Published in final edited form as:
Expert Rev Proteomics. 2017 December ; 14(12): 1059–1071. doi:10.1080/14789450.2017.1394189.

Advances in the Chromosome-Centric Human Proteome Project: Looking to the Future

Young-Ki Paik[1], **Gilbert S. Omenn**[2], **Lydie Lane**[3,4], and **Christopher M. Overall**[5]

[1]Yonsei Proteome Research Center and Department of Biochemistry, Yonsei University, Sudaemoon-ku, Seoul, Korea

[2]Department of Computational Medicine & Bioinformatics, University of Michigan, Ann Arbor, MI, USA

[3]Department of Human Protein Sciences, Faculty of Medicine, University of Geneva, Geneva

[4]Switzerland and 4SIB Swiss Institute of Bioinformatics

[5]Centre for Blood Research, Departments of Oral Biological & Medical Sciences, and Biochemistry & Molecular Biology, Faculty of Dentistry, University of British Columbia, Vancouver, Canada

Abstract

Introduction: The mission of the Chromosome-Centric Human Proteome Project (C-HPP), is to map and annotate the entire predicted human protein set (~20,000 proteins) encoded by each chromosome. The initial steps of the project are focused on “missing proteins (MPs)”, which lacked documented evidence for existence at protein level. In addition to remaining 2,579 MPs, we also target those annotated proteins having unknown functions, uPE1 proteins, alternative splice isoforms and post-translational modifications. We also consider how to investigate various protein functions involved in cis-regulatory phenomena, amplicons lncRNAs and smORFs.

Areas covered: We will cover the scope, historic background, progress, challenges and future prospects of C-HPP. This review also addresses the question of how we can best improve the methodological approaches, select the optimal biological samples, and recommend stringent protocols for the identification and characterization of MPs. A new strategy for functional analysis of some of those annotated proteins having unknown function will also be discussed.

Expert commentary: If the project moves well by reshaping the original goals, the current working modules and team work in the proposed extended planning period, it is anticipated that a progressively more detailed draft of an accurate chromosome-based proteome map will become available with functional information.

Correspondence: Young-Ki Paik, Yonsei Proteome Research Center and Department of Biochemistry, Yonsei University, Sudaemoon-ku, Seoul, Korea, paikyk@yonsei.ac.kr.

Declaration of interest
The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

Keywords

Chromosome-Centric Human Proteome Project; Functional annotation; HUPO; Human Protein Atlas; long non-coding RNA; Mass Spectrometry; Missing Proteins; neXtProt; PeptideAtlas; protein evidence; ProteomeXchange; targeted proteomics; smORF; SRMAtlas

1. Scope, goals and and deliverables

The Human Proteome Project (HPP) of the Human Proteome Organization (HUPO) was established in 2010 as a global initiative. The scientific goals of HPP were to map the entire human proteome in a systematic effort using currently available and emerging techniques. To facilitate these goals, two complementary scientific projects were launched: the Chromosome-Centric Human Proteome Project (C-HPP) and the Biology/Disease-driven-Human Proteome Project (B/D-HPP). The C-HPP and B/D-HPP interact with each other and are supported by the HUPO resource pillars for mass spectrometry (MS), antibody profiling (Ab), and public knowledge bases of genomics, transcriptomics and proteomics [1].

The original scientific goals of the C-HPP were to identify, quantify, and localize expression of at least one representative protein from each of the ~20,000 predicted protein-coding genes in a chromosome-by-chromosome manner. To accomplish this a 25-member international C-HPP consortium that covers 24 chromosomes and mitochondria DNA was formed initially for a 10-year period but recently extended by 15 years (2012–2027) [2,3] (**Fig. 1**). One of the most important missions is to detect and annotate all missing proteins (MPs) which lacked sufficient documented evidence for their existence at protein level (protein existence [PE] in neXtProt) (see **Box 1**). As of neXtProt version 2017–01-23, there are 17,008 PE1 proteins and 2,579 PE2–4 proteins (MPs). In addition, C-HPP will pursue chromosome-based functional characterization of 1,232 uPE1 proteins which have no known function at present.

The C-HPP consortium has established a standard workflow which starts surveying the public DBs (e.g., Ensembl, PRIDE, UniProt, neXtProt and GPMDB) and identifying MPs using proteogenomic approaches (**Fig. 2**). Besides the original objective across the C-HPP to focus on representative proteins, we determined that it is appropriate to include the dynamics and functions of the proteoforms that participate in various metabolic pathways, networks and disease processes (**Fig. 2**). To this end, the C-HPP is also focusing on alternative splice isoforms and post-translational modifications (PTMs) (phosphorylations, glycosylations, acetylations). In addition to the expected products of the ~20,000 predicted protein-coding genes, claims of small open reading frame ORF (smORF) translation products (smORFs), and long non-coding (lnc) RNAs have recently gained attention [4, 5], and are actively investigated using new proteomics technologies.

Given that the C-HPP consortium placed its early priorities on the annotation of ~6000 MPs in neXtProt as of 2012, the end point of the C-HPP shall be the time when there is the best possible match between the numbers of protein-coding genes and those of the validated proteins based on the consensus of information across the public knowledge bases (e.g., PRIDE, PeptideAtlas, neXtProt and GPMDB) [3] (**Fig. 2**). It should be kept in mind that a substantial number of predicted proteins (PE1–4, neXtProt) will not be detectable by mass spectrometry due to lack of tryptic peptides or uniquely-mapping peptides. Current limitations in detecting MPs based on low abundance or restricted expression in time and/or space may be overcome by appropriate selection of tissues or cells at the optimal developmental time or after pathology or stress, some of which can be guided by mRNA transcript analyses. We anticipate that these globally coordinated efforts that employ a stepwise experimental workflow and data sharing process will bring many potential deliverables that will be useful for biological and biomedical research and translation to human health and disease [2,3].

