Journal of proteome research

Editorial

pubs.acs.org/jpr

Recent Advances in the Chromosome-Centric Human Proteome Project: Missing Proteins in the Spot Light

INTRODUCTION

The HUPO Human Proteome Project (HPP) was launched in September 2011 with dual aims: (1) to identify and characterize at least one protein product from each of the approximately 20 000 predicted protein-coding genes as well as many isoforms of those proteins and (2) to develop reagents, methods, and databases that would facilitate the incorporation of proteomics in biological studies throughout the life sciences community and put proteins and proteoforms in the context of biological networks and pathways in health and disease (Legrain et al.[1]). A leading component of the HPP is the Chromosome-centric HPP (C-HPP), led by Young-Ki Paik of Korea, Bill Hancock of the USA, Gyorgy Marko-Varga of Sweden, and now Lydie Lane of Switzerland, with chromosome-based teams distributed throughout the world (see cover). This September issue of *JPR* (and manuscripts in the October issue to follow) constitutes the third annual C-HPP special issue since 2013, with about 40 articles each year. Complementary components are the Biology/Disease-HPP and the Mass Spectrometry, Antibody, and Knowledgebase resource pillars. The HPP has had a major impact on the field through collaborative research, through the development of the ProteomeXchange in conjunction with PRIDE and PASSEL to capture and make available all data sets (Vizcaino et al.[2]), and through standardized generation of reanalyzed peptide data and protein matches from those experimental data sets, led by Peptide Atlas (Deutsch et al.[3]) and GPMDB (Fenyo and Beavis[4]). The PeptideAtlas results are then incorporated into the neXtProt curation of evidence of protein existence, combining mass spectrometry and multiple other types of protein data (Lane et al.;[5] Deutsch et al.;[3] Omenn et al.[6]). Such reanalysis was particularly informative for the large-scale experimental data sets of Kim et al.[7] and Wilhelm et al.[8]), including reanalysis by Savitski et al.[9]

Over the course of the past 3 years the number of highly confident protein identifications in neXtProt (protein existence level PE1) has grown dramatically from 13 664 in December 2012 to 15 646 in September 2013 to 16 491 in October 2014, which served as the baseline for the studies and new data reported in this special issue. Conversely, that means that there were, as of late 2014, 2948 "missing proteins" from genes classified having protein existence level PE 2, 3, or 4 as well as 616 uncertain or dubious proteins at PE 5 in the UniProt/SwissProt/neXtProt scheme (Omenn et al.[6]).

The C-HPP was planned and funded with a 10-year perspective. During the first half of Phase I (2012-09 to 2018-09) (Paik et al.[10]), the consortium has made meaningful progress in establishing teams for each of the 24 chromosomes plus mitochondria, building working relationships through 12 scientific workshops, adopting standard metrics, and publishing these three special issues of *JPR*, with many interesting findings and methodological advances. This year the emphasis has been on finding credible evidence of additional "missing proteins" (Horvatovich et al.[11]) among the 2948 lacking any or sufficient protein-level evidence as curated by neXtProt 2014-09-19, down from >6000 2 years earlier (Marko-Varga et al.[12]). There has emerged a complementary interest in the reasons why many predicted proteins have not yet been detected or, indeed, may not be detectable by current sample preparation and mass spectrometry methods. We have lately realized how even excellent mass spectra and peptides confirmed with synthetic peptides and multiple reaction monitoring may be challenging to match to the missing proteins due to alternative explanations. These include additional reference genomes, sequence variants, and post-translational modifications that match the m/z ratios of the features of the reference protein or explain novel peptides attributed to translation from a lncRNA sequence (Deutsch et al.;[3] Nesvizhskii[13]). We must be alert to such alternative explanations when a protein never before observed in the same types of specimens is identified, generally with automated search engines. The acceptance of one-hit wonders, long excluded, especially with a single spectrum, reflects overconfidence in the methods and requires additional data and scrutiny of the spectral features and alternative matches.

All manuscripts submitted for review for the special issues are required to make available the experimental data sets via ProteomeXchange. ProteomeXchange has grown from 114 MS/MS data submissions at PRIDE in 2012 to 1089 data sets in July 2015. Meanwhile, the Human Protein Atlas continues its rapid growth using immunohistochemistry and immunofluorescence and RNA-seq on a broad array of tissues (Uhlen et al.[14]).

HIGHLIGHTS

We have organized this brief overview of the 30 articles in the September 2015 C-HPP special issue under four major topics:

(i) *Genome complexity, proteogenomics, integrated analysis of transcriptomics and proteomics and functional studies of splice variant proteins*: Modulation of neuronal differentiation by the Y chromosome male-specific region gene DDX3Y (Vakilian et al.[15]); functional networks of highest-connected splice isoforms in a genome-wide analysis of 6157 multi-isoform genes, illustrated with ABCC3, RBMB34, and ERBB2 (Li et al.[16]); splice variants of Y chromosome-coded lysine-specific demethylase 5D in prostate cancer cells (Jangravi et al.[17]); analysis of pathways and protein−protein interactions for the 66 Y chromosome-coded proteins (Rengaraj et al.[18]); functions of noncanonical splice isoforms specific to HER2+/ER−/PR− breast cancers such as highly expressed DMXL2 isoform 3 (Menon et al.[19]); and integrative analysis of transcriptome and proteome of a B-cell lymphoma, with 82% overlap in proteins identified by the two omics approaches (Diez et al.[20]) are explored. Tay et al.[21] utilized their PGNexus pipeline to integrate genomic and proteomic data from mesenchymal stem cells on

