The Human Proteoform Project: Bringing Proteoforms to Life

A Plan to Define the Human Proteome

Lloyd M. Smith[8], Jeffrey N. Agar[2], Julia Chamot-Rooke[3], Paul O. Danis[4] Ying Ge[5] Joseph A. Loo[6], Ljiljana Paša-Tolić[7], Yury O. Tsybin[8], Neil L. Kelleher[*9]

[1]Department of Chemistry, University of Wisconsin, Madison, Wisconsin, USA
[2]Departments of Chemistry and Chemical Biology and Pharmaceutical Sciences, Northeastern University, Boston, Massachusetts, USA
[3]Department of Structural Biology and Chemistry, Institut Pasteur, CNRS, Paris France
[4]Consortium for Top-Down Proteomics, Cambridge, Massachusetts, USA
[5]Department of Cell and Regenerative Biology, Department of Chemistry, Human Proteomics Program, University of Wisconsin-Madison, Madison, Wisconsin, USA
[6]Department of Chemistry and Biochemistry, University of California, Los Angeles, Los Angeles, California, USA
[7]Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, Washington, USA
[8]Spectroswiss, Lausanne, Switzerland
[9*]Departments of Chemistry, Molecular Biosciences and the Feinberg School of Medicine, Northwestern University, Evanston, Illinois, USA; n-kelleher@northwestern.edu

ABSTRACT

Proteins are the primary effectors of function in biology, and thus complete knowledge of their structure and properties is fundamental to deciphering function in basic and translational research. The chemical diversity of proteins is expressed in their many proteoforms, which result from combinations of genetic polymorphisms, RNA splice variants and post-translational modifications. This knowledge is foundational for the biological complexes and networks that control biology, yet remains largely unknown. We propose here an ambitious initiative to define the human proteome; that is to generate a definitive reference set of the proteoforms produced from the genome. Several examples of the power and importance of proteoform-level knowledge in disease-based research are presented, along with a call for improved technologies in a two-pronged strategy to accomplish the Human Proteoform Project.

Keywords: proteoform, human genome project, proteomics, post-translational modification, human proteome

Table of Contents

The Human Proteoform Project: Bringing Proteoforms to Life

The Consortium for Top-Down Proteomics

1. Context and Central Goals

The Human Genome Project was a remarkable and unqualified success, profoundly transforming and accelerating biological and medical research, converting a ~$4B public investment into over $700B of economic activity and new industries [1]. The challenge of revealing the "Blueprints of Life", however, is surpassed by the challenge we face today: deriving from these blueprints an understanding of the structures they dictate and how these function within biological systems.

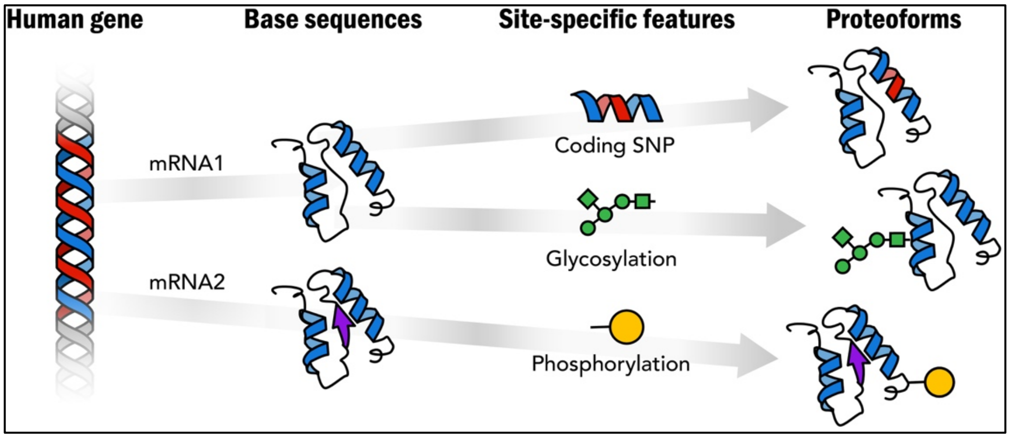

Figure 1. Proteoforms: distinct protein forms arising from a single human gene.

Proteins are primary effectors of function in biology, and thus complete knowledge of their structure and behavior is fundamental to deciphering function in basic and translational research [2]. The richness of protein structure and function goes far beyond the linear amino acid sequence dictated by the genetic code. Multigene families, alternative splicing, coding polymorphisms, and post-translational modifications (PTMs) work together to create a rich variety of different proteoforms arising from our genes (**Figure 1**) [3]. The chemical diversity of proteins is foundational for the biological complexes and networks that control biology, yet remains largely unknown. Genome sequence alone does not provide the needed information – only direct analysis of the proteoforms themselves can reveal their composition, enabling studies of their temporal dynamics and spatial distributions in biological systems. We propose here an ambitious initiative to define the human proteome; that is to generate a definitive reference set of the proteoforms produced from the genome (see **Box 1**).

Box 1. What is a proteome?

A standard answer to this question is that a proteome is the set of proteins expressed by an organism. This idea clearly depends on what is meant by a "protein". Proteins from even a single gene can vary widely in their amino acid sequence and post-translational modifications, giving rise to a variety of proteoforms. Thus, the proteome is necessarily the set of all proteoforms expressed by an organism. The initiative proposed here is founded upon this simple idea.

1.1. Significance of Proteoforms

Proteins are the central intermediaries between genotype and phenotype [2-4]. It is not possible to understand the functioning of a biological system if one does not know what protein molecules are present, and the nature and abundances of their proteoforms. Knowledge of where the proteoforms are located within cells or tissues, what other proteoforms they interact with to form the powerful multi-functional complexes that carry out critical functions in cell biology, and how they change in response to stimuli, is essential. Innovative new tools are needed to comprehensively define the proteome, allowing proteoform abundances, interactors, and locations to be assessed with far greater depth and facility. The foundational premise of the Human Genome Project, that knowledge of the Genome Sequence will provide a fundamental understanding of biological systems, will not be realized in the absence of detailed proteoform-level information.

1.2. Central Goal and Strategy

We propose here a plan to elucidate the complete human proteome– a community effort to identify a definitive reference set of the expressed proteoforms derived from the ~20,000 gene blueprint encoded in the human genome. We outline a two-pronged strategy: on the one hand, we will pursue deep proteoform-level analysis of critical medically-relevant systems (neurodegeneration, cardiovascular health, infectious disease, cancer, immunobiology, see **Figure 2** and **Section 2**); this will open up fundamental new insights into these critical medical targets. In parallel, we will invest heavily in the accelerated development of proteoform discovery and characterization technologies, and apply these technologies to global proteoform-wide analysis (see **Section 3**).

