Journal of proteome research

pubs.acs.org/jpr

Editorial

The HUPO High-Stringency Inventory of Humanity's Shared Human Proteome Revealed

Cite This: https://dx.doi.org/10.1021/acs.jproteome.0c00794

Read Online

ACCESS | Metrics & More | Article Recommendations

In the midst of a pandemic, in the midst of a global effort to develop effective vaccines and antivirals for SARS-CoV-2—yet paradoxically, also in the midst of a surreal moment in history when the very science that can save millions is assailed if the facts and truth conflict with political mantra—we nonetheless can celebrate. Reminding us all of the importance and relevance of science, one of humanity's greatest scientific achievements occurred 20 years ago on June 26, 2000 with the completion of the draft sequence of the human genome. Whereas the genome is the genetic blueprint of humans, the proteome—all proteins encoded by the human genome—is our working architecture. Today, October 19, 2020, we celebrate the release of the draft human proteome[1] by the international Human Proteome Organization (HUPO) at the 19th Human Proteome Organization World Congress, connecting virtually, with this Virtual Issue of the *Journal of Proteome Research*, "Celebrating 90% Completion of the Human Proteome". Here we compile 60 of the most significant papers published in the Journal over the past decade of the Human Proteome Project (HPP).

In the year of the release of the draft of the human genome and in recognition of the importance of the expression and functions of the human proteome—then estimated to be encoded by 32 000 genes—HUPO was established in 2001. 20 years later and 10 years after the launch of the HPP by HUPO on September 23, 2010, we have much to celebrate with the reporting of the HUPO high-stringency draft inventory of humanity's shared proteome. The neXtProt database posted the landmark human proteome data release covering 90% of the human proteome on January 17, 2020,[2] which is now reported by the HPP Consortium in *Nature Communications* by Adhikari et al.[1] The human proteome was identified by HPP global research teams and scientists from the wider scientific community and assembled by the Chromosome-Centric HPP (C-HPP) and the HPP Knowledge-Base Pillar data curators from neXtProt,[3,4] PeptideAtlas,[5,6] and MassIVE.[7] The C-HPP[8] was established in 2010 as the major initiative of the HPP to identify at least one protein form (proteoform) from each of the protein-encoding genes in the human genome. neXtProt is the official HPP knowledgebase of the human proteome, developed and curated by Dr. Lydie Lane's group at SIB Swiss Institute of Bioinformatics. NeXtprot provides a readily accessible framework that translates the peptide and protein identifications assembled from the UniProt Swiss-Prot database with added proteomic data derived from the Human Peptide Atlas, led by Drs. Eric Deutsch and Rob Moritz (Chair, Human Proteome Project), Institute of Systems Biology, Seattle, and, starting this year, also from MassIVE, developed by Dr. N. Bandeira, UCSD.

In 2011, just 70% of human proteins were credibly identified with protein existence (PE) level 1 evidence. Despite little governmental financial support for the HPP and proteomics, in general, compared with the Human Genome Project and genomics, just 10 years after the launch of the HPP, Adhikari et al. have reported the identification of 17 874 PE1 proteins translated from the 19 773 protein-encoding genes, which represents 90% of the human proteome now rigorously identified at the protein level. In the companion annual human proteome metrics paper by Omenn et al.[9] reporting this year's progress of the HPP, the underlying data are presented in depth. The metrics paper will be published in the eighth special issue of the *Journal of Proteome Research* dedicated to the HPP in December 2020, "Human Proteome Project 2020", and was published ASAP today,[9] leading this HPP Virtual Issue of the Journal.

The HPP international consortium is now structured on two initiatives: the Chromosome-Centric HPP (C-HPP)[10] and the Biology/Disease-Driven HPP (B/D-HPP),[11] supported by four resource pillars: Antibody Resource Pillar, Pathology Pillar, Mass Spectrometry Pillar, and Knowledge-Base Pillar. The HPP has welcomed teams of collaborating scientists from all around the world, including China, Switzerland, Japan, Taiwan, Netherlands, Canada, United States, Australia, New Zealand, Korea, India, Brazil, France, Spain, Russia, Mexico, Iran, and Italy, who participate in this global enterprise by contributing data and expertise. The second principle of HUPO has been free and open access to proteomic data. On publication, all proteomic data are uploaded to databases in the ProteomeXchange Consortium, which PeptideAtlas and MassIVE scrape to aggregate the data that are annotated in neXtProt, to be made publicly available for free.

