specimens, pre-analytical protocols, and knowledge bases for identification, quantification, and characterization of proteins in network context in a broad array of biological systems[1,2]. The HPP comprises

Corresponding Author Gilbert S. Omenn, 2065 Palmer Commons, 100 Washtenaw Avenue, Ann Arbor, MI, 48109-2218, USA, Ph: 1-734-763-7583, Fx: 1-734-615-6553. Corresponding Author: Gilbert S. Omenn, University of Michigan, Ann Arbor, MI, 48109-2218, USA; gomenn@umich.edu; fax 734-615-6553.

Author Contributions The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

about 50 teams organized in the Chromosome-centric C-HPP, the Biology and Disease-driven B/D-HPP, and the Antibody, Mass Spectrometry, and Knowledgebase resource pillars. Our Grand Challenge is to use proteomics to bridge major gaps between evidence of genomic, epigenomic, and transcriptomic variation and diverse phenotypes[3].

The purpose of this article is to ensure common ground for all the C-HPP and B/D-HPP teams for the assessment of progress on the protein parts list, updated approximately annually, for our search for “missing proteins”, and for extensive characterization of proteins in networks and pathways. Understanding the extensive information available in the key data resources is valuable to many other researchers interested in knowing what proteins and what protein isoforms have been identified and characterized in various cell types, organs, and biofluids.

The September 2013 Update of the HPP Metrics for this Special Issue

For the initial JPR C-HPP special issue in January 2013, the HPP executive committee and investigators agreed on five standard baseline metrics for the whole proteome and for each chromosome, as of October 2012, and the respective thresholds for credible evidence[1]. These resources, metrics, and thresholds were Ensembl v69 for numbers of protein-coding genes; PeptideAtlas (canonical/1% FDR) and GPMdb (green) for standardized analyses of mass spectrometry datasets using TransProteomicPipeline and X!Tandem methods, respectively; Human Protein Atlas (high/medium score) for antibody-based protein identifications and expression profiles; and neXtProt (validated at “gold” protein level, corresponding to 1% FDR) for combined mass spectrometry, immunohistochemical, structural, and/or Edman sequence evidence[5]. Each resource has provided a chromosome-by-chromosome analysis as part of their engagement with the Human Proteome Project and C-HPP. The numbers across those five resources last year were 20 059 for Ensembl v69, 12 509 for PeptideAtlas, 14 300 for GPMdb, 10 794 for Human Protein Atlas, and 13 664 for neXtProt.

Here we update those metrics, chromosome-by-chromosome, to the time of the Yokohama HUPO Congress in September 2013. These metrics were useful for HPP discussions and workshops in Yokohama and for the many manuscripts being prepared for this January 2014 second C-HPP special issue of J Proteome Research. As shown in summary rows at the bottom of Table 1, there has been a substantial increase in the numbers of proteins identified: with a denominator of 20,115 neXtProt entries for presumed protein-coding genes, there are 15 646 entries validated at the protein expression level PE1 in neXtProt (78%). The corresponding figures are 14 012 in PeptideAtlas, 14 869 in GPMdb, and 10 976 in HPA. The HPA number reflects a new combination of high and moderate antibody-based protein identifications, now called “supportive” evidence, released as HPA version 12 on 5 December 2013 at www.proteinatlas.org [Stadler et al, this issue]. Last year we used a very rough estimation of “missing proteins” which was the mean of neXtProt, PA, and GPMdb subtracted from the Ensembl number of genes, or 6568 (33%). That approach has been replaced by our Pie Chart analysis (see below).

Both neXtProt and PeptideAtlas had notable increases in numbers in 2013, with 1982 and 1503 additional high-confidence entries, respectively. In the 2013 JPR special issue, Farrah et al reported that the Human Proteome PeptideAtlas lacked major datasets for liver, muscle, and kidney and membrane fractions, which were enriched in the “unseen proteins” category [4]; PeptideAtlas has been expanded in part by many new datasets and in part by the major project comparing the Kidney, Urine, and Plasma Proteomes [Farrah et al, this issue].