Special Issue: The Chromosome-Centric Human Proteome Project 2015

Published: September 4, 2015

DOI: 10.1021/acs.jproteome.5b00785
J. Proteome Res. 2015, 14, 3409−3414

exons and splice junctions to identify protein isoforms. Querying the TranscriptCoder-derived or Ensembl databases, they identified ~450 protein isoforms and their proteotypic peptides, including candidate hMSC-specific isoforms for DPYSL2 and FXR1 (Tay et al.[21]). Woo et al.[22] from the TCGA Clinical Proteomic Tumor Analysis Consortium (CPTAC) applied proteogenomics to identify multiple mutational or structural variants and gene fusions, including those in immune system genes, through customized mining of RNA-seq data. The Chromosome 12 Consortium from India and South Asia has focused on differentially expressed transcripts and proteins in gliomas and their coexpression, coregulation, and colocalization on chromosome 12. Overexpression of genes maps onto amplicon regions; colocalization suggests common determinants of coexpression and coregulation; close proximity to functionally related genes may help predict their functions; and integration of gene−protein sets with ontologies of medical terms can reveal the disease network (Jayaram et al.[23]).

(ii) *Choice of little-studied biological or clinical specimens to identify and characterize tissue-specific or tissue-enriched low-abundance missing proteins, guided by evidence of transcript expression (PE level 2)*: Eckhard et al.[24] applied the N-terminomics approach with TAILS and shotgun proteomics to identify enriched stable proteolytic cleavage proteoforms in the human dental pulp proteome; this paper provides a striking example of the importance of stringent standards for validation of initial findings (see case study, below). Testis has been predicted from transcript (PE2) data to be by far the richest organ for tissue-specific expression (Uhlen et al.;[14] Djureinovic et al.[25]). Zhang et al.[26] used two different SDS-PAGE separation systems on three individual unpooled testis specimens and found 166 protein groups listed in neXtProt as missing proteins (138 with transcript expression) with 1% protein FDR and stringent review of the spectra. They noted that 364 genes have 50 times higher abundance in testis than in 26 other tissues, according to HPA, including 28 cancer-testis antigens. Ahmadi Rastegar et al.[27] investigated testicular tissue from men with azoospermia to generate isoform-level gene expression profiles of Y chromosome genes and their X chromosome paralogues, with a focus on 14 AZF family genes. A Franco-Swiss team (Jumeau et al.[28]) analyzed the proteome of ejaculated spermatozoa; initial screens identified 89 matches for missing proteins, of which ten were identified as coded on chromosome 2 and two were identified as coded on chromosome 14, including some inferred to be involved in ciliation and flagellar mechanics; two (C2orf57 and TEX37) were validated by immunohistochemistry of testis tissue samples. A separate Franco-Swiss effort on chromosomes 2 and 14 using a diverse array of specimens, not including testis or semen, reported 52 PE2-4 missing proteins and 6 PE5 uncertain protein candidates, of which 13 (8 on chromosome 2, 5 on chromosome 14) were validated with FDR and similarity scores for reference versus endogenous peptides in LC−SRM assays (Carapito et al.[29]). There was no overlap between the two groups' final lists. Another special set of proteins is the family of 34 beta-defensins, inducible antimicrobial proteins, of which only three had detectable expression, with high levels for DEFB 1 (Chr 8) and low levels for DEFB 112 (Chr 6) and DEFB126 (Chr 20) (Fan et al.[30]). DNase I hypersensitivity, transcription factor binding, and histone and CpG modifications in ENCODE data were characterized to explain rare expression of the beta-defensins and other proteins in the same chromosome region. A clever method for activating unexpressed genes and transcripts is epigenetic manipulation of fully differentiated adult human cells in culture through interference with histone acetylation or methylation (Yang et al.[31]); 29 missing or uncertain proteins predicted to be involved in development or spermatogenesis were identified with Mascot and 13 with MaxQuant. This approach might be useful for GPCRs and olfactory receptor genes.

(iii) *Development of improved analytical techniques and related experimental methods to solubilize, capture, and characterize proteins that are difficult to measure, especially membrane proteins*: Horvatovich et al.[11] summarized a wide variety of C-HPP approaches for the quest to identify missing proteins. For example, a consortium of chromosome teams 5, 10, 16, and 19 utilized the in vitro transcription/translation system of LaBaer to develop LC−SRM assays for 18 missing proteins (Horvatovich[32]). High-pH reversed-phase StageTip fractionation and MRM−MS were used to analyze nonsmall cell lung cancers and cell lines (Kitata[33]). Differentially expressed membrane proteins, such as DSG3, CD109, and CD14, from the 11q13 amplicon prominently associated with head, neck, and breast cancers were functionally validated in vitro (Hoover[34]). Su et al.[35] developed systematic enrichment strategies to identify missing proteins that fell into four classes: (1) low-molecular-weight (LMW) proteins, (2) membrane proteins, (3) proteins that contained various post-translational modifications (PTMs), and (4) nucleic-acid-associated proteins. Of the 8845 proteins identified in seven data sets, 79 proteins were classified as missing proteins, of which 30 missing and 6 uncertain proteins were confirmed. A Chinese consortium (Chen[36]) prepared a workflow for enriching detergent-insoluble cytoplasmic proteins (DIPs) from three human lung and three human hepatoma cell lines via differential centrifugation. A total of 23 missing proteins were identified by MS, of which 18 had translation evidence from nascent RNC complexes. Cytoplasmic DIPs were not an enrichment of transmembrane proteins, were chromosome-, cell type-, and tissue-specific, and were biologically and physical−chemically different from the soluble proteome.

(iv) *Upgrading bioinformatics knowledge base and related informatics tools to enable credible identification and validation of new protein observations*: A regular feature of the annual special issue of the C-HPP is the paper on HPP Metrics, with 16 491 PE level 1 neXtProt and 14 928 canonical PeptideAtlas entries as well as an extensive discussion about quality assurance and guidelines for evaluating claims of missing proteins or novel proteins (Omenn et al.[6]). Deutsch et al.[3] present the latest 2015-03 build of PeptideAtlas, showing the increments during 2014 from TCGA, Kim et al.[7] and Wilhelm et al.[8] data sets, and refinements of PeptideAtlas, including the addition of Atlas-Prophet, new categories for peptide-to-protein matches, and a higher bar for protein matches, requiring two proteotypic peptides of nine or more amino acids and search of the reference proteomes for more likely matches from the available peptides.