The draft of the Human Genome was published on February 15, 2001 in two versions by the International Human Genome Sequencing Consortium[12] and by the U.S. biotechnology company Celera.[13] Despite many parallels with today's human proteome, there are also key differences in the quality of data released in the first drafts of the human genome and the human proteome. Sequencing the human genome started on October 1, 1990, and on June 26, 2000, the International Human Genome Sequencing Consortium announced the rough draft of 90% of the human genome sequence. With an error rate of 1/1000 and with 148 000 gaps from the estimated ~32 000 genes, the shotgun phase of the project then transitioned to the finishing phase, which progressed quickly. In April 2003, the accurate sequence of 99% of the human genome was announced with an error rate 1/10 000 and just 341 gaps in the revised estimate of 20 000−25 000 protein-coding genes. In contrast, the human proteome first draft is highly accurate for 90% of the proteins (designated PE1), with some 1596 additional candidate proteins having mRNA transcript evidence of their existence (PE2) but, to date, remaining entirely missing from the known proteome. The gaps lie not in the protein sequences themselves, which are known from start to end, but instead from the 10% of the proteome "parts list" that has so far escaped detection at the protein level. These proteins lack credible evidence of existence at the protein level and have come to be known as "Missing Proteins" (PE2, PE3, PE4), now numbering 1899 and spread across all 22 autosomes and the X and Y chromosomes. The 15 proteins encoded on the mitochondrial DNA were all previously identified by the Italian C-HPP team—the first completion of a "chromosome"—and now contain no missing proteins.

The finishing phase of the human proteome is more difficult than that for the human genome, which, in fact, was accelerated due to the maturity of the sequencing technologies, computing power, and bioinformatics advances. Proteomics is more difficult due to the higher complexity of the polypeptide chain composed of 20 amino acids compared with the 4 nucleotides of DNA, the >400 post-translational modifications of amino acids,[14] the splice forms, the alternate start sites, and the proteolytic processing of all polypeptides,[15] which together generate millions of proteoforms and a dynamic proteome. These factors, coupled to the sequencing and bioinformatics limitations arising from the shorter protein peptides generated by trypsin compared with DNA fragments generated from *Hin*dIII cleavage used in the initial phase of the human genome project, render the shotgun assembly of proteins from tryptic peptides often more difficult than that of genomic sequences. Furthermore, whereas genomic DNA resides in almost all human cells, the repertoire of proteins expressed by any cell or tissue type is restricted to a core set of proteins necessary for the essential cellular functions plus cell- or task-specific proteins. Thus the expression of missing proteins is not universal to all cells, which would simplify their detection. Rather, missing proteins are limited in abundance, time of expression, spatial distribution, cell or tissue of origin, and amenability to mass spectrometric detection. This renders missing protein detection challenging, and increasingly so, by the law of diminishing returns.

■ LESSONS LEARNED

So, what have we learned from the human proteome? Proteomes evolve through natural selection on evolutionary time scales. DNA was shuffled between bacteria and humans, between viruses and us, and within our own genes. By the latter, exon shuffling generated new protein architectures. This is especially successful when gene duplication occurs, as it allows the parental protein to maintain the essential characteristic functions of the protein, while the duplicated daughter proteins are free to evolve according to new selective pressures. Thereby, evolution assembled and evolved new proteins from discrete smaller functional protein modules to generate new functions in new locations in new cells—some of which we have learned have a critical role in protection from disease.

Deficiencies in the proteome "parts" can stem from inherited genetic mutations, leading to genetic diseases or manifest only by environmental, nutritional, and infection stressors that lead to defective or inadequate immune and cellular responses, putting such individuals at greater risk of disease or its pathobiological consequences. Knowledge of the individual proteins that are key to protection from disease and their deficiencies in expression or activity that are hallmarks of disease can inform individualized medicine relevant to the particular defective protein or pathway in the affected individual, enabling treatment with the optimal specific drugs, where available.

Amazingly, humans share 99.9% identity in their DNA between individuals, yet one base-pair change in our genes can lead to individuality and predispositions to disease. But how? Immunodeficiencies—the boy in the bubble—are such an example, where one DNA base-pair change in a gene leads to a single amino acid substitution in the translated protein. Where this protein is essential to an immune signaling pathway, this can lead to immunodeficiency disease.[16] Proteomics provides this essential missing functional information, which genomics cannot. Knowing the amino acid, proteomics can decipher the effects of this substitution on the proteoform and protein and cellular function and expression and lead to the development of new therapies. Novel molecular corrector drugs have demonstrated the correction of the defective amino acid. Such drugs function like a molecular prosthetic to restore critical protein function.[17]