The neXtProt and UniProt/SwissProt Scheme of Levels of Evidence

The “protein evidence levels” used by neXtProt (and UniProt/SwissProt) are classified in five categories (PE 1-5). PE1 signifies credible evidence of protein expression and identification by mass spectrometry, immunohistochemistry, 3D structure, and/or amino acid sequencing. PE2 recognizes transcript expression evidence, without evidence of protein expression. PE3 signifies lack of protein or transcript evidence, but presence of protein evidence for a homologous protein in a related species. PE4 hypothesizes from gene models. And PE5 consists of “dubious” or “uncertain” genes that seemed to have some protein-level evidence in the past, but since has been deemed doubtful. We have clear thresholds for PE1 protein evidence, as stated in the Introduction and Table 1 (neXtProt gold, PeptideAtlas 1% FDR, GPMdb green, and HPA supportive). Below we propose a threshold for PE2 transcript evidence. We recommend and implement a fresh approach to PE5, for which evidence is poor to absent, namely, to remove category PE5 from the denominator of protein-coding genes and expected proteins (below). As of September 2013, neXtProt has 15 646 at PE1, 3570 at PE2, 187 at PE3, 87 at PE4, and 638 at PE5.

All of our database resources have chromosome-specific displays to assist the C-HPP investigators; see www.neXtProt.org, www.peptideatlas.org, www.gpm.org, and www.proteinatlas.org, as well as www.thehpp.org, www.c-hpp.org and the C-HPP wiki at http://c-hpp.webhosting.rug.nl/tiki-index.php.

Pie Chart Framing the Quest for Missing Proteins

In Yokohama one of us (AB) presented the pie chart shown in Figure 1 guiding our quest for the missing proteins. This chart also reveals some of the interactions and gaps between pairs of our core databases. Two-thirds of the pie (13 359) are proteins validated at the protein level in neXtProt (and UniProt/SwissProt) and present in PeptideAtlas; 1071 are validated in neXtProt and by additional proteomic (mass spectrometric) data, but not in PeptideAtlas; 1216 are validated in neXtProt from antibody-capture, 3D structures in Protein Data Bank, or Edman sequencing, but have no high-quality mass spec evidence reviewed by neXtProt; these include polypeptide hormones and cytokines and other proteins hard to solubilize or digest with trypsin or otherwise hard to detect by MS. That leaves two sectors. The first is called PE5 (above); these listed genes actually have very little credible evidence and are commonly described as “uncertain” or “dubious”. After considerable discussion and a caution that a manual curation of each would be desirable, we decided to remove these 638 PE5 “genes” from the denominator of protein-coding genes. We also reviewed the discrepancies between protein entries in UniProtKB/SwissProt and neXtProt and gene entries in Ensembl, as discussed below. The final sector of proteins targets those awaiting experimental validation at the protein level, i.e. PE 2 + 3 + 4 = 3844.

Table 2 summarizes the results from all the databases, by chromosome. It also reveals a few anomalies in the matching of proteins and genes. There are 25 nextProt entries not confidently mappable to any chromosome, including 7 PE1, 17 PE2+3+4, and 1 PE5 entries; in addition, there are 7 entries which are shared by more than 1 chromosome, including NX_Q6VEP3 (FAM138A, B, C, F) coded on chromosomes 1, 2, 9, and 19, and NX_Q6iEY1 (OReF16, F29, F3) on chromosomes 1 and 5. There are many other complications. Identical histone H4 proteins are produced from 14 genes and ubiquitin protein by 4 genes. There are bicistronic mRNAs; there are somatic rearrangements that generate immunoglobulins; there are proteins with two or more polypeptide chains from two or more genes (and sometimes chromosomes), representing “many to one” relationships; and there are a great many complexities arising from alternative splicing and gene fusions. Also, we should remember that “the human genome” is a “consensus” sequence; individuals

vary in the number of copies of various genes, and there is evidence for active proteins from certain pseudogenes in certain individuals. In sum, Ensembl is not completely in sync with neXtProt annotation of protein-coding genes.