These two papers provide a thorough examination of the challenges of claiming, confirming, and validating peptide findings and protein matches. We recommend that all investigators scrutinize the discussion sections of these papers and apply the guidelines to their own data sets and other publicly available data sets. Such quality assurance will be subjected to open discussion at the HUPO 2015 Congress in Vancouver. Claims of novel translated products from pseudogenes or long noncoding RNAs require at least as great scrutiny as missing proteins from genes with transcripts or homologies (PE 2,3,4), including use of class-specific FDRs (Omenn et al.;[6] Nesvizhskii[13]). These papers also guide investigators to use of

DOI: 10.1021/acs.jproteome.5b00785
J. Proteome Res. 2015, 14, 3409−3414

the valuable data in neXtProt, PeptideAtlas, GPMDB, and Human Protein Atlas.

GenomewidePDB 2.0 by Jeong[37] describes new features that integrate transcriptomic information (e.g., alternatively spliced transcripts), annotated peptide information, and an advanced search interface, which can find proteins of interest when applying a targeted proteomics strategy. GenomewidePDB2.0, may also enhance the exchange of information among the proteome community. The authors have a thorough discussion of the missing proteins challenge; for example, 92 of the 2948 missing proteins lack any uniquely mappable tryptic peptide, suggesting that a different protease or a top-down analysis may be needed. CAPER3.0 is the latest version of the analytical resource, now a cloud-based platform, for data sets from the Chinese C-HPP chromosomes 1, 8, and 20 Consortium (Yang[38]); data-intensive applications are demonstrated for identifying novel peptides, single amino acid variants (SAVs) from known missense mutations, sample-specific SAVs, and exon-skipping events. PPLine is a Python-based proteogenomic pipeline from the Russian consortium (Krasnov[39]) providing automated discovery of single-amino-acid polymorphisms, indels, and alternatively spliced variants from raw transcriptome and exome sequence data (using both Illumina HiSeq and SOLiD for HepG2 cells and liver tissue), single-nucleotide polymorphism annotation and filtration, and the prediction of proteotypic peptides.

dasHPPboard facilitates the analysis of a variety of proteogenomic data sets from the Spanish Chromosome 16 researchers (Tabas-Madrid[40]). MI-PVT is a new web-based Proteome Visualization Tool for displays by chromosome or by protein family (Panwar[41]). Finally, I-TASSER and COFACTOR algorithms, well-established tools in structural biology for predicting protein folding and protein functions, were applied to the list of 616 PE5 uncertain/dubious predicted proteins to evaluate their prospects for conformational folding and potential functions (Dong et al.[42]).

A CASE STUDY OF CLOSE EXAMINATION OF INITIAL FINDINGS OF MISSING PROTEINS

As noted above, an attractive approach to finding missing proteins is to combine the use of orthogonal methods of peptide and protein identification with a selection of understudied tissues. The experience of one of us (C.M.O.) is illustrative of the difficulties of validating missing protein findings. The Overall Lab employed a N-terminomics approach, Terminal Amino Isotopic Labeling of Substrates (TAILS), to enrich and identify protein N-terminal peptides. Such information is intrinsically valuable for the HPP effort as knowledge of the actual protein termini captures a basic characteristic of every protein. However, in addition to mature protein N terminal peptide identification, previous TAILS analyses had repeatedly made the surprising observation that >50% of proteins in a tissue sample displayed various proteoforms having truncated termini, both N and C-terminal. This number increased to over 60 and 70% in erythrocytes and human dental pulp, as reported by Eckhard et al.[24] Thus, precision proteolysis had generated one or more neo-termini per protein in vivo, therefore generating a variety of proteolytic proteoforms of each protein. Recognizing this observation as a potential new route to increasing proteome coverage, Eckhard et al.[24] reasoned that identification of protein semitryptic peptides may render some regions of a protein amendable for proteomic identification by overcoming unfavorable peptide fragmentation or ionization properties. Thus, semitryptic internal peptides may be more readily identified compared with their spanning tryptic peptides.

Human dental pulp, the tissue residing inside teeth that reacts to dental caries and thermal or pH extremes with pain (toothache), also lays down the organic protein matrix of mineralized dentine. Thus, dentine-specific proteins that direct mineralization ought to be found and may be proteins not found elsewhere. Many such specific proteins had been known from published biochemical studies, such as dentin sialophosphoprotein, but not found proteomically simply because pulp and dentine had been subject to so few proteomics analyses by modern techniques. We identified 4332 proteins at a protein FDR <0.7% with 9079 protein N termini identified in dental pulp, so the prospect for finding a trove of missing proteins seemed high. Indeed, 174 missing proteins were identified by the community-accepted criteria for high confidence protein identification (multiple PSMs identified by multiple search engines at <1% FDR, with further analysis by the ISB PeptideProphet, iProphet, and ProteinProphet), yet when manual inspection of each PSM (1159) was performed and peptides were excluded in the case of isobaric Leu/Ile ambiguities and if fewer than 7 amino acid residues in length, the number dwindled all the way to 17 from the previous 174. Moreover, we now recognize that these still only represent candidate missing proteins, because synthetic peptide SRMs have yet to be analyzed to match with the these peptides. Finding missing proteins is hard! We appreciate the guidance of the reviewers and editors, and we support the guidelines of the C-HPP.

Several other papers in this issue also present details of the application of more stringent guidelines to the confirmation or validation of missing protein claims (Carapito et al.;[29] Jumeau et al.;[28] and Yang et al.[31]).