The plan is modeled roughly after the successful roadmap provided by the Human Genome Project (HGP), which provided a reference sequence for the Human Genome [2, 3, 5]. An

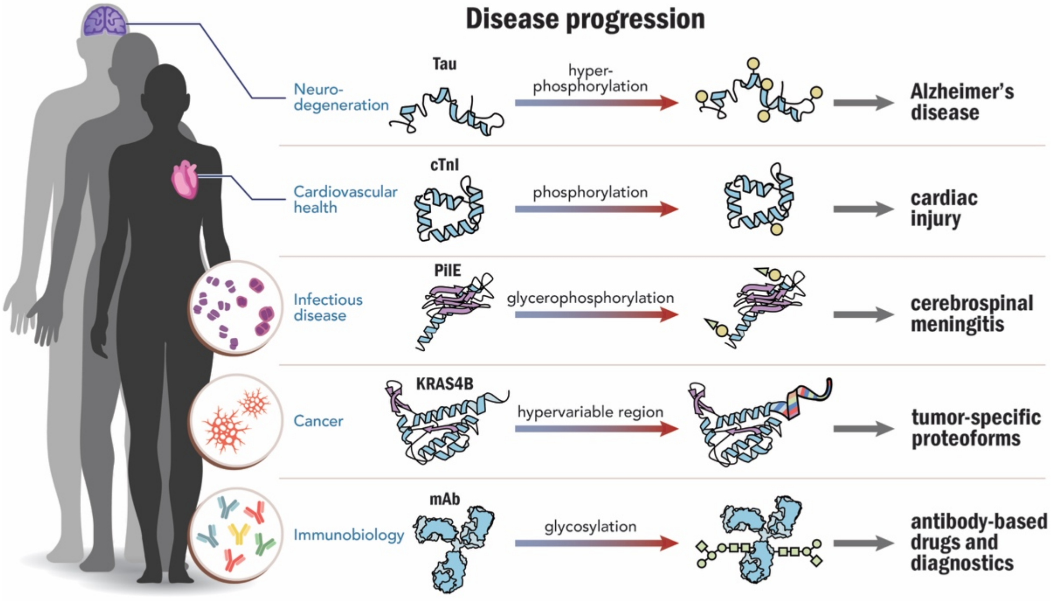

Figure 2. Proteoforms in human disease. Five important disease areas are depicted. Selected examples where proteoforms have been identified and linked to the progression of human disease are discussed in **Section 2**.

international effort on the scale of the HGP in both funding and time will reveal for humanity the presently hidden aspects of the proteoforms within the human proteome. It will make far more accessible the chemical and functional complexity of our proteins, drive the frontiers of research and medicine well beyond what is currently possible, and be critical in the assignment of function to proteins and their PTMs in the years ahead.

Background, Technologies and Project Scope. Proteomics technologies today fall into three major categories: traditional protein electrophoresis methodologies, bottom-up proteomics, and top-down proteomics (**Figure 3**). Each of these has strengths and limitations: protein electrophoresis provides a good overview of the proteins present, but does not provide molecular definition. Bottom-up proteomics allows deep coverage of the proteins present, but is not able to yield proteoform-level knowledge. Top-down proteomics is the most powerful proteoform-level analysis technology in existence, filling gaps in our knowledge regarding RNA isoform translation and combinatorial patterns of protein variation, but is limited in depth and throughput [4, 6].

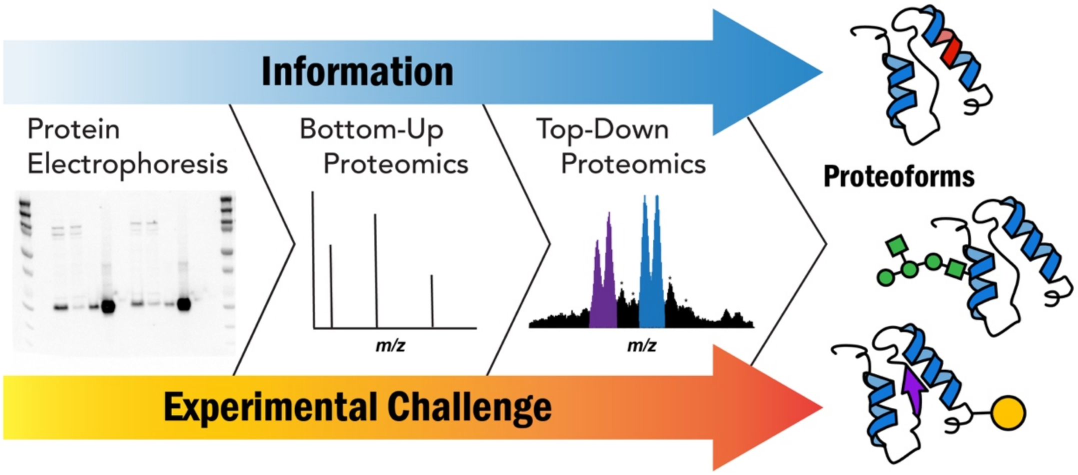

Figure 3. Three analytical techniques used in the analysis of protein primary structure, which includes knowledge of the position and occupancy of post-translational modifications (PTMs).

The dominant bottom-up paradigm of proteomics sacrifices information on proteoforms by cleaving proteins into peptides; this is done for a pragmatic reason - it works, as the resultant peptides are much easier to identify than their parent intact proteoforms. Top-down Proteomics, in contrast, analyzes the entire intact proteoform. At present, our ability to identify and quantify whole proteoforms is in its early stage of development and is poised to mature rapidly (*cf.* **Section 3.1**). Eventually it will also be critical to know proteoform interaction partners and where and when proteoforms are expressed in human systems; however such activities are ancillary to the main scope of this project aimed at the protein-level analog of the Human Genome Project – i.e., 'sequencing' the human proteome as defined in **Box 1**.

2. The Human Proteoform Project: Disease-based Discovery of Proteoforms and Their Modifications

Detailed knowledge of proteoforms is fundamental to both basic and translational biomedical research. The functional roles of proteoforms cannot be gleaned until their existence is known. There are now dozens of examples of the power and importance of proteoform-level knowledge [4], along with emergent platforms and instrumentation increasingly able to discover and characterize proteoforms [7]. The proposed two-pronged structure of the Human Proteoform Project includes both discovery and characterization of proteoform diversity in specific disease areas (targeted analysis), and comprehensive discovery and characterization of human proteomes at the proteoform level (global analysis). Select examples of the importance of targeted analysis are summarized below in the context of five disease areas (see **Figure 2**), and **Section 3** presents the path forward for global analysis.