With extensive studies of presumed “non-coding RNAs”, new small proteins are being found hidden in the 5′ or 3′ regions or made from the opposite strand of DNA; some of these are deduced from ENCODE evidence of transcription from regions of the genome where no gene had been detected. Most of these transcripts and a few resulting polypeptides lack evidence of conservation and fit no functional class, making their recognition suspicious. Bairoch’s conclusion is appropriate: “extraordinary claims require extraordinary evidence”.

There is a vast amount of annotation in neXtProt and the combined databases used by C-HPP and HPP. neXtProt as of September 2013 had 20 128 protein entries (a number which keeps declining) and 39 325 isoforms (a number which keeps rising). As noted at the bottom of Table 2, the number of protein entries is slightly larger (by 15) than the sum of entries assigned to individual chromosomes.

neXtProt contains 45 000 PTMs (glycosylation, phosphorylation, ubiquitination, sumoylation, nitrosylation, and acetylation); it has subcellular localization from the DKFZ GFP-cDNA datasets and the Weizmann Institute of Science Kahn Dynamic Proteomics Database[5]. The next major dataset incorporation will be from Version 12 of Human Protein Atlas. Sequence variants number approximately 850 000, about 40 per protein, drawing upon dbSNP and COSMIC. As demonstrated by Gaudet et al[5], neXtProt offers multiple views and means of data export. GPMdb also has extensive data on proteins with PTMs.

neXtProt does not incorporate findings from GPMdb, so searching GPMdb, as some of the CHPP teams have shown, can reveal some proteins not present in PeptideAtlas and not present in neXtProt. Comparisons of protein lists are made awkward by the fact that different analytical pipelines and different choices of “representative proteins” for “protein groups”, especially highly homologous protein families or subfamilies, lead to matching for different proteins from the same peptides. GPMdb is built on Ensembl, while PeptideAtlas is mapped on SwissProt. Thus, we have resisted the standard practice of creating a Venn diagram with the protein lists from each database. GPMdb has a far larger number of peptides and datasets than does PeptideAtlas. As part of the comparative analysis of Kidney, Urine, and Plasma proteomes, Farrah et al (this issue) compared PeptideAtlas with a computed Swiss-Prot complete mapping for the green-level proteins in GPMdb. Of 12 934 non-redundant Swiss-Prot entries in the 2013 HumanAllPA, 95% are present in GPMdb; conversely 88% of the 14 841 identifiers in the computed GPMdb mapping are present in HumanAllPA. Combined, the two resources covered 15 912 Swiss-Prot entries, 79% of the predicted human proteome in SwissProt.

We recognize that there are additional data resources and data browsers being created around the world, some of which are highlighted in the C-HPP papers of 2013, including Proteome Browser[6], CAPER[7] and CAPER2.0 [Wang D, et al, this issue], GenomewidePDB[8], and Gene-centric Knowledgebase[9].

Strategies for Searching for “Missing Proteins”

There are many explanations for missing proteins. Here we give an outline of approaches (Table 3).

First, we systematically miss proteins expressed significantly only in unusual organs or cell types. Examples might be various regions of the brain, with extreme histologic and functional heterogeneity; nasal epithelium and olfactory cortex, with the very large family of

(generally untranscribed) olfactory receptor genes; testis; and placenta. However, Martins-de-Souza et al[10] reported only five proteins not identified previously by MS in an impressive study of the anterior temporal lobe and corpus callosum in human brains (mapped to chromosomes 4, 6, 11, 14, and 15). Multiple cultured cells from different cell lineages have shown relatively small numbers of cell-line specific proteins, with about 10 000 proteins in common and 20-100 proteins specific per cell line[11,12,13,14]. Fagerberg et al[15] performed antibody-based protein profiling on 11 cell lines and found 27-98 present in only one cell line and not in the others, compared with 50% of 13 985 proteins found in all 11 cell lines. Placenta was reported in the 2013 JPR C-HPP special issue to have 33 previously unreported proteins, of which only 13 were uniquely expressed in placenta[16]. Such studies show synergies between the C-HPP and B/D-HPP consortia.