SETTING OUT THE FUTURE DIRECTION OF THE C-HPP/ADDRESSING KEY PROBLEMS

(1) *What is sufficient evidence of publishing claims of detection of MPs*? From the claims of 904 discovered MPs in the manuscripts submitted for this special issue, we realized that we need more stringent guidelines for our investigators and for the journals. These findings might be considered candidate missing proteins. There is no ambiguity that spectra for claims of detection of missing proteins must be scrutinized for the presence of ion ladder features and the absence of anomalies; that more than one search engine should be used, but the results must agree between the search engines; and that reliance on a single proteotypic peptide, especially with a single spectrum, is highly risky, which led PeptideAtlas to raise the bar to two or more distinct peptides of nine or more amino acids and neXtProt to raise its bar to either two proteotypic peptides of seven or more amino acids or one of nine amino acids. When similar specimens in good studies have never found the newly identified protein match, it behooves us to check for a match to an abundant protein with a single amino acid substitution or an isobaric post-translational modification. See Deutsch et al.[3] for examples. The same applies in spades for peptides that appear to represent translation products from lncRNAs or from pseudogenes. We recommend assay with SRM or SWATH using high-quality synthetic peptides and comparison with spectral libraries and databases (Schubert et al.[43]). However, it is essential to recognize that these steps enhance confidence in the identification of the peptide but cannot validate the match to the protein.

DOI: 10.1021/acs.jproteome.5b00785
J. Proteome Res. 2015, 14, 3409−3414

(2) *How best can we improve methodological approaches, select optimal biological samples, and recommend stringent informatics protocols for the characterization of challenging protein families, such as the olfactory receptors?* Protein families, especially those with highly homologous sequences, generate challenges for peptide-to-protein matches when the available peptides generate indistinguishable matches to two or more proteins. See Deutsch et al.[3] for the classification of protein matches.

Olfactory receptor proteins are a particular dilemma. Many investigators are attracted to their study because they are G-protein coupled receptors (GPCRs) and may have a variety of sensory functions, not limited to smell and taste, especially in nonhuman animals. There are about 900 genes and pseudogenes for olfactory receptors. As documented by Omenn et al.,[6] PeptideAtlas has reevaluated its olfactory receptor entries and has eliminated the two remaining entries as of 2014 (Deutsch et al.[3]). GPMDB has reduced its very long list of entries to six with high-quality entries then recognized that even these may match better to other proteins. Ezkurdia et al.[44] examined the spectra made available by Kim et al.[7] for 108 reported olfactory receptor proteins and by Wilhelm et al.[8] for 200 reported olfactory receptor proteins and concluded not one passed muster for quality spectra or valid protein match. For several years, C-HPP teams have been searching for surgical collaborators to obtain good tissue specimens from olfactory epithelium in the upper nasal passages or olfactory cortex in the brain. Even then, expression of these genes, if active, may be clonal and at low levels because only a few OR transcripts have been reported and confirmed.

(3) *The initial focus of characterizing the missing proteins can now be expanded to the study of protein variants*. Recognizing the huge space of protein proteoforms, the community needs to agree on a viable set of next targets, such as the as alternative splice variants (ASVs) of disease-important proteins.

The HPP and C-HPP have articulated from the beginning that the characterization of protein products from protein-coding genes should include deep analyses for sequence variants, splice variants, and post-translational modifications. Glycan and glycoprotein analysis have been a major feature of the HUPO initiatives from the beginning. PeptideAtlas has now created a Phospho-PeptideAtlas. Many of the articles in this special issue address the detection and functional characterization of splice variants. We are certain that these directions will become even more prominent in the coming years. They represent additional technical and biological complexity.

(4) *Technology and methodology will continue to advance.* We look forward to the development of integrated bioinformatics tools as well as experimental methods and to crossover analyses of mass spectrometry and antibody-profiling. A major outcome from the B/D-HPP has been the identification of proteotypic peptides for essentially all known and predicted protein sequences, the preparation of synthetic peptides, and the generation of algorithms for SRM/MRM−MS transitions and SWATH analyses. The utilization of these resources has been limited to date except in expert hands. Facilitating the use of proteomics platforms and data for integrated omics studies widely in the life sciences and biomedical community remains a major challenge. We shall encourage greater interaction between the C-HPP, B/D-HPP, and Resource Pillars of the HPP.

(5) *As we begin a new phase of the long-term C-HPP program, how can we enhance the capabilities and deliverables of the C-HPP teams?* One approach emerging naturally among the teams and the leadership is the formation of clusters of teams with overlapping interests and the welcoming of additional investigators with fresh ideas and new methods. Because major data sets cover the entire proteome, it is natural to share the data, to collectively identify new kinds of organs or tissues to analyze, and to compare the methods, databases, and browsers developed in isolation. Many teams may be interested in organizing studies of families of proteins often occurring in clusters on several chromosomes, of amplicons, which are quintessential cis-regulated chromosomal features, and alternative splicing, a key evolutionary development in multicellular organisms with multiexonic genes to generate greater protein diversity. It is likely that isoforms will be much more specific as diagnostic biomarkers or molecular therapeutic targets than are the protein and transcript mixtures from individual genes.

To establish cross-chromosome analysis and collaboration networks within the consortium, we should make greater efforts to restructure certain C-HPP teams according to the target disease (e.g., cancers, reproductive, neurodegenerative, or metabolic disease), sharing common resources (e.g., In Vitro Transcription & Translation/IVTT, bioinformatics pillars, biobanks, antibody reagents), method development (e.g., membrane protein, OR, computational programs, bioinformatics tools), biological insight into gene activity for regulation of protein production (e.g., lncRNAs, pseudogenes, nsSNPs), and tissues (e.g., brain, liver, heart). These steps would naturally lead to stronger collaborations with the B/D-HPP groups.