2.1 Infectious Disease

In the field of infectious diseases, the study of proteoforms has followed two main axes. The first axis took advantage of the power of proteoform-level knowledge to reveal essential PTMs involved in bacterial pathogenesis. For the causative pathogen in cerebrospinal meningitis (*N. meningitidis*), proteoforms of the PilE protein carrying phosphoglycerol groups on serines were tightly associated with crossing of the epithelial barrier and access to the blood stream [8]. PilE is the major component of Type IV pili, which are filamentous organelles protruding from the bacterial membrane and major virulence factors for many Gram-negative pathogens. For the same protein, proteoforms containing multiple glycerophosphorylations were observed in meningitis patients and linked to immune escape [9]. Another study determined that rare O-mycoloylations on several proteins were critical for their localization to the outer membrane of *C. glutamicum* [10] and other unusual glycosylations on rare strains have also been observed by top-down proteomics [11]. Another example of proteoform detection leading to the identification of critical PTMs was reported for *Salmonella typhimurium*, the most common foodborne pathogen; specific and labile S-cysteinylated proteoforms were found in response to infection-like conditions [12].

The second major axis has used proteoforms as the unit of measurement in the routine identification of bacterial pathogens in clinical laboratories. In the past ten years, proteoform-level analysis has been successfully deployed in thousands of hospitals worldwide [13], achieving major impact via robust commercialization based on measurement of ribosomal and other abundant proteoforms [14]. This dramatic clinical impact underscores the impact that large-scale discovery of bacterial proteoforms can bring to basic and clinical microbiology. Similarly impactful clinical assays are under development in mammalian systems, in areas such as hemoglobinopathies and neurodegenerative disease (see below).

2.2 Neurodegenerative Disease

The global estimate of people currently suffering from dementia (mainly Alzheimer's disease) is in excess of 47 million, and is expected to increase to 135 million by 2050 [15]. Despite proteins being at the root of neurodegenerative diseases—even to the extent that they are referred to clinically as "proteinopathies" [16] —we still have not taken an accurate census of the molecular players involved. Aberrant modifications of one or more proteins and associated protein aggregation appear to play an important role in neurological disorders such as Alzheimer's and Parkinson's diseases [17]. Six of the most highly-cited proteins in the scientific literature (**Figure 4**) are either associated with neurodegenerative disease (tau, APOE, SOD1) or are involved in the inflammatory cascade occurring during disease progression (TNF, CASP, CytC). For example, tau phosphorylated at threonine 181 (P-tau181), the 42 amino acid proteoform of alpha β-amyloid (Aβ), and APOE polymorphisms are used to support clinical Alzheimer's diagnosis, and specific apolipoprotein A-I (Apo A-I) proteoforms have diagnostic potential [15]. Phosphorylation of tau is linked to amyloidogenic propensity, but the structural details are not well understood because of the lack of characterization of complex tau phosphoproteoforms. The role of amyloid-β and α-synuclein proteoforms in the pathogenesis of Alzheimer's and Parkinson's diseases remains puzzling but has led to excitement recently with the success of Biogen's amyloid aggregate-targeting Alzheimer's therapy, aducanumab, in clinical trials [18].

Dysregulated PTMs may influence the propensity for protein aggregation in neurodegenerative disease. Numerous PTMs have been identified, that, in isolation or in combination, act as modulators of proteinopathy in neurodegenerative diseases: isoaspartate formation in amyloid-β, phosphorylation of amyloid-β or tau in Alzheimer's Disease; acetylation, 4-hydroxy-2-neonal modification, O-GlcNAcylation or phosphorylation of α-synuclein in Parkinson's Disease; acetylation or phosphorylation of TAR DNA-binding protein-43 in Amyotrophic lateral sclerosis (ALS), SUMOylation of Superoxide Dismutase-1 (SOD1) in ALS; and phosphorylation of huntingtin in Huntington's Disease [19]. Studies with SOD1 highlight the utility of top-down proteomics; whereas bottom-up studies from *in vitro* modified SOD1 characterized tens of modifications, top-down studies of ALS-patient immunopurified SOD1 revealed there were only three prevalent SOD1 proteoforms, one of which had a neurotoxic function, one of which had a neuroprotective function—and none of which were inferred from the bottom-up studies [20]. These studies illustrate how top-down proteomics is able to unravel the relationships among sequence variants, PTMs, and protein complexes involved in proteinopathies, information that cannot be obtained from bottom-up proteomics alone. This knowledge is essential to reveal the mysteries and mechanisms underlying these devastating diseases, enabling informed design and discovery of novel diagnostic and therapeutic strategies for neurodegenerative disease [21].

2.3 Cancer

The biology of cancer has long been understood to involve proteins and their PTMs, particularly in signaling pathways involving intracellular phosphorylation. The flagship efforts of the Cancer Proteomics Consortium, CPTAC, over the years have brought targeted proteomics and proteogenomics into regular use and resulted in publication of major studies on ovarian [22], breast [23] and colorectal cancer [24]. Using the bottom-up approach to proteomics, CPTAC noted in 2016 that "*the aggregated NCI-60 proteomics data set covers only 12% of the whole encoded proteome, and only ~5% of the genes had sequence coverage >50% of their protein coding regions."* [25]. Regarding alternative splicing, "*There is yet a major gap between the number of alternative transcripts asserted by RNA-seq and that detectable by proteomics (e.g., <0.1% of putative novel splice junctions in cancer xenografts).*" [25].

Because it measures the entire proteoform, top-down proteomics closes these gaps by determining the oncoproteoforms derived from combinations of driving mutations, RNA splice variants, and PTMs. This has been demonstrated in the area of RAS biology. RAS family genes encode small GTPases, mutations of which drive over 40% of all cancers and over 90% of pancreatic tumors. Three RAS genes give rise to four isoforms (KRAS4A, KRAS4B, HRAS, and NRAS), which share high sequence identity within their first 165 residues and whose PTMs can be distinguished precisely using immunoprecipitation with top-down mass spectrometry (IP-TDMS)

[26]. The complete molecular definition and abundance of *wild-type* vs. mutant RAS proteoforms was revealed in proteoform-level studies of cancer cell lines and tumors, providing combinations and occupancies for diverse PTMs that are inaccessible by bottom-up approaches. This illustrates how proteoform-resolved measurements can drive the detection of PTMs and help assign their function in both basic and translational research. The expanded use of targeted IP-TDMS offers the possibility of deep proteoform-level knowledge of signaling cascades, which will be transformative to the field of cancer research.

2.4 Proteoforms in Cardiovascular Disease

Cardiovascular diseases (CVD) are the leading cause of death globally and the afflicted population is expected to increase as the population ages [27]. There has been tremendous progress in the application of proteomics in CVD medicine [28]. Altered cardiac proteoforms have been linked to disease phenotypes in heart failure using both animal models and human clinical samples [29]. Phosphorylated proteoforms of cardiac troponin I (cTnI), a gold standard biomarker for detection of cardiac injury, were identified by top-down quantitative proteomics to be candidate biomarkers for chronic heart failure [30]. Functionally significant site-specific phosphorylation alterations of cTnI have been identified in human heart failure [31]. Top-down proteomics has also identified actin proteoforms as potential cardiac disease markers [32] and novel phosphorylation of a critical Z-disc protein, enigma homolog isoform 2, in acute myocardial infarction [33].