A strong clue to the potential expression of proteins in particular tissues is the presence of substantial levels of the corresponding mRNA. Zhong et al[17] and Wang Q et al [this issue] have emphasized the usefulness of measuring poly-A-rich-mRNA in nascent chain complexes on polysomes, which they call the “translatome”, reflecting the first and rate-limiting step for translation to proteins. They report 3-7% fewer transcripts in the translatome than in the transcriptome. The Human Protein Atlas has combined analysis of transcripts and detection of protein expression through immunohistochemistry with polyclonal antibodies raised against predicted epitopes of proteins. In the panel of cell lines and tissues used in the Human Protein Atlas transcripts for a total of 18 845 protein-coding genes were detected. Danielsson et al[18] emphasized that there are at least 411 cell types in the human body, including 145 different neuronal cell types. They also note that it is important to recognize a threshold for transcript evidence (typically RPKM >1 or >3, as used by the chromosome 17 team[19](Menon et al, this issue), so as not to be misled by noisy results for very low abundance transcripts. As they stated, low abundance proteins have much higher risk of antibody cross-reactivity/off-target binding at low affinity generating false-positives. Work presented in Yokohama by Uhlen showed a surprisingly high correlation between RNA and protein expression (Spearman correlation 0.96) with 46% of proteins omnipresent, “housekeeping” proteins; 28% mixed tissue expression; 16% tissue or group specific; and 9% of genes lacking both transcripts and proteins. The tissue-specific or tissue-enriched set included 928 in testis, 340 in brain, 194 in adipose tissue, 104 in heart, and 68 in kidney.

Searching for mass spec evidence of such proteins can employ Multiple Reaction Monitoring MS methods, as demonstrated by Chen et al [this issue], who reported confirming 57 targeted proteins in normal human liver, of which 7 had no previous MS-based evidence. Proteogenomic approaches may be complementary, mapping individual peptides to their “genomic coding locations”; for example, an estimated 25% of tryptic peptides span exon boundaries and generally would be false-negatives[20]. Islam et al [this issue] use Protannotator to suggest protein candidates based on non-human homologous proteins.

Second, there may be whole sets of proteins expressed only in early developmental stages in the embryo or fetus. Heck and collaborators[21] reported deep proteomic analyses of embryonic and induced pluripotent stem cell lines, but found only relatively few unknown proteins. It is possible that a burst of additional proteins might occur at critical points in lineage-specific growth and development. Again, transcript expression and antibody-based detection might be guides for mass spectrometry methods.

Third, it is quite likely that some families of genes and proteins are silent but can be activated under certain stresses. The Chinese Chromosome Proteome Consortium investigating chromosomes 1, 8, and 20 has studied the beta-defensin family of 38 genes;

they encode antimicrobial polypeptides of considerable sequence diversity [Liu Y, et al, this issue; Wang Q, et al, this issue]. Thirty of these genes are clustered in chromosomal regions 6p12.3, 8p23.1 (two blocks), 20p13, and 20q11.21. DEFB136 was detected by immunohistochemistry in 25/80 hepatic samples. Essentially all the others were silent. DEFB1 and DEFBA104 were detected at mRNA and RNC-mRNA levels, but no protein was detected. DEFB4A and DEFB4B have been reported to be induced by infectious and inflammatory stimuli.

Olfactory receptor genes are similarly silenced in humans. Olfaction is far more important in rodents and other mammalian species, yet humans have hundreds of olfactory receptor genes, of which a substantial percentage have degraded to pseudogene status.