■ CONCLUSIONS

This third C-HPP special issue further demonstrates the important role of HUPO in the evolution of the complex goal of defining the human proteome. The C-HPP initiative has successfully promoted global collaborations in pursuing the full protein parts list based on known protein coding genes, has required the deposition of proteomics data sets and provided standardized reanalyses of those data sets, has integrated transcriptomic and proteomic information, has emphasized the importance of amplicons in protein expression, and has promoted novel analytical approaches and informatics tools. Much has been achieved over the past few years. We look forward to the next phase of better understanding the complexity of human biology enabled when the draft proteomes become the functional proteome map.

Young-Ki Paik*

Yonsei Proteome Research Center, Yonsei University, Seoul 120-749, Korea

Gilbert S. Omenn

Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan 48109, United States

Christopher M. Overall

Department of Biochemistry and Molecular Biology, University of British Columbia, Vancouver, British Columbia V6T 1Z3, Canada

Eric W. Deutsch

Institute for Systems Biology, Seattle, Washington 98109, United States

William S. Hancock*

Department of Chemical Biology, Northeastern University, Boston, Massachusetts 02115, United States

DOI: 10.1021/acs.jproteome.5b00785
J. Proteome Res. 2015, 14, 3409−3414

AUTHOR INFORMATION

Corresponding Authors

*Y.-K.P.: E-mail: paikyk@yonsei.ac.kr. Tel: +82-2-2123-4242. Fax: +82-2-393-6589.

*W.S.H.: E-mail: wi.hancock@neu.edu. Tel: +1-617-869-8458. Fax: +1 617-373-2855.

Notes

Views expressed in this editorial are those of the authors and not necessarily the views of the ACS.

ACKNOWLEDGMENTS

Part of this work was supported by a grant from the Korean Ministry of Health and Welfare for the Global C-HPP Research (to Y.K.P., HI13C2098).

RELATED READINGS

(1) Legrain, P.; Aebersold, R.; Archakov, A.; Bairoch, A.; Bala, K.; Beretta, L.; Bergeron, J.; Borchers, C. H.; Corthals, G. L.; Costello, C. E.; Deutsch, E. W.; Domon, B.; Hancock, W.; He, F.; Hochstrasser, D.; Marko-Varga, G.; Salekdeh, G. H.; Sechi, S.; Snyder, M.; Srivastava, S.; Uhlen, M.; Wu, C. H.; Yamamoto, T.; Paik, Y. K.; Omenn, G. S. The Human Proteome Project: Current State and Future Direction. *Mol. Cell. Proteomics* **2011**, *10* (7), M111 009993.

(2) Vizcaino, J. A.; Deutsch, E. W.; Wang, R.; Csordas, A.; Reisinger, F.; Rios, D.; Dianes, J. A.; Sun, Z.; Farrah, T.; Bandeira, N.; Binz, P. A.; Xenarios, I.; Eisenacher, M.; Mayer, G.; Gatto, L.; Campos, A.; Chalkley, R. J.; Kraus, H. J.; Albar, J. P.; Martinez-Bartolome, S.; Apweiler, R.; Omenn, G. S.; Martens, L.; Jones, A. R.; Hermjakob, H. ProteomeXchange Provides Globally Coordinated Proteomics Data Submission and Dissemination. *Nat. Biotechnol.* **2014**, *32* (3), 223−6.

(3) Deutsch, E. W.; Sun, Z.; Campbell, D.; Kusebauch, U.; Chu, C. S.; Mendoza, L.; Shteynberg, D.; Omenn, G. S.; Moritz, R. L. State of the Human Proteome in 2014/2015 As Viewed through PeptideAtlas: Enhancing Accuracy and Coverage through the AtlasProphet. *J. Proteome Res.* **2015**, DOI: 10.1021/acs.jproteome.5b00500.

(4) Fenyo, D.; Beavis, R. C. The GPMDB REST Interface. *Bioinformatics* **2015**, *31*, 2056.

(5) Lane, L.; Bairoch, A.; Beavis, R. C.; Deutsch, E. W.; Gaudet, P.; Lundberg, E.; Omenn, G. S. Metrics for the Human Proteome Project 2013−2014 and Strategies for Finding Missing Proteins. *J. Proteome Res.* **2014**, *13* (1), 15−20.

(6) Omenn, G. S.; Lane, L.; Lundberg, E. K.; Beavis, R. C.; Nesvizhskii, A. I.; Deutsch, E. W. Metrics for the Human Proteome Project 2015: Progress on the Human Proteome and Guidelines for High-Confidence Protein Identification. *J. Proteome Res.* **2015**, DOI: 10.1021/acs.jproteome.5b00499.

(7) Kim, M. S.; Pinto, S. M.; Getnet, D.; Nirujogi, R. S.; Manda, S. S.; Chaerkady, R.; Madugundu, A. K.; Kelkar, D. S.; Isserlin, R.; Jain, S.; Thomas, J. K.; Muthusamy, B.; Leal-Rojas, P.; Kumar, P.; Sahasrabuddhe, N. A.; Balakrishnan, L.; Advani, J.; George, B.; Renuse, S.; Selvan, L. D.; Patil, A. H.; Nanjappa, V.; Radhakrishnan, A.; Prasad, S.; Subbannayya, T.; Raju, R.; Kumar, M.; Sreenivasamurthy, S. K.; Marimuthu, A.; Sathe, G. J.; Chavan, S.; Datta, K. K.; Subbannayya, Y.; Sahu, A.; Yelamanchi, S. D.; Jayaram, S.; Rajagopalan, P.; Sharma, J.; Murthy, K. R.; Syed, N.; Goel, R.; Khan, A. A.; Ahmad, S.; Dey, G.; Mudgal, K.; Chatterjee, A.; Huang, T. C.; Zhong, J.; Wu, X.; Shaw, P. G.; Freed, D.; Zahari, M. S.; Mukherjee, K. K.; Shankar, S.; Mahadevan, A.; Lam, H.; Mitchell, C. J.; Shankar, S. K.; Satishchandra, P.; Schroeder, J. T.; Sirdeshmukh, R.; Maitra, A.; Leach, S. D.; Drake, C. G.; Halushka, M. K.; Prasad, T. S.; Hruban, R. H.; Kerr, C. L.; Bader, G. D.; Iacobuzio-Donahue, C. A.; Gowda, H.; Pandey, A. A Draft Map of the Human Proteome. *Nature* **2014**, *509* (7502), 575−81.