As PTMs and alternative splicing are critically important in the development of CVD, dissection of proteoforms and their interactions is essential to improve our understanding of its molecular mechanisms. Various PTMs of both structural and signaling proteins in the cell create proteoforms that regulate function in organelles such as mitochondria in currently unknown ways as they respond to stress and aging. Top-down proteomics will play a critical role in characterization of stem cell-derived cardiomyocytes for regenerative medicine [34]. This will provide new opportunities for uncovering proteoform complexity underlying central mechanistic details involved in cardiac development and disease, contributing toward the use of patient-specific cell therapy for precision medicine [34, 35].

2.5 Immunobiology and Antibody Proteoforms

Much of the adoption, and therefore the impact, of measuring whole proteoforms has been in the biopharmaceutical industry. Protein-based drugs represent an increasing proportion of pharmaceutical sales, comprising roughly half of current drug development pipelines and new FDA approvals in the U.S. from 2014-2018 [36]. Such biotherapeutics are largely comprised of monoclonal antibodies (mAbs), antibody-drug conjugates (ADCs), and fusion proteins that have been developed as novel therapies for a wide variety of clinical indications, including cancers, autoimmunity/inflammation, and genetic disorders. Antibodies are critical to clinical diagnostics and academic research alike. Top-down MS is currently widely used to provide comprehensive structural characterization of antibody-based drugs, particularly those carrying a low-to-medium complexity of glycoforms [37-39] and other types of engineered or natural PTMs.

Intact protein mass measurements (a facet of top-down MS) are being integrated into a widening number of contexts in the industrial pipeline. This industry will continue to expand, but would benefit greatly from accelerated development of proteoform discovery and characterization technologies. Exciting new areas such as personalized cellular immunotherapies (e.g., CART cells), gene therapies (e.g., adenovirus-mediated gene delivery), and RNA-vaccines like that for SARS-CoV-2 will benefit significantly from accelerated proteoform-level technologies by providing critical knowledge of the structures of the key proteoforms involved. An exciting extension of this area combines mass spectrometry with nucleic acid sequencing to characterize the repertoire of endogenous antibodies generated in response to a pathogen [40-47] or to gauge immune response to an organ transplant [48]. The exceptional selectivity and specificity of mass spectrometry in general and of top-down proteomics in particular, will play a growing role in immunobiology and can even be envisaged to deal with the complexity of the innate and adaptive immune responses within the next ten years.

2.6 Proteoform-level Knowledge is Essential to Understand Biological Function

The examples above from five important disease areas illustrate the critical role of proteoform-level knowledge in disease and health. This message is further underscored by consideration of the most highly cited work in the biomedical literature. **Figure 4** shows the top human gene targets ranked by their number of citations. Tumor necrosis factor (TNF), at the top of the list, has 206,000 citations; given such a level of resources expended on study of even a single gene, it is clearly essential moving forward to underpin biomedical research with definitive proteoform-level information. Notably, even the most-studied proteins have proteoforms that have not yet been revealed and are essential to understanding of their biological or disease-related functions. The economies of scale afforded by a concerted project to obtain comprehensive

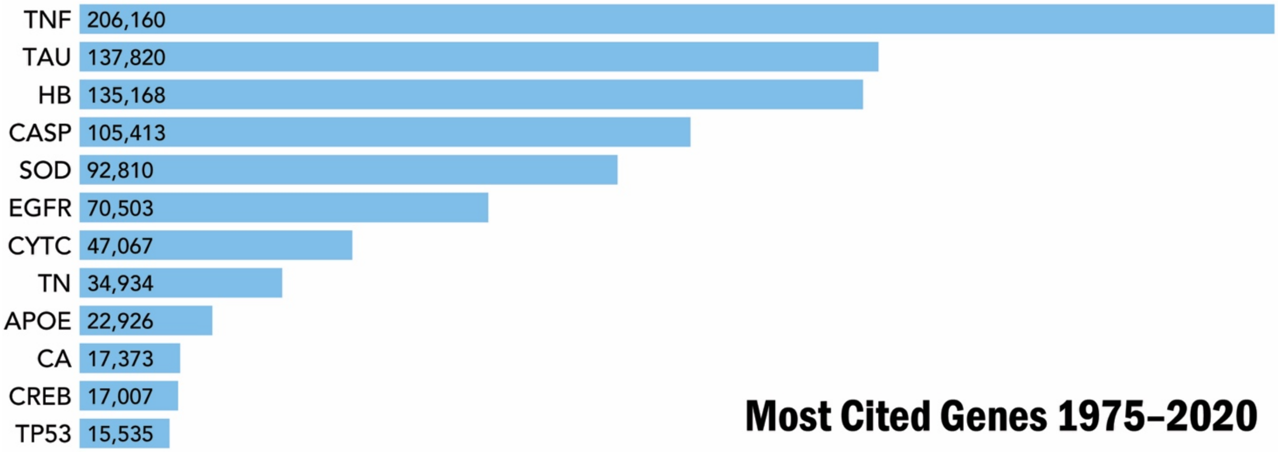

Figure 4. The most studied proteins have essential proteoforms that contain common post-translational modifications such as phosphorylation, methylation, acetylation, as well as other important variations of primary structure such as disulfide bond formation, metal attachment, and proteolytic processing. Protein abbreviations: TNF, tumor necrosis factor; TAU, tubulin-associated unit; HB, hemoglobin subunits (HbA, HbB, *etc.*); CASP, cysteine-aspartic proteases (Casp1-9); SOD, superoxide dismutases (SOD-1,-2,-3); EGFR, estrogen growth factor receptor; CYTC, cytochrome c; TN, troponin (Tn-C,I,T); APOE, apolipoprotein E; CA, carbonic anhydrase; CREB, cyclic-AMP response element-binding protein; TP53, cellular tumor antigen p53. This figure is adapted from a 2017 feature in *Nature* [49], and has been updated using the Web of Science Core Collection (1975-2020).

proteoform-level knowledge will make possible the acquisition of such information for the 20,000 proteoform families derived from the human genome. Just as the $1/base estimate for the Human Genome Project provided an important target to spur technology competition and development, so will a $1/proteoform goal for the Human Proteoform Project as enunciated previously for 1 billion proteoforms [50].

Detailed proteoform knowledge is critical to illuminating the functional roles of gene products and realizing the vision of the Human Genome Project, as clearly articulated by F. Collins *et al.* in 2003, "*A critical step towards gaining a complete understanding . . . will be to take an accurate census of the proteins present in particular cell types. It will be a major challenge to catalogue proteins present in low abundance or in membranes. Determining the absolute abundance of each protein, including all modified forms, will be an important next step.*" [2] The Human Proteoform Project we present here is the critical next step in the quest to understand human health and disease.