Fourth, we recognize the importance of limit of detection with present methods, especially for highly complex tissue and biofluid specimens. All technologies and fractionation or depletion protocols that increase the sensitivity of detection will increase the prospects of identifying proteins at low concentrations. MRM or SISCAPA-MRM can be used to target specific proteins and extend the limit of detection. For many membrane-embedded proteins, solubilization is difficult and the protein sequence may not expose many or any tryptic cleavage sites, making detection of tryptic peptides unlikely or impossible. Examples of solute carrier channel proteins were highlighted in the comparative analysis of kidney, urine, and plasma proteomes[4]. There are many proteins with unusual conformations and unusual sequences, as well as peptides that may not be detected by standard methods. Farrah et al[4] extensively annotated the “PA-unseen proteins”, which were enriched for very hydrophobic, very basic, and membrane proteins. They described the striking case of the second most abundant transcript, 60S ribosomal protein L41, which has so many lysine and arginine residues that there are no tryptic peptides longer than two amino acids. Nevertheless, Muraoka et al[22] recovered and identified 3282 membrane-embedded proteins and mapped them to the 24 chromosomes as a service to all C-HPP teams.

Finally, there are large families of proteins or subfamilies with high sequence homology which may give overlapping matches to available tryptic peptides and thereby be excluded when one of the group is chosen as a “representative protein” in the analysis. PeptideAtlas excludes additional protein matches if there is >80% sequence homology. Pfam, the leading protein family database (http://pfam.sanger.ac.uk/), defines protein family by sequence-based hidden Markov model search using HMMER3 alignment[23]. Pfam release 26.0 has 13,000 protein families. Detailed annotations on those proteins which have structures experimentally solved are available at SCOP, the Structural Classification of Proteins (http://scop.mrc-lmb.cam.ac.uk/scop/), with a hierarchy of species, protein, family, superfamily, fold, and class[24].

There is one more way in which we can reduce the number of missing proteins to be found. Earlier we discussed the 638 dubious genes or pseudogenes which we have decided should be removed from the denominator of protein-coding genes. That reduces the number of missing proteins to the proteins with evidence levels 2, 3, or 4, a total of 3844 (Table 2). The new denominator is 20 128 (column 3 of Table 2) – 638 (PE5, column 8 of Table 2) = 19 490.

There is also a path that requires special caution, which is dipping into the less confidently identified proteins in silver or bronze neXtProt, yellow or red GPMdb, 5% FDR in PeptideAtlas, or low-reliability immunostaining in HPA. Such reports may be precursors to definitive findings with more sensitive or more accurate methods, as has been demonstrated with “one-hit wonders” when a more advanced mass spectrometer was utilized[25]. For the

HPP, however, we want to avoid any confusion about the HPP thresholds for “counting” proteins.

Dataset Submissions

The HPP and C-HPP leadership published guidelines and secured agreement from the participating investigators to upload their raw spectra, full datasets, and metadata via ProteomeXchange to EBI/PRIDE for MS/MS data and ISB/SRM Atlas/PASSEL [Vizcaino, JA, et al., Nature Biotechnol, in press]. Authors were advised to place the PXD identifier at the end of the Abstract and the relevant details in Methods; many have done so. It is very important to have full access to the datasets from the HPP, of course.

Concluding Remark

We are greatly encouraged by the progress during this first full year of operation of the CHPP/HPP presented in the articles of the JPR 2014 C-HPP special issue and the many oral reports at the Yokohama Congress and HPP follow-on meeting in Kyoto.

Acknowledgments

Funding Sources GSO acknowledges support from NIH grants RM-08-029, P30U54ES017885, and UL1RR24986.