(8) Wilhelm, M.; Schlegl, J.; Hahne, H.; Gholami, A. M.; Lieberenz, M.; Savitski, M. M.; Ziegler, E.; Butzmann, L.; Gessulat, S.; Marx, H.; Mathieson, T.; Lemeer, S.; Schnatbaum, K.; Reimer, U.; Wenschuh, H.; Mollenhauer, M.; Slotta-Huspenina, J.; Boese, J. H.; Bantscheff, M.; Gerstmair, A.; Faerber, F.; Kuster, B. Mass-Spectrometry-Based Draft of the Human Proteome. *Nature* **2014**, *509* (7502), 582−7.

(9) Savitski, M. M.; 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**, mcp.M114.046995 DOI: 10.1074/mcp.M114.046995.

(10) Paik, Y. K.; Omenn, G. S.; Uhlen, M.; Hanash, S.; Marko-Varga, G.; Aebersold, R.; Bairoch, A.; Yamamoto, T.; Legrain, P.; Lee, H. J.; Na, K.; Jeong, S. K.; He, F.; Binz, P. A.; Nishimura, T.; Keown, P.; Baker, M. S.; Yoo, J. S.; Garin, J.; Archakov, A.; Bergeron, J.; Salekdeh, G. H.; Hancock, W. S. Standard Guidelines for the Chromosome-Centric Human Proteome Project. *J. Proteome Res.* **2012**, *11* (4), 2005−13.

(11) Horvatovich, P.; Lundberg, E. K.; Chen, Y. J.; Sung, T. Y.; He, F.; Nice, E. C.; Goode, R. J.; Yu, S.; Ranganathan, S.; Baker, M. S.; Domont, G. B.; Velasquez, E.; Li, D.; Liu, S.; Wang, Q.; He, Q. Y.; Menon, R.; Guan, Y.; Corrales, F. J.; Segura, V.; Casal, J. I.; Pascual-Montano, A.; Albar, J. P.; Fuentes, M.; Gonzalez-Gonzalez, M.; Diez, P.; Ibarrola, N.; Degano, R. M.; Mohammed, Y.; Borchers, C. H.; Urbani, A.; Soggiu, A.; Yamamoto, T.; Salekdeh, G. H.; Archakov, A.; Ponomarenko, E.; Lisitsa, A.; Lichti, C. F.; Mostovenko, E.; Kroes, R. A.; Rezeli, M.; Vegvari, A.; Fehniger, T. E.; Bischoff, R.; Vizcaino, J. A.; Deutsch, E. W.; Lane, L.; Nilsson, C. L.; Marko-Varga, G.; Omenn, G. S.; Jeong, S. K.; Lim, J. S.; Paik, Y. K.; Hancock, W. S. Quest for Missing Proteins: Update 2015 on Chromosome-Centric Human Proteome Project. *J. Proteome Res.* **2015**, DOI: 10.1021/pr5013009.

(12) Marko-Varga, G.; Omenn, G. S.; Paik, Y. K.; Hancock, W. S. A First Step toward Completion of a Genome-Wide Characterization of the Human Proteome. *J. Proteome Res.* **2013**, *12* (1), 1−5.

(13) Nesvizhskii, A. I. Proteogenomics: Concepts, Applications and Computational Strategies. *Nat. Methods* **2014**, *11* (11), 1114−25.

(14) Uhlen, M.; Fagerberg, L.; Hallstrom, B. M.; Lindskog, C.; Oksvold, P.; Mardinoglu, A.; Sivertsson, A.; Kampf, C.; Sjostedt, E.; Asplund, A.; Olsson, I.; Edlund, K.; Lundberg, E.; Navani, S.; Szigyarto, C. A.; Odeberg, J.; Djureinovic, D.; Takanen, J. O.; Hober, S.; Alm, T.; Edqvist, P. H.; Berling, H.; Tegel, H.; Mulder, J.; Rockberg, J.; Nilsson, P.; Schwenk, J. M.; Hamsten, M.; von Feilitzen, K.; Forsberg, M.; Persson, L.; Johansson, F.; Zwahlen, M.; von Heijne, G.; Nielsen, J.; Ponten, F. Tissue-Based Map of the Human Proteome. *Science* **2015**, *347* (6220), 1260419.

(15) Vakilian, H.; Mirzaei, M.; Sharifi Tabar, M.; Pooyan, P.; Habibi Rezaee, L.; Parker, L.; Haynes, P. A.; Gourabi, H.; Baharvand, H.; Salekdeh, G. H. DDX3Y, a Male-Specific Region of Y Chromosome Gene, May Modulate Neuronal Differentiation. *J. Proteome Res.* **2015**, DOI: 10.1021/acs.jproteome.5b00512.

(16) Li, H.; Menon, R.; Govindarajoo, B.; Panwar, B.; Zhang, Y.; Omenn, G. S.; Guan, Y. Functional Networks of Highest-Connected Splice Isoforms, from the Chromosome 17 Human Proteome Project. *J. Proteome Res.* **2015**, DOI: 10.1021/acs.jproteome.5b00494.