Post-translational modification differences of key proteins can turn a protein activity on or off in a cell-specific manner or change a protein's cellular expression, intracellular localization, or half life. Aberrations in the regulation of post-translational modifications can form the molecular mechanism that both underlies and is diagnostic of a multitude of human diseases. In infection, changes in the proteomes of the infected cell and tissue wrought by microorganisms or viruses cannot be determined by genomics. Only proteomics can decipher these. In COVID-19, there are two proteomes involved, that of the SARS-CoV-2 virus and that of the infected cells, both of which likely interact with the other, modify the other, and change the function of the other. This interconnection needs to be understood, in particular, the post-translational modifications altering the form and function of both proteomes, rendering some cells and individuals more resilient to COVID-19 and others, sadly, more vulnerable. Notably, the impact of the viral enzymes, especially the two SARS-CoV-2 proteases, 3CLpro and PLpro, in decapitating essential cell proteins and pathways while keeping the infected cell alive, enables the virus to infect, circumvent cellular protection, and therefore replicate and spread. The key to deciphering COVID-19 pathobiology is proteomics, one subfield of which is degradomics, whereby the N-terminome composed of the intact and protease-cleaved neo-N-termini of proteins identifies the viral protease substrates and host-cell targets in infection.[18] Knowledge of the exact cleavage site and protein substrates enables rational choices to be made in devising new antiviral treatments aimed at restoring the targeted protein essential functions and thereby dampening or mitigating the infection. Hence genetic mutations and polymorphisms, overlaid by proteome post-translational modifications, splice forms, and proteolytic proteoforms of the same genetically encoded protein, form a framework to understand human individuality and the risk and propensity of disease. This information cannot be derived by genomics and individual DNA sequences. Only proteomics can provide this higher order level of knowledge at the protein and protein-complex levels that is so critical in understanding and diagnosing disease. Deciphering this new "proteome code" is the challenge that lies ahead for the proteomics and HPP communities and for addressing the broken hyperbole springing from the euphoria of the publication of the human genome papers 20 years ago, when the media and pundits predicted the curing of some, if not all, human diseases within a few years.

MIND THE GAP

Today's published draft of the human proteome is a triumph, sufficient for the deeper understanding of human individuality and disease, yet there is a huge amount of work left to extend it. The 10% missing protein gap in completing the overall coverage of the human proteome will hold further keys to understanding human embryonic and childhood development, cell differentiation, and less frequent yet essential responses to disease and environmental and dietary challenges that were essential for hominid survival and evolution from ~2.8 million years ago to today's modern human. The HPP "minds the gap" and so aims to provide evidence of all human protein-encoding genes and is committed to closing the 10% proteome gap with high fidelity.[19] The HPP also aims to probe the function of the 1254 individual proteins with no known function or predicted function—Donald Rumsfeld's "known unknowns"—many of which will prove to be essential for normal physiological and pathological processes and some of which will prove to be unexpected and promising new drug targets.[20] But how many functional "unknown unknowns" or, indeed, unknown proteins lurk, some in plain sight and others subtly present, that need to be discerned and deciphered? How many functions emerge only upon the generation of higher order protein complexes in and out of cells? A cat, a millisecond after death, will have an indistinguishable genome and proteome from that a moment before, yet there is a vast difference in the emergent properties that arose from the same collection of genes and proteins. How is this possible? Comparative proteomics will build different protein interactors and interaction groups between individuals and define proteome differences among the diverse human populations of the world. Such insight will provide clues and answers as to disease susceptibility and responses to treatment. For example, protein and protein-complex alterations and differences between susceptible and resistant patients that are critically important will point to potential new treatments for which drugs already exist for these proteins or become new drug targets for the pharmaceutical industry.

Unlike the human genome, where the polishing phase proceeded quickly, for the proteome, this is predicted to proceed slower. To complete this task, we need new technology for improved coverage, higher sensitivity for single-cell proteomics, and machine-learning-aided bioinformatics to provide an accessible framework for data access for scientists and clinicians to make sense of the vast new information and knowledge sets and data records for each patient. Finally, the proteomics community needs government and institutional awareness to provide support and resources for these essential research challenges. It is ironical that the very existence of neXtProt, which released the data revealing the 90% completion of the human proteome, is now in doubt due to the lack of financial support. For the next high-fidelity compendium of the full human proteome and to develop a broader understanding of life, human conscience, and disease, proteomics needs more data, more patients, more scientists—biochemists, geneticists, engineers, mathematicians, and bioinformaticians, and more doctors to understand life, individuality, personality, and disease. Science needs us all, but now, more than ever, humanity needs more science.