REFERENCES

1. 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]
2. Aebersold R, Bader GD, Edwards AM, van Eyk JE, Kussmann M, Qin J, Omenn GS. The biology/disease-driven human proteome project (B/D-HPP): enabling protein research for the life sciences community. J Proteome Res. 2013; 12(1):23–7. [PubMed: 23259511]
3. Hood LE, Omenn GS, Moritz RL, Aebersold R, Yamamoto KR, Amos M, Hunter-Cevera J, Locascio L. New and improved proteomics technologies for understanding complex biological systems: addressing a grand challenge in the life sciences. Proteomics. 2012; 12(18):2773–83. [PubMed: 22807061]
4. Farrah T, Deutsch EW, Hoopmann MR, Hallows JL, Sun Z, Huang CY, Moritz RL. The State of the Human Proteome in 2012 as viewed through PeptideAtlas. J Proteome Res. 2013; 12(1):162–71. [PubMed: 23215161]
5. 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]
6. Goode RJ, Yu S, Kannan A, Christiansen JH, Beitz A, Hancock WS, Nice E, Smith AI. The proteome browser web portal. J Proteome Res. 2013; 12(1):172–8. [PubMed: 23215242]
7. Guo F, Wang D, Liu Z, Lu L, Zhang W, Sun H, Zhang H, Ma J, Wu S, Li N, Jiang Y, Zhu W, Qin J, Xu P, Li D, He F. CAPER: a chromosome-assembled human proteome browsER. J Proteome Res. 2013; 12(1):179–86. [PubMed: 23256906]
8. Jeong SK, Lee HJ, Na K, Cho JY, Lee MJ, Kwon JY, Kim H, Park YM, Yoo JS, Hancock WS, Paik YK. GenomewidePDB, a proteomic database exploring the comprehensive protein parts list and transcriptome landscape in human chromosomes. J Proteome Res. 2013; 12(1):106–11. [PubMed: 23252913]
9. Zgoda VG, Kopylov AT, Tikhonova OV, Moisa AA, Pyndyk NV, Farafonova TE, Novikova SE, Lisitsa AV, Ponomarenko EA, Poverennaya EV, Radko SP, Khmeleva SA, Kurbatov LK, Filimonov AD, Bogolyubova NA, Ilgisonis EV, Chernobrovkin AL, Ivanov AS, Medvedev AE, Mezentsev YV, Moshkovskii SA, Naryzhny SN, Ilina EN, Kostrjukova ES, Alexeev DG, Tyakht AV, Govorun VM, Archakov AI. Chromosome 18 transcriptome profiling and targeted proteome

mapping in depleted plasma, liver tissue and HepG2 cells. J Proteome Res. 2013; 12(1):123–34. [PubMed: 23256950]