(17) Jangravi, Z.; Tabar, M. S.; Mirzaei, M.; Parsamatin, P.; Vakilian, H.; Alikhani, M.; Shabani, M.; Haynes, P. A.; Goodchild, A. K.; Gourabi, H.; Baharvand, H.; Salekdeh, G. H. Two Splice Variants of Y Chromosome-Located Lysine-Specific Demethylase 5D Have Distinct Function in Prostate Cancer Cell Line (DU-145). *J. Proteome Res.* **2015**, DOI: 10.1021/acs.jproteome.5b00333.

(18) Rengaraj, D.; Kwon, W. S.; Pang, M. G. Bioinformatics Annotation of Human Y Chromosome-Encoded Protein Pathways and Interactions. *J. Proteome Res.* **2015**, DOI: 10.1021/acs.jproteome.5b00491.

(19) Menon, R.; Panwar, B.; Eksi, R.; Kleer, C.; Guan, Y.; Omenn, G. S. Computational Inferences of the Functions of Alternative/Non-canonical Splice Isoforms Specific to HER2+/ER-/PR- Breast Cancers, a Chromosome 17 C-HPP Study. *J. Proteome Res.* **2015**, DOI: 10.1021/acs.jproteome.5b00498.

(20) Diez, P.; Droste, C.; Degano, R. M.; Gonzalez-Munoz, M.; Ibarrola, N.; Perez-Andres, M.; Garin-Muga, A.; Segura, V.; Marko-Varga, G.; LaBaer, J.; Orfao, A.; Corrales, F. J.; De Las Rivas, J.; Fuentes, M. Integration of Proteomics and Transcriptomics Data Sets for the Analysis of a Lymphoma B-Cell Line in the Context of the