Christopher M. Overall, Chair, Chromosome-Centric Human Proteome Project orcid.org/0000-0001-5844-2731

AUTHOR INFORMATION

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jproteome.0c00794

Notes

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

The author declares no competing financial interest.

ACKNOWLEDGMENTS

C.M.O. is supported by a Canadian Institutes of Health Research 7-year Foundation Grant (FDN148408), a 2019 Novel Coronavirus (COVID-19) Rapid Research Funding Opportunity (F20-01013), and a Canada Research Chair in Protease Proteomics and Systems Biology.

REFERENCES

(1) Adhikari, S.; Nice, E.; Deutsch, E.; Lane, L.; Omenn, G.; Pennington, S.; Paik, Y.-K.; Overall, C. M.; Corrales, F.; Cristea, I.; Van Eyk, J.; Uhlen, M.; Lindskog, C.; Chan, D.; Bairoch, A.; Waddington, J.; Justice, J.; LaBaer, J.; Rodriguez, H.; He, F.; Kostrzewa, M.; Ping, P.; Gundry, R.; Stewart, P.; Srivastava, S.; Srivastava, S.; Nogueira, F.; Domont, G.; Vandenbrouck, Y.; Lam, M.; Wennersten, S.; Vizcaino, J. A.; Wilkins, M.; Schwenk, J.; Lundberg, E.; Bandeira, N.; Marko-Varga, G.; Weintraub, S.; Pineau, C.; Kusebauch, C.; Moritz, R.; Ahn, S. B.; Palmblad, M.; Snyder, M.; Aebersold, R.; Baker, M. A High-Stringency Blueprint of the Human Proteome. *Nat. Communications* **2020**, DOI: 10.1038/s41467-020-19045-9.

(2) neXtProt Relase Statistics. https://www.nextprot.org/about/statistics (accessed October 7, 2020).

(3) Lane, L.; Argoud-Puy, G.; Britan, A.; Cusin, I.; Duek, P. D.; Evalet, O.; Gateau, A.; Gaudet, P.; Gleizes, A.; Masselot, A.; Zwahlen, C.; Bairoch, A. neXtProt: A Knowledge Platform for Human Proteins. *Nucleic Acids Res.* **2012**, *40* (D1), D76−D83.

(4) Gaudet, P.; Michel, P. A.; Zahn-Zabal, M.; Britan, A.; Cusin, I.; Domagalski, M.; Duek, P. D.; Gateau, A.; Gleizes, A.; Hinard, V.; Rech de Laval, V.; Lin, J.; Nikitin, F.; Schaeffer, M.; Teixeira, D.; Lane, L.; Bairoch, A. The neXtProt knowledgebase on human proteins: 2017 update. *Nucleic Acids Res.* **2017**, *45* (D1), D177−D182.

(5) Peptide Atlas. http://www.peptideatlas.org (accessed October 7, 2020).

(6) Desiere, F.; Deutsch, E. W; Nesvizhskii, A. I; Mallick, P.; King, N. L; Eng, J. K; Aderem, A.; Boyle, R.; Brunner, E.; Donohoe, S.; Fausto, N.; Hafen, E.; Hood, L.; Katze, M. G; Kennedy, K. A; Kregenow, F.; Lee, H.; Lin, B.; Martin, D.; Ranish, J. A; Rawlings, D. J; Samelson, L. E; Shiio, Y.; Watts, J. D; Wollscheid, B.; Wright, M. E; Yan, W.; Yang, L.; Yi, E. C; Zhang, H.; Aebersold, R. Integration with

the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol.* **2004**, *6*, R9.

(7) MassIVE. https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp (accessed October 7, 2020).

(8) Paik, Y.; Jeong, S.; Omenn, G.; et al. The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. *Nat. Biotechnol.* **2012**, *30*, 221−223.

(9) Omenn, G. S.; Lane, L.; Overall, C. M.; Cristea, I. M.; Corrales, F. J.; Lindskog, C.; Paik, Y.-K.; Van Eyk, Eyk; Liu, S.; Pennington, S.; Snyder, M. P.; Baker, M.; Bandeira, N.; Aebersold, R.; Moritz, R. L.; Deutsch, E. W. Research on The Human Proteome Reaches a Major Milestone: >90% of Predicted Human Proteins Now Credibly Detected, According to the HUPO Human Proteome Project. *J. Proteome Res.* **2020**, DOI: 10.1021/acs.jproteome.0c00485.

(10) HUPO C-HPP. https://hupo.org/C-HPP (accessed October 7, 2020).

(11) HUPO B/D-HPP. https://www.hupo.org/B/D-HPP (accessed October 7, 2020).

(12) International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **2001**, *409*, 860−921.