10. Martins-de-Souza D, Carvalho P, Schmitt A, Junqueira M, Nogueira F, Turck C, Domont G. Deciphering the human brain proteome as part of the chromosome 15-centric human proteome project. J Proteome Res. 2013
11. Beck M, Schmidt A, Malmstroem J, Claassen M, Ori A, Szymborska A, Herzog F, Rinner O, Ellenberg J, Aebersold R. The quantitative proteome of a human cell line. Mol Syst Biol. 2011; 7:549. [PubMed: 22068332]
12. Nagaraj N, Wisniewski JR, Geiger T, Cox J, Kircher M, Kelso J, Paabo S, Mann M. Deep proteome and transcriptome mapping of a human cancer cell line. Mol Syst Biol. 2011; 7:548. [PubMed: 22068331]
13. Geiger T, Wehner A, Schaab C, Cox J, Mann M. Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. Mol Cell Proteomics. 2012; 11(3) M111 014050.
14. Neuhauser N, Nagaraj N, McHardy P, Zanivan S, Scheltema R, Cox J, Mann M. High performance computational analysis of large-scale proteome data sets to assess incremental contribution to coverage of the human genome. J Proteome Res. 2013; 12(6):2858–68. [PubMed: 23611042]
15. Fagerberg L, Oksvold P, Skogs M, Algenas C, Lundberg E, Ponten F, Sivertsson A, Odeberg J, Klevebring D, Kampf C, Asplund A, Sjostedt E, Al-Khalili Szigyarto C, Edqvist PH, Olsson I, Rydberg U, Hudson P, Ottosson Takanen J, Berling H, Bjorling L, Tegel H, Rockberg J, Nilsson P, Navani S, Jirstrom K, Mulder J, Schwenk JM, Zwahlen M, Hober S, Forsberg M, von Feilitzen K, Uhlen M. Contribution of antibody-based protein profiling to the Human Chromosome-centric Proteome Project (C-HPP). J Proteome Res. 2013; 12(6):2439–48. [PubMed: 23276153]
16. Lee HJ, Jeong SK, Na K, Lee MJ, Lee SH, Lim JS, Cha HJ, Cho JY, Kwon JY, Kim H, Song SY, Yoo JS, Park YM, Hancock WS, Paik YK. Comprehensive genome-wide proteomic analysis of human placental tissue for the chromosome-centric human proteome project. J Proteome Res. 2013; 12(6):2458–66. [PubMed: 23362793]
17. Zhong J, Cui Y, Guo J, Chen Z, Yang L, He Q, Wang T. Resolving chromosome centric human proteome with translating mRNA analysis: a strategic demonstration. J Proteome Res. Nov 15.2013 Epub.
18. Danielsson F, Wiking M, Mahdessian D, Skogs M, Ait Blal H, Hjelmare M, Stadler C, Uhlen M, Lundberg E. RNA deep sequencing as a tool for selection of cell lines for systematic subcellular localization of all human proteins. J Proteome Res. 2013; 12(1):299–307. [PubMed: 23227862]
19. Liu S, Im H, Bairoch A, Cristofanilli M, Chen R, Deutsch EW, Dalton S, Fenyo D, Fanayan S, Gates C, Gaudet P, Hincapie M, Hanash S, Kim H, Jeong SK, Lundberg E, Mias G, Menon R, Mu Z, Nice E, Paik YK, Uhlen M, Wells L, Wu SL, Yan F, Zhang F, Zhang Y, Snyder M, Omenn GS, Beavis RC, Hancock WS. A chromosome centric human proteome project (C-HPP) to characterize the sets of proteins encoded in chromosome 17. J Proteome Res. 2013; 12(1):45–57. [PubMed: 23259914]
20. Risk BA, Spitzer WJ, Giddings MC. Peppy: proteogenomic search software. J Proteome Res. 2013; 12(6):3019–25. [PubMed: 23614390]
21. Munoz J, Low TY, Kok YJ, Chin A, Frese CK, Ding V, Choo A, Heck AJ. The quantitative proteomes of human-induced pluripotent stem cells and embryonic stem cells. Mol Syst Biol. 2011; 7:550. [PubMed: 22108792]
22. Muraoka S, Kume H, Adachi J, Shiromizu T, Watanabe S, Masuda T, Ishihama Y, Tomonaga T. In-depth membrane proteomic study of breast cancer tissues for the generation of a chromosome-based protein list. J Proteome Res. 2013; 12(1):208–13. [PubMed: 23153008]
23. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer EL, Eddy SR, Bateman A, Finn RD. The Pfam protein families database. Nucleic Acids Res. 2012; 40(Database issue):D290–301. [PubMed: 22127870]
24. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res. 2008; 36(Database issue):D419–25. [PubMed: 18000004]

25. Tang HY, Ali-Khan N, Echan LA, Levenkova N, Rux JJ, Speicher DW. A novel four-dimensional strategy combining protein and peptide separation methods enables detection of low-abundance proteins in human plasma and serum proteomes. Proteomics. 2005; 5(13):3329–42. [PubMed: 16052622]

