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

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 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, 2020,¹ 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,² PeptideAtlas,³,⁴ and MassIVE.⁵ The C-HPP 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.⁶ 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, 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: Antbody 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 Consortium¹² 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 HindIII cleavage used in the initial phase of the human genome project, render the shotgun assembly of proteins from trypptic 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 character-
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 anchoring in the emergent properties 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 of 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 (FDNI48408), a 2019 Novel Coronavirus (COVID-19) Rapid Research Funding Opportunity (F20-01013), and a Canada Research Chair in Protease Proteomics and Systems Biology.

**REFERENCES**


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


(9) Omenn, G. S.; Lane, L.; Overall, C. M.; Cristea, I. M.; Corrales, F. J.; Lindskog, C.; Paik, Y.-K.; Van Eyk, Eyr; Liu, S.; Pennington, S.; Snyder, M. P.; Baker, M.; Bandeira, N.; Abeirosld, 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.