(13) Venter, J. C.; Adams, M. D.; Myers, E. W.; Li, P. W.; Mural, R. J.; Sutton, G. G.; Smith, H. O.; Yandell, M.; Evans, C. A.; Holt, R. A.; et al. The Sequence of the Human Genome. *Science* **2001**, *291*, 1304−1351.

(14) Aebersold, R.; Agar, J. N; Amster, I J.; Baker, M. S; Bertozzi, C. R; Boja, E. S; Costello, C. E; Cravatt, B. F; Fenselau, C.; Garcia, B. A; Ge, Y.; Gunawardena, J.; Hendrickson, R. C; Hergenrother, P. J; Huber, C. G; Ivanov, A. R; Jensen, O. N; Jewett, M. C; Kelleher, N. L; Kiessling, L. L; Krogan, N. J; Larsen, M. R; Loo, J. A; Ogorzalek Loo, R. R; Lundberg, E.; MacCoss, M. J; Mallick, P.; Mootha, V. K; Mrksich, M.; Muir, T. W; Patrie, S. M; Pesavento, J. J; Pitteri, S. J; Rodriguez, H.; Saghatelian, A.; Sandoval, W.; Schluter, H.; Sechi, S.; Slavoff, S. A; Smith, L. M; Snyder, M. P; Thomas, P. M; Uhlen, M.; Van Eyk, J. E; Vidal, M.; Walt, D. R; White, F. M; Williams, E. R; Wohlschlager, T.; Wysocki, V. H; Yates, N. A; Young, N. L; Zhang, B. How Many Human Proteoforms are There? *Nat. Chem. Biol.* **2018**, *14* (3), 206−214.

(15) Klein, T.; Eckhard, U.; Dufour, A.; Solis, N.; Overall, C. M. Proteolytic Cleavage—Mechanisms, Function, and "Omic" Approaches for a Near-Ubiquitous Posttranslational Modification. *Chem. Rev.* **2018**, *118*, 1137−1168.

(16) Turvey, S. E.; Durandy, A.; Fischer, A.; Fung, S.-Y.; Geha, R. S.; Gewies, A.; Giese, T.; Greil, J.; Keller, B.; McKinnon, M. L.; Neven, B.; Rozmus, J.; Ruland, J.; Snow, A. L.; Stepensky, P.; Warnatz, K. The CARD11-BCL10-MALT1 (CBM) Signalosome Complex: Stepping Into the Limelight of Human Primary Immunodeficiency. *J. Allergy Clin. Immunol.* **2014**, *134*, 276−284.

(17) Quancard, J.; Klein, T.; Fung, S.-Y.; Renatus, M.; Hughes, N.; Israël, L.; Priatel, J. J.; Kang, S.; Blank, M. A.; Viner, R. I.; Blank, J.; Schlapbach, A.; Erbel, P.; Kizhakkedathu, J.; Villard, F.; Hersperger, R.; Turvey, S. E.; Eder, J.; Bornancin, F.; Overall, C. M. An Allosteric MALT1 Inhibitor is a Molecular Corrector Rescuing Function in an Immunodeficient Patient. *Nat. Chem. Biol.* **2019**, *15*, 304−313.

(18) Jagdeo, J. M.; Dufour, A.; Klein, T.; Solis, N.; Kleifeld, O.; Kizhakkedathu, J. N.; Luo, H.; Overall, C. M.; Jan, E. N-Terminomics TAILS Identifies Host Cell Substrates of Poliovirus and Coxsackievirus B3 3C Proteinases that Modulate Virus Infection. *J. Virol.* **2018**, *92*, No. e02211-17.

(19) Deutsch, E. W.; Lane, L.; Overall, C. M.; Bandeira, N.; Baker, M. S.; Pineau, C.; Moritz, R. L.; Corrales, F.; Orchard, S.; Van Eyk, J. E.; Paik, Y.-K.; Weintraub, S. T.; Vandenbrouck, Y.; Omenn, G. S. Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 3.0. *J. Proteome Res.* **2019**, *18*, 4108−4116.

(20) Paik, Y.-K.; Lane, L.; Kawamura, T.; Chen, Y. J.; Cho, J. Y.; LaBaer, J.; Yoo, J. S.; Domont, G.; Corrales, F.; Omenn, G. S.; Archakov, A.; Encarnación-Guevara, S.; Lui, S.; Salekdeh, G. H.; Cho, J. Y.; Kim, C. Y.; Overall, C. M. Launching the C-HPP neXt-CP50 Pilot Project for Functional Characterization of Identified Proteins with No Known Function. *J. Proteome Res.* **2018**, *17*, 4042−4050.