3. The Human Proteoform Project: Assembling the Proteoform Atlas

We propose the Human Proteoform Project, an ambitious program to aggressively develop new technologies for comprehensive proteoform analysis, and to generate a definitive reference set of human proteoforms. We envision next generation proteomics in humans to be based upon ~20,000 proteoform families [6], one for each gene in the genome. Deep catalogs of proteoforms compiled for important model organisms and widely characterized mammalian cell lines will dramatically accelerate proteoform level studies. This more profound knowledge of the central molecules of biology will provide an essential cornerstone for 21st century biology. New technologies will be central to this effort as today's ability to comprehensively identify proteoforms in complex systems is limited.

3.1 New Technologies

Today's proteoform-level analysis is almost solely based upon top-down mass spectrometry [51, 52]. To achieve the objectives outlined above, it is critical to expand our technological abilities through a concerted long-term and multi-faceted research and development effort. This effort should pursue both the continued development of mass spectrometry-based technologies for proteoform analysis, as well as the exploration of potential paradigm-shifting new ideas and approaches which offer the possibility of transformative change. The development of increasingly powerful and effective nucleic acid sequencing has demonstrated the importance of investing heavily in ambitious new efforts to drive technology development. Similarly, single molecule mass spectrometry [7, 53, 54], nanopore sequencing [55, 56], cryoelectron microscopy and visual proteomics [57, 58], single cell proteomics [59-63], single molecule protein arrays [64], and other ideas yet to be conceived need to be encouraged, supported, and developed to advance proteoform biology. The outstanding success of the technology development program in the Human Genome Project and the associated private sector engagement provide an inspiring model for how this can be done well.

3.2 The Human Proteoform Atlas

In this Human Proteoform Project, it is essential to recognize that proteoform expression varies across cells and tissues. Studies of proteoform expression can be either global or targeted (**Figure 5**). In global studies all proteoforms present at detectable levels are characterized; in targeted studies, specific proteoform families could be enriched using, for example, protein affinity capture reagents, in order to reveal the deeper levels of proteoform diversity present. The development of affinity reagents to capture the proteins encoded by each human gene will be invaluable to enrich and then characterize their proteoform families in a selection of human specimens. An important thrust of the project is the delineation of proteoform expression patterns in human cell types [50]. This work will be greatly empowered by the ongoing HuBMAP [65] and Human Cell Atlas projects [66], and affiliated consortia, which seek to define all of the human cell types. The depth of proteoform analysis possible is dependent upon the technologies employed; achieving detection limits of ~25 copies per cell [50] is dependent upon technology advances, underscoring the importance of aggressive technology investment.

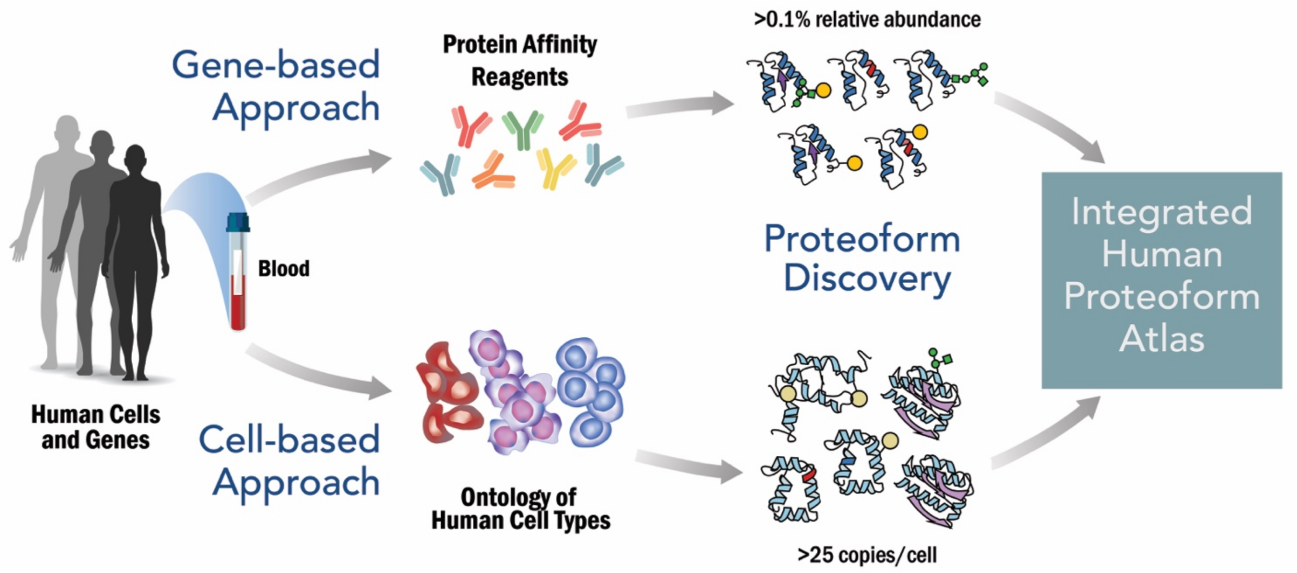

Figure 5. Strategy to create an integrated Human Proteoform Atlas. The upper approach illustrates the use of protein affinity reagents to capture proteoform families derived from targeted genes. The lower approach illustrates the in-depth analysis of human cell types for proteoform discovery and characterization. Relative abundance refers to the ratio of a given proteoform to the sum of all proteoforms in that family.

A central principle in comprehensive proteoform analysis concerns the distinction between discovery and scoring. Detailed analysis of protein primary structure requires the generation of highly complex data, necessary in the discovery phase of proteoform analysis. However, once we have in hand a comprehensive index of such proteoforms for the system under study, efforts can shift to a scoring mode, informed by the previous knowledge. This transition from discovery to scoring is central to many fields: in genomics, for example, the initial discovery of single nucleotide polymorphisms (SNPs) led to the generation of SNP databases and technologies for their scoring at scale. The scoring technology enabled cost-effective functional studies and disease-based research across human populations. Similarly, in mass spectrometry, initial work to develop small molecule identification from gas-phase fragmentation patterns led to the establishment of rich databases of molecular fragmentation spectra allowing the rapid identification of already known compounds. This venerable principle will also be invaluable to driving increased throughput and decreased cost in the Human Proteoform Project.

Deep knowledge of the proteoforms present in human and model organisms, with the invaluable associated information as to their PTMs along with their parent genetic variants and the RNA isoforms from which they were transcribed and translated, will empower studies to directly probe the expression of these key molecules and their roles in health and disease. Beyond improving the use of proteins as biomarkers, it will open the study of their temporal and spatial distributions within cells and tissues, information which is presently impossible to obtain. This process will often involve protein affinity capture reagents enabling diverse types of readouts (**Figure 6**), enhanced and made more specific based upon the detailed knowledge afforded by the comprehensive proteoform index. These new tools will make it possible for studies in which cells, tissues, or whole animals are subjected to perturbations, such as mutations, disease, viral infection, or drug treatment, to yield far richer and more detailed knowledge than has previously been possible.