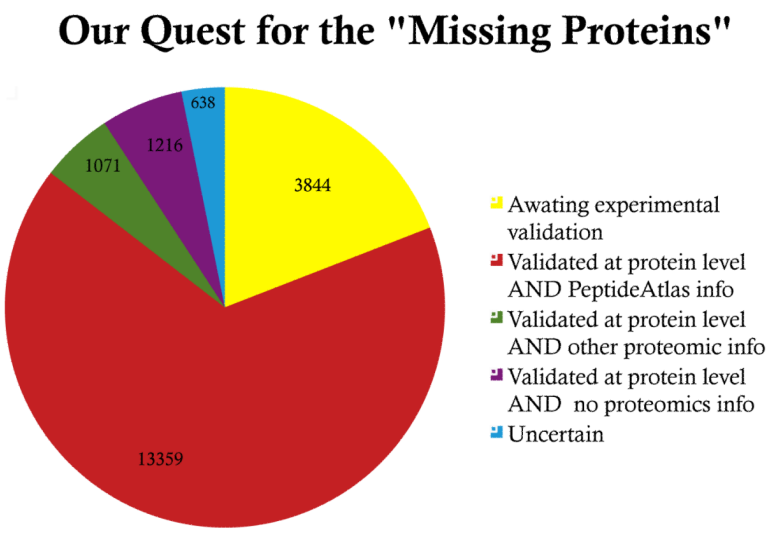

Figure 1.

Quantitative Scheme showing the nature of mass spectrometry and other data underlying neXtProt proteins with Protein Evidence level 1. The final sector or wedge of 3844 proteins constitutes our target of “missing proteins”, while the fourth sector of 638 “dubious/ uncertain genes” represents the set of genes we are removing from the denominator of projected protein-coding genes in the whole human genome, now set at 19 490 in neXtProt.

Table 1

Summary of Baseline (December 2012) and Updated (September 2013) C-HPP Master Table of Metrics
Numbers of highly confident protein identifications in each of the major data resources as of Dec 2012 (Marko-Varga et al, JPR Jan 2013) & Sept 2013

	Chr.	Ensembl Protein-coding Genes	neXtProt (gold)	Human PeptideAtlas (canonical)	GPMdb (green)	Human Protein Atlas Evidence (high or medium)
Dec-12	Total	20,059	13,664	12,509	14,300	10,794
Sep-13	Total	20,123	15,646	14,012	14,869	10,976

Inputs from Pascale Gaudet, Lydie Lane, Amos Bairoch–neXtProt; Terry Farrah, Eric Deutsch–Peptide Atlas; Ron Beavis–GPMdb; Emma Lundberg, Mathias Uhlen–Human Protein Atlas; and all C-HPP teams

Table 2

Chromosome-by-chromosome Updated HPP Metrics Table as Guide to Search for Missing Proteins

Chromosome	Ensembl Protein-coding Genes v 72	neXtProt entries (Sept 2013)	neXtProt PE1 (Sept 2013)	PeptideAtlas (July 2013)	GPMdb green (Aug 2013)	HPA Evidence Supportive (Sept 2013)
1	2059	2061	1600	1415	1521	1119
2	1240	1239	1029	909	962	685
3	1071	1076	877	764	818	634
4	769	763	613	529	566	398
5	862	867	696	624	668	496
6	1094	1108	902	795	784	616
7	938	944	706	660	680	520
8	709	701	553	472	500	383
9	821	821	616	544	567	403
10	762	763	592	528	574	452
11	1314	1321	926	791	845	659
12	1026	1030	842	730	787	609
13	328	328	266	236	251	186
14	625	626	503	445	472	372
15	614	609	457	427	456	328
16	839	831	660	614	671	498
17	1167	1167	951	847	907	648
18	276	278	227	195	219	175
19	1414	1424	1018	1028	1076	716
20	547	552	442	384	398	276
21	254	254	167	142	148	122
22	458	464	356	337	339	254
X	866	827	613	554	626	398
Y	56	47	21	30	22	20
MT	14	14	13	12	12	
Total	20123	20115	15646	14012	14869	10967
Added since Oct 2012 (Marko-Varga 2013)		Total neXtProt entries (20128) include 15 not mapped to chromosome	1982	1503	569	See Stadler et al [this issue]

Table 3

Strategies for Searching for Missing Proteins

1.	Start from the tissue distribution of reported transcriptome expression
2.	Consider early stages of life
3.	Consider special stresses or other perturbations
4.	Recognize low abundance of many proteins or transmembrane helical structures and sequences poor in tryptic cleavage sites; all of these will require more sensitive or different analytical methods.
5.	Seek more detailed information about highly homologous families of proteins; and increase the sequence coverage, if feasible.