DOI: 10.1021/acs.jproteome.5b00785
J. Proteome Res. 2015, 14, 3409−3414

Chromosome-Centric Human Proteome Project. *J. Proteome Res.* **2015**, DOI: 10.1021/acs.jproteome.5b00474.
(21) Tay, A. P.; Pang, C. N.; Twine, N. A.; Hart-Smith, G.; Harkness, L.; Kassem, M.; Wilkins, M. R. Proteomic Validation of Transcript Isoforms, Including Those Assembled from RNA-Seq Data. *J. Proteome Res.* **2015**, DOI: 10.1021/pr5011394.
(22) Woo, S.; Cha, S. W.; Bonissone, S.; Na, S.; Tabb, D. L.; Pevzner, P. A.; Bafna, V. Advanced Proteogenomic Analysis Reveals Multiple Peptide Mutations and Complex Immunoglobulin Peptides in Colon Cancer. *J. Proteome Res.* **2015**, DOI: 10.1021/acs.jproteome.5b00264.
(23) Jayaram, S.; Gupta, M. K.; Shivakumar, B. M.; Ghatge, M.; Sharma, A.; Vangala, R. K.; Sirdeshmukh, R. Insights from Chromosome-Centric Mapping of Disease-Associated Genes: Chromosome 12 Perspective. *J. Proteome Res.* **2015**, DOI: 10.1021/acs.jproteome.5b00488.
(24) Eckhard, U.; Marino, G.; Abbey, S. R.; Tharmarajah, G.; Matthew, I.; Overall, C. M. 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**, DOI: 10.1021/acs.jproteome.5b00579.
(25) 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.
(26) 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**, DOI: 10.1021/acs.jproteome.5b00435.
(27) Ahmadi Rastegar, D.; Sharifi Tabar, M.; Alikhani, M.; Parsamatin, P.; Sahraneshin Samani, F.; Sabbaghian, M.; Sadighi Gilani, M. A.; Mohammad Ahadi, A.; Mohseni Meybodi, A.; Piryaei, A.; Ansari-Pour, N.; Gourabi, H.; Baharvand, H.; Salekdeh, G. H. Isoform-Level Gene Expression Profiles of Human Y Chromosome Azoospermia Factor Genes and Their X Chromosome Paralogs in the Testicular Tissue of Non-Obstructive Azoospermia Patients. *J. Proteome Res.* **2015**, DOI: 10.1021/acs.jproteome.5b00520.
(28) Jumeau, F.; Com, E.; Lane, L.; Duek, P.; Lagarrigue, M.; Lavigne, R.; Guillot, L.; Rondel, K.; Gateau, A.; Melaine, N.; Guevel, B.; Sergeant, N.; Mitchell, V.; Pineau, C. Human Spermatozoa as a Model for Detecting Missing Proteins in the Context of the Chromosome-Centric Human Proteome Project. *J. Proteome Res.* **2015**, DOI: 10.1021/acs.jproteome.5b00170.
(29) Carapito, C.; Lane, L.; Benama, M.; Opsomer, A.; Mouton-Barbosa, E.; Garrigues, L.; Gonzalez de Peredo, A.; Burel, A.; Bruley, C.; Gateau, A.; Bouyssie, D.; Jaquinod, M.; Cianferani, S.; Burlet-Schiltz, O.; Van Dorsselaer, A.; Garin, J.; Vandenbrouck, Y. Computational and Mass-Spectrometry-Based Workflow for the Discovery and Validation of Missing Human Proteins: Application to Chromosomes 2 and 14. *J. Proteome Res.* **2015**, DOI: 10.1021/pr5010345.
(30) 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**, DOI: 10.1021/acs.jproteome.5b00565.
(31) Yang, L.; Lian, X.; Zhang, W.; Guo, J.; Wang, Q.; Li, Y.; Chen, Y.; Yin, X.; Yang, P.; Lan, F.; He, Q. Y.; Zhang, G.; Wang, T. Finding Missing Proteins from the Epigenetically Manipulated Human Cell with Stringent Quality Criteria. *J. Proteome Res.* **2015**, DOI: 10.1021/acs.jproteome.5b00480.
(32) Horvatovich, P.; Vegvari, A.; Saul, J.; Park, J. G.; Qiu, J.; Syring, M.; Pirrotte, P.; Petritis, K.; Tegeler, T. J.; Aziz, M.; Fuentes, M.; Diez, P.; Gonzalez-Gonzalez, M.; Ibarrola, N.; Droste, C.; De Las Rivas, J.; Gil, C.; Clemente, F.; Hernaez, M. L.; Corrales, F. J.; Nilsson, C. L.; Berven, F. S.; Bischoff, R.; Fehniger, T. E.; LaBaer, J.; Marko-Varga, G. In Vitro Transcription/Translation System: A Versatile Tool in the Search for Missing Proteins. *J. Proteome Res.* **2015**, DOI: 10.1021/acs.jproteome.5b00486.
(33) Kitata, R. B.; Dimayacyac-Esleta, B. R.; Choong, W. K.; Tsai, C. F.; Lin, T. D.; Tsou, C. C.; Weng, S. H.; Chen, Y. J.; Yang, P. C.; Arco, S. D.; Nesvizhskii, A. I.; Sung, T. Y.; Chen, Y. J. Mining Missing Membrane Proteins by High-pH Reverse-Phase StageTip Fractionation and Multiple Reaction Monitoring Mass Spectrometry. *J. Proteome Res.* **2015**, DOI: 10.1021/acs.jproteome.5b00477.
(34) Hoover, H.; Li, J.; Marchese, J.; Rothwell, C.; Borawoski, J.; Jeffery, D. A.; Gaither, L. A.; Finkel, N. Quantitative Proteomic Verification of Membrane Proteins as Potential Therapeutic Targets Located in the 11q13 Amplicon in Cancers. *J. Proteome Res.* **2015**, DOI: 10.1021/acs.jproteome.5b00508.
(35) Su, N.; Zhang, C.; Zhang, Y.; Wang, Z.; Fan, F.; Zhao, M.; Wu, F.; Gao, Y.; Li, Y.; Chen, L.; Tian, M.; Zhang, T.; Wen, B.; Sensang, N.; Xiong, Z.; Wu, S.; Liu, S.; Yang, P.; Zhen, B.; Zhu, Y.; He, F.; Xu, P. Special Enrichment Strategies Greatly Increase the Efficiency of Missing Proteins Identification from Regular Proteome Samples. *J. Proteome Res.* **2015**, DOI: 10.1021/acs.jproteome.5b00481.
(36) Chen, Y.; Li, Y.; Zhong, J.; Zhang, J.; Chen, Z.; Yang, L.; Cao, X.; He, Q. Y.; Zhang, G.; Wang, T. Identification of Missing Proteins Defined by Chromosome-Centric Proteome Project in the Cytoplasmic Detergent-Insoluble Proteins. *J. Proteome Res.* **2015**, DOI: 10.1021/pr501103r.
(37) Jeong, S. K.; Hancock, W. S.; Paik, Y. K. GenomewidePDB 2.0: A Newly Upgraded Versatile Proteogenomic Database for the Chromosome-Centric Human Proteome Project. *J. Proteome Res.* **2015**, DOI: 10.1021/acs.jproteome.5b00541.
(38) Yang, S.; Zhang, X.; Diao, L.; Guo, F.; Wang, D.; Liu, Z.; Li, H.; Zheng, J.; Pan, J.; Nice, E. C.; Li, D.; He, F. CAPER 3.0: A Scalable Cloud-Based System for Data-Intensive Analysis of Chromosome-Centric Human Proteome Project Data Sets. *J. Proteome Res.* **2015**, DOI: 10.1021/pr501335w.
(39) Krasnov, G. S.; Dmitriev, A. A.; Kudryavtseva, A. V.; Shargunov, A. V.; Karpov, D. S.; Uroshlev, L. A.; Melnikova, N. V.; Blinov, V. M.; Poverennaya, E. V.; Archakov, A. I.; Lisitsa, A. V.; Ponomarenko, E. A. PPLine: An Automated Pipeline for SNP, SAP, and Splice Variant Detection in the Context of Proteogenomics. *J. Proteome Res.* **2015**, DOI: 10.1021/acs.jproteome.5b00490.
(40) Tabas-Madrid, D.; Alves-Cruzeiro, J.; Segura, V.; Guruceaga, E.; Vialas, V.; Prieto, G.; Garcia, C.; Corrales, F. J.; Albar, J. P.; Pascual-Montano, A. Proteogenomics Dashboard for the Human Proteome Project. *J. Proteome Res.* **2015**.
(41) Panwar, B.; Menon, R.; Eksi, R.; Omenn, G. S.; Guan, Y. MI-PVT: A Tool for Visualizing the Chromosome-Centric Human Proteome. *J. Proteome Res.* **2015**, DOI: 10.1021/acs.jproteome.5b00525.
(42) Dong, Q.; Menon, R.; Omenn, G. S.; Zhang, Y. Structural Bioinformatics Inspection of neXtProt PE5 Proteins in the Human Proteome. *J. Proteome Res.* **2015**, DOI: 10.1021/acs.jproteome.5b00516.
(43) Schubert, O. T.; Gillet, L. C.; Collins, B. C.; Navarro, P.; Rosenberger, G.; Wolski, W. E.; Lam, H.; Amodei, D.; Mallick, P.; MacLean, B.; Aebersold, R. Building High-Quality Assay Libraries for Targeted Analysis of SWATH MS Data. *Nat. Protoc.* **2015**, *10* (3), 426−41, DOI: 10.1038/nprot.2015.015.
(44) Ezkurdia, I.; Vazquez, J.; Valencia, A.; Tress, M. Analyzing the First Drafts of the Human Proteome. *J. Proteome Res.* **2014**, *13* (8), 3854−55.

DOI: 10.1021/acs.jproteome.5b00785
J. Proteome Res. 2015, 14, 3409−3414