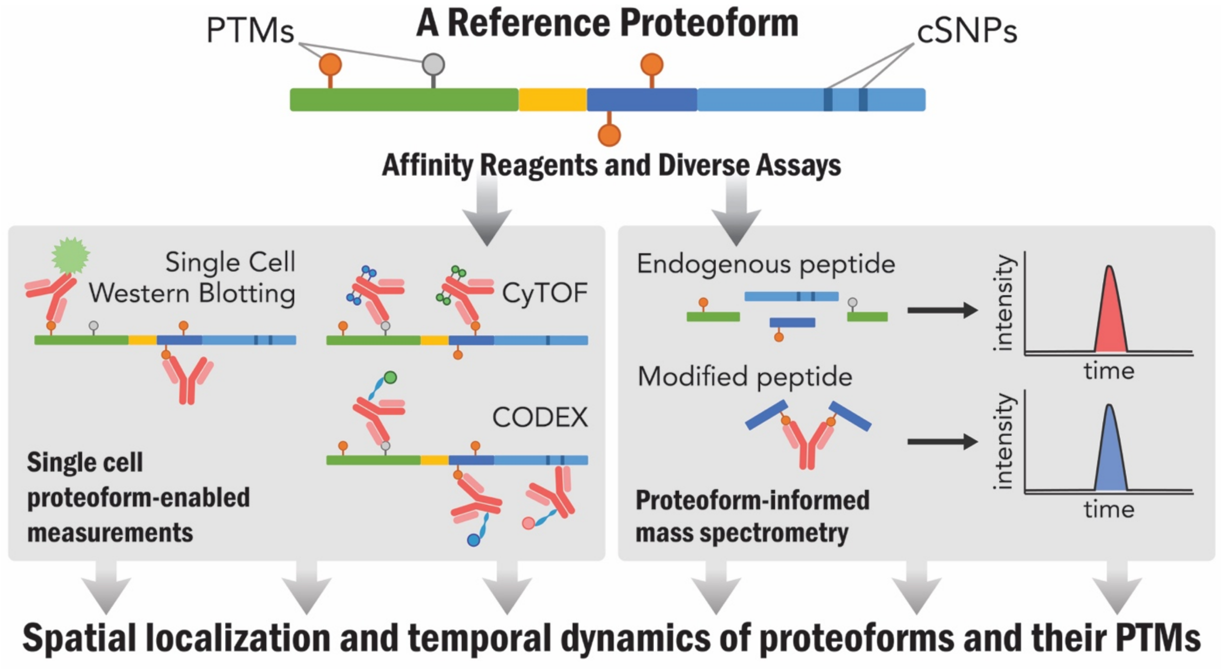
PTMs
A Reference Proteoform
cSNPs
Affinity Reagents and Diverse Assays
Single Cell Western Blotting
CyTOF
CODEX
Single cell proteoform-enabled measurements
Endogenous peptide
Modified peptide
intensity
time
intensity
time
Proteoform-informed mass spectrometry
Spatial localization and temporal dynamics of proteoforms and their PTMs

Figure 6. Once proteoforms have been identified, affinity reagents and targeted assays enable powerful new strategies to enable delineation of spatial distribution and temporal dynamics of proteoforms and their PTMs; CyTOF, mass cytometry; CODEX, CO-Detection by indEXing.

3.3 Role of Government, Foundations and the Private Sector

For the necessary transformation of technology and knowledge to take place over the coming decade, numerous stakeholders will be needed to engage and align with the Project to bring it to fruition [67]. Within the emergent proteomics ecosystem that we envisage, three categories of organizations can be identified – those focused on creating new knowledge (universities and research institutes), those creating new value for customers (instrument, biopharma, and diagnostics companies), and those providing financial and other resource support for the creation of knowledge or customer value (government agencies, philanthropies, non-profit foundations, and well-funded companies) [68]. The role of the knowledge creators is paramount for a research-intensive area like this, and the major universities and research institutes will generate the structural, large-scale data to drive this effort. This will require substantial funding; for comparison, genomics research worldwide was publicly funded at about $3B per year from 2003 to 2006, with the U.S. contributing about 35% of this [69].

The companies and institutions that commercialize the tools, technologies, and services to advance the field also play a pivotal role in this endeavor, often collaborating with academic researchers to bring new technologies to the marketplace. This cycle of innovation and commercialization was a fundamental enabler of the Human Genome Project. The biopharmaceutical and diagnostic companies invest heavily in research and development (for example, having spent $97 billion in R&D in the U.S. in 2017 [70]), and so are well poised to participate in such efforts. As noted above, the definitive proteoform set for the expressed human proteome will constitute a major economic opportunity for the private sector.

Bringing alignment and finding common goals for the various members of the emerging 'proteoform ecosystem' is already underway, with organizations starting to forge bridges across the boundaries. Increasing the cooperation between the public agencies, organizations and international institutions will hasten the discovery and understanding of human proteoforms, and provide dramatic growth in therapeutics, diagnostics, and the life sciences.

4 Conclusion

The Human Proteoform Project will revolutionize our understanding of human health and disease. This ambitious project to develop and apply powerful new technologies in order to reveal the molecular complexity that underlies human biology will be transformative. Detailed proteoform-level knowledge will drive the understanding of fundamental biology while also opening new frontiers in human diagnostic and therapeutic strategies. We hope that the roadmap presented here will serve to inspire investment in its realization by public and private entities.

Acknowledgements:

YG was supported by the National Institutes of Health (NIH) under awards R01 HL096971, GM117058, and GM125085. NLK was supported by the Paul G. Allen Frontiers Program under award 11715, the National Institute of General Medical Sciences (NIGMS) at NIH under award P41GM1085698 and the NIH Common Fund under award UH3CA246635. JAL was supported by the NIH under award R01GM103479 and the Department of Energy (DOE) under award DE-FC02-02ER63421. LMS was supported by the NIGMS under award R35GM126914. LPT was supported under the Intramural program at the Environmental Molecular Sciences Laboratory (a DOE Office of Science User Facility sponsored by the Office of Biological and Environmental Research and operated under contract DE-AC05-76RL01830) and the NIH Common Fund under award UG3CA256959. JCR and YOT were supported by the European Horizon 2020 program under award 829157, and JCR was also supported by the Institut Pasteur, the CNRS and EPIC-XS under award 823839. The authors acknowledge Michael W. Mullowney for assistance in generating figures.

Conflicts of Interest:

The authors declare the following competing financial interest(s): YT is an employee of Spectroswiss, which develops and commercializes FTMS data processing and data analysis software. NLK is involved with commercialization of hardware and software for proteoform analysis. PD is the Founder of Eastwoods Consulting, providing business advisory services to life science companies.

Background on the Consortium for Top Down Proteomics (CTDP):

The Consortium was formed on March 3, 2012 and is a non-profit 501(c)3 organization. More information on current members can be found at http://www.topdownproteomics.org/. The CTDP mission is to promote innovative research, collaboration and education to accelerate the comprehensive analysis of intact proteoforms and their complexes in diverse biological systems. The Consortium for Top-Down Proteomics gratefully acknowledges the support of its corporate sponsors ThermoFisher Scientific Inc., Bruker Corporation, Waters Corporation, Pfizer Inc., and New Objective Inc.

References:

1. Dranke, N., *What is the Human Genome Worth?* Nature, 2011: https://doi.org/10.1038/news.2011.281.
2. Collins, F.S., et al., *A vision for the future of genomics research.* Nature, 2003. **422**(6934): p. 835-47.
3. Smith, L.M., N.L. Kelleher, and P. Consortium for Top Down, *Proteoform: a single term describing protein complexity.* Nat Methods, 2013. **10**(3): p. 186-7.
4. Aebersold, R., et al., *How many human proteoforms are there?* Nat Chem Biol, 2018. **14**(3): p. 206-214.
5. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.
6. Smith, L.M. and N.L. Kelleher, *Proteoforms as the next proteomics currency.* Science, 2018. **359**(6380): p. 1106-1107.
7. Kafader, J.O., et al., *Multiplexed mass spectrometry of individual ions improves measurement of proteoforms and their complexes.* Nat Methods, 2020. **17**(4): p. 391-394.

8. Chamot-Rooke, J., et al., *Posttranslational modification of pili upon cell contact triggers N. meningitidis dissemination.* Science, 2011. **331**(6018): p. 778-82.
9. Gault, J., et al., *Neisseria meningitidis Type IV Pili Composed of Sequence Invariable Pilins Are Masked by Multisite Glycosylation.* PLoS Pathog, 2015. **11**(9): p. e1005162.
10. Carel, C., et al., *Identification of specific posttranslational O-mycoloylations mediating protein targeting to the mycomembrane.* Proc Natl Acad Sci U S A, 2017. **114**(16): p. 4231-4236.
11. Schirm, M., et al., *Identification of unusual bacterial glycosylation by tandem mass spectrometry analyses of intact proteins.* Anal Chem, 2005. **77**(23): p. 7774-82.
12. Ansong, C., et al., *Top-down proteomics reveals a unique protein S-thiolation switch in Salmonella Typhimurium in response to infection-like conditions.* Proc Natl Acad Sci U S A, 2013. **110**(25): p. 10153-8.
13. Nassif, X., *A revolution in the identification of pathogens in clinical laboratories.* Clin Infect Dis, 2009. **49**(4): p. 552-3.
14. Levesque, S., et al., *A Side by Side Comparison of Bruker Biotyper and VITEK MS: Utility of MALDI-TOF MS Technology for Microorganism Identification in a Public Health Reference Laboratory.* PLoS One, 2015. **10**(12): p. e0144878.
15. Shrivastava, S.R., P.S. Shrivastava, and J. Ramasamy, *Dementia in middle- and low-income nations: A public health priority.* J Res Med Sci, 2016. **21**: p. 5.
16. Bayer, T.A., *Proteinopathies, a core concept for understanding and ultimately treating degenerative disorders?* Eur Neuropsychopharmacol, 2015. **25**(5): p. 713-24.
17. Boland, B., et al., *Promoting the clearance of neurotoxic proteins in neurodegenerative disorders of ageing.* Nat Rev Drug Discov, 2018. **17**(9): p. 660-688.
18. Schneider, L., *A resurrection of aducanumab for Alzheimer's disease.* Lancet Neurol, 2020. **19**(2): p. 111-112.
19. Schaffert, L.N. and W.G. Carter, *Do Post-Translational Modifications Influence Protein Aggregation in Neurodegenerative Diseases: A Systematic Review.* Brain Sci, 2020. **10**(4): p. 232.
20. Schmitt, N.D. and J.N. Agar, *Parsing disease-relevant protein modifications from epiphenomena: perspective on the structural basis of SOD1-mediated ALS.* J Mass Spectrom, 2017. **52**(7): p. 480-491.
21. Nshanian, M., et al., *Native Top-Down Mass Spectrometry and Ion Mobility Spectrometry of the Interaction of Tau Protein with a Molecular Tweezer Assembly Modulator.* J Am Soc Mass Spectrom, 2019. **30**(1): p. 16-23.
22. Zhang, H., et al., *Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer.* Cell, 2016. **166**(3): p. 755-765.
23. Mertins, P., et al., *Proteogenomics connects somatic mutations to signalling in breast cancer.* Nature, 2016. **534**(7605): p. 55-62.
24. Vasaikar, S., et al., *Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities.* Cell, 2019. **177**(4): p. 1035-1049.
25. Ruggles, K.V., et al., *An Analysis of the Sensitivity of Proteogenomic Mapping of Somatic Mutations and Novel Splicing Events in Cancer.* Mol Cell Proteomics, 2016. **15**(3): p. 1060-71.
26. Ntai, I., et al., *Precise characterization of KRAS4b proteoforms in human colorectal cells and tumors reveals mutation/modification cross-talk.* Proc Natl Acad Sci U S A, 2018. **115**(16): p. 4140-4145.

27. Mensah, G.A., G.A. Roth, and V. Fuster, *The Global Burden of Cardiovascular Diseases and Risk Factors: 2020 and Beyond.* J Am Coll Cardiol, 2019. **74**(20): p. 2529-2532.

28. Lam, M.P., P. Ping, and E. Murphy, *Proteomics Research in Cardiovascular Medicine and Biomarker Discovery.* J Am Coll Cardiol, 2016. **68**(25): p. 2819-2830.

29. Cai, W., et al., *Top-down Proteomics: Technology Advancements and Applications to Heart Diseases.* Expert Rev Proteomics, 2016. **13**(8): p. 717-30.

30. Zhang, J., et al., *Top-down quantitative proteomics identified phosphorylation of cardiac troponin I as a candidate biomarker for chronic heart failure.* J Proteome Res, 2011. **10**(9): p. 4054-65.

31. Zhang, P., et al., *Multiple reaction monitoring to identify site-specific troponin I phosphorylated residues in the failing human heart.* Circulation, 2012. **126**(15): p. 1828-37.

32. Chen, Y.C., et al., *Effective top-down LC/MS+ method for assessing actin isoforms as a potential cardiac disease marker.* Anal Chem, 2015. **87**(16): p. 8399-8406.

33. Peng, Y., et al., *Top-down proteomics reveals concerted reductions in myofilament and Z-disc protein phosphorylation after acute myocardial infarction.* Mol Cell Proteomics, 2014. **13**(10): p. 2752-64.

34. Cai, W., et al., *An Unbiased Proteomics Method to Assess the Maturation of Human Pluripotent Stem Cell-Derived Cardiomyocytes.* Circ Res, 2019. **125**(11): p. 936-953.

35. Chen, I.Y., E. Matsa, and J.C. Wu, *Induced pluripotent stem cells: at the heart of cardiovascular precision medicine.* Nat Rev Cardiol, 2016. **13**(6): p. 333-49.

36. Walsh, G., *Biopharmaceutical benchmarks 2018.* Nat Biotechnol, 2018. **36**(12): p. 1136-1145.

37. Fornelli, L., et al., *Middle-down analysis of monoclonal antibodies with electron transfer dissociation orbitrap fourier transform mass spectrometry.* Anal Chem, 2014. **86**(6): p. 3005-12.

38. Fornelli, L., et al., *Top-down analysis of immunoglobulin G isotypes 1 and 2 with electron transfer dissociation on a high-field Orbitrap mass spectrometer.* J Proteomics, 2017. **159**: p. 67-76.

39. Mao, Y., et al., *Top-down structural analysis of an intact monoclonal antibody by electron capture dissociation-Fourier transform ion cyclotron resonance-mass spectrometry.* Anal Chem, 2013. **85**(9): p. 4239-46.

40. Cheung, W.C., et al., *A proteomics approach for the identification and cloning of monoclonal antibodies from serum.* Nat Biotechnol, 2012. **30**(5): p. 447-52.

41. Srzentic, K., et al., *Multiplexed Middle-Down Mass Spectrometry as a Method for Revealing Light and Heavy Chain Connectivity in a Monoclonal Antibody.* Anal Chem, 2018. **90**(21): p. 12527-12535.

42. Wine, Y., et al., *Molecular deconvolution of the monoclonal antibodies that comprise the polyclonal serum response.* Proc Natl Acad Sci U S A, 2013. **110**(8): p. 2993-8.

43. Cha, S.W., et al., *The Antibody Repertoire of Colorectal Cancer.* Mol Cell Proteomics, 2017. **16**(12): p. 2111-2124.

44. Safonova, Y. and P.A. Pevzner, *De novo Inference of Diversity Genes and Analysis of Non-canonical V(DD)J Recombination in Immunoglobulins.* Front Immunol, 2019. **10**: p. 987.

45. Georgiou, G., et al., *The promise and challenge of high-throughput sequencing of the antibody repertoire.* Nat Biotechnol, 2014. **32**(2): p. 158-68.

46. Lee, J., et al., *Molecular-level analysis of the serum antibody repertoire in young adults before and after seasonal influenza vaccination.* Nat Med, 2016. **22**(12): p. 1456-1464.

47. Wang, Z., et al., *Top-down Mass Spectrometry Analysis of Human Serum Autoantibody Antigen-Binding Fragments.* Sci Rep, 2019. **9**(1): p. 2345.
48. Toby, T.K., et al., *Proteoforms in Peripheral Blood Mononuclear Cells as Novel Rejection Biomarkers in Liver Transplant Recipients.* Am J Transplant, 2017. **17**(9): p. 2458-2467.
49. Dolgin, E., *The most popular genes in the human genome.* Nature, 2017. **551**(7681): p. 427-431.
50. Kelleher, N.L., *A cell-based approach to the human proteome project.* J Am Soc Mass Spectrom, 2012. **23**(10): p. 1617-24.
51. LeDuc, R.D., et al., *ProForma: A Standard Proteoform Notation.* J Proteome Res, 2018. **17**(3): p. 1321-1325.
52. Smith, L.M., et al., *A five-level classification system for proteoform identifications.* Nat Methods, 2019. **16**(10): p. 939-940.
53. Dominguez-Medina, S., et al., *Neutral mass spectrometry of virus capsids above 100 megadaltons with nanomechanical resonators.* Science, 2018. **362**(6417): p. 918-922.
54. Kafader, J.O., et al., *Measurement of Individual Ions Sharply Increases the Resolution of Orbitrap Mass Spectra of Proteins.* Anal Chem, 2019. **91**(4): p. 2776-2783.
55. Ouldali, H., et al., *Electrical recognition of the twenty proteinogenic amino acids using an aerolysin nanopore.* Nat Biotechnol, 2020. **38**(2): p. 176-181.
56. Restrepo-Perez, L., C. Joo, and C. Dekker, *Paving the way to single-molecule protein sequencing.* Nat Nanotechnol, 2018. **13**(9): p. 786-796.
57. Beck, M., et al., *Visual proteomics of the human pathogen Leptospira interrogans.* Nat Methods, 2009. **6**(11): p. 817-23.
58. Xu, M., et al., *De Novo Structural Pattern Mining in Cellular Electron Cryotomograms.* Structure, 2019. **27**(4): p. 679-691 e14.
59. Specht, H. and N. Slavov, *Transformative Opportunities for Single-Cell Proteomics.* J Proteome Res, 2018. **17**(8): p. 2565-2571.
60. Budnik, B., et al., *SCoPE-MS: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation.* Genome Biol, 2018. **19**(1): p. 161.
61. Slavov, N., *Single-cell protein analysis by mass spectrometry.* Curr Opin Chem Biol, 2020. **60**: p. 1-9.
62. Zhou, M., et al., *Sensitive Top-Down Proteomics Analysis of a Low Number of Mammalian Cells Using a Nanodroplet Sample Processing Platform.* Anal Chem, 2020. **92**(10): p. 7087-7095.
63. Zhu, Y., et al., *Proteomic Analysis of Single Mammalian Cells Enabled by Microfluidic Nanodroplet Sample Preparation and Ultrasensitive NanoLC-MS.* Angew Chem Int Ed Engl, 2018. **57**(38): p. 12370-12374.
64. Swaminathan, J., et al., *Highly parallel single-molecule identification of proteins in zeptomole-scale mixtures.* Nat Biotechnol, 2018. **36**, p. 1076–1082.
65. Consortium, H., *The human body at cellular resolution: the NIH Human Biomolecular Atlas Program.* Nature, 2019. **574**(7777): p. 187-192.
66. Regev, A., et al., *The Human Cell Atlas.* Elife, 2017. **6**.
67. Adner, R., *Ecosystem as Structure: An Actionable Construct for Strategy.* Journal of Management, 2016. **43**(1): p. 39-58.
68. Valkokari, K., *Business, Innovation, and Knowledge Ecosystems: How They Differ and How to Survive and Thrive within Them.* Technology Innovation Management Review, 2015: p. 17-24.

69. Pohlhaus, J.R. and R.M. Cook-Deegan, *Genomics research: world survey of public funding.* BMC Genomics, 2008. **9**: p. 472.

70. *Biopharmaceutical Industry Profile.* PhRMA - Pharmaceutical Research and Manufacturers of America, 2019.