2. Background, history and rationale

When the Human Genome Project (HGP) [6] reported in 2001 and 2003, it was quite a surprise that protein-coding DNA occupies less than 2% of whole genomes [7]. Due to the prime importance of proteins to human biology, the C-HPP focused on the approximately 20,000 protein-coding genes. In the early 2000s the proteomics community had no clear blueprint for a long-term project that aims to elucidate the exact protein numbers encoded by genes, MPs and their modifications (e.g., PTMs, truncation, fusion etc.) [8].

From HUPO's founding on February 9, 2001 there were many initiatives and debates on how to accomplish the HPP (**Fig. 3**). The scientific concept for "HPP" was first coined by Samir Hanash, the inaugural president of the HUPO who launched the Human Plasma Proteome Project (HPPP) as the first HUPO-sponsored initiative [9]. This project initially aimed to map all proteins present in human plasma through the active cooperation of proteomics scientists, industrial sectors and some biomedical resource providers (e.g., for blood, cells and reagents). The HPPP was developed under the leadership of Gilbert Omenn from the United States, which resulted in more than three-dozen publications [9]. Consequently, additional HUPO-sponsored initiatives were created. However, unlike the current C-HPP, other HUPO initiatives (e.g., Human Plasma Proteome Project, Human Liver Proteome Project, Human Brain Proteome Project, etc.) might not have been obligated to set a finite end point due to the nature of the research. Below we summarize how the C-HPP was built along with the HUPO development.

The name for the C-HPP originated from a Gene-Centric HPP (GC-HPP), initially proposed by John Bergeron of Canada in January, 2008, when HUPO hosted a workshop in Barbados (the HPP white paper, https://www.hupo.org/Publications; see Table 1 for more details) [10]. The GC-HPP evolved into the current C-HPP when Young-Ki Paik and colleagues in Korea declared an HPP based on chromosome 13 protein-coding genes on May 30, 2008 (News Focus in '*Science*', Sept 26, 2008) [11] and 'Online News in *JPR*', Oct 7, 2008) [12]. Two more teams (Russia for Chr 18 and Iran for Chr Y) joined the Korean team (Chr 13). At that time, the term "HPP" was used to mean both the GC-HPP and the pre-existing HUPO initiatives.

In early 2009, Sam Hanash suggested the formation of a working group with the aim of establishing a formal HPP and Young-Ki Paik of Korea, 4th HUPO president, appointed Pierre Legrain of France to manage this task. This move indeed stimulated the preparation process for shaping HPP groups. Late in 2009, at the HUPO congress in Toronto, six more teams (Chr 3, 8, 19, 17, 21 and X) joined the C-HPP consortium. The first working group meeting was held in Seattle, Washington, United States, in January, 2010; leaders in proteomics research gathered and discussed the rationale for the HPP and the HPP deliverables. This meeting was followed by the 2010 Busan Korean Human Proteome Organization (KHUPO) meeting in which the major issues and framework of the HPP were refined through public discussion.

In February 2010, the HUPO working group published a HUPO View article [7] that described the concept for the GC-HPP and a rationale for why the HUPO needed an internationally coordinated and gene-centric project. Similar to the human gene “parts” list produced by the HGP, the HUPO wished to produce a human protein parts list that integrated all of the representative proteins encoded by the genes in each chromosome. Critics noted that functional networks and pathways of proteins did not track chromosome locations of the coding genes. However, support for the C-HPP concept was maintained because of the rational division of labor between participating groups and because of the importance of amplicons and cis-regulatory elements leading to co-expression of proteins from certain sets of co-located genes [2,3].

On June 16, 2010, the HPP working group decided to formally organize the C-HPP consortium, which took the basic concept of the GC-HPP, and Paik of Korea was elected as the inaugural chair of the consortium. A major task of the C-HPP was to detect and characterize the MPs coded in chromosomes, chromosome-by-chromosome using a global consortium. At the 2010 HUPO congress in Sydney,the HPP was officially launched with Omenn as chair. At that point, the HPP was composed of two main scientific branches, the C-HPP (formerly GC-HPP) and the B/D-HPP (formerly the various HUPO organ and biofluid-specific initiatives), while the Protein Standards Initiative (PSI) remained a supportive resource group for the HPP. On November 4, 2010, a newly formed C-HPP working group held a teleconference and made decisions on (i) soliciting JPR endorsement of the C-HPP [12]; (ii) nominating co-chairs and a steering committee, which became the Principal Investigator Council (PIC); (iii) defining the C-HPP working period; (iv) data formatting; and (v) plans for future meetings. The results of the various discussions about the HPP, which covered both the C-HPP and the B/D-HPP for a few years, were published as the second HUPO View Paper [1]. This paper described the use of three main technologies, MS, antibodies and bioinformatics, for performing chromosome-based and B/D-based scientific missions, and discussed the deliverables (e.g., protein parts list, reagents, and tools) and strategy for the operation of the HPP.

During the 2011 Geneva HUPO congress, the C-HPP consortium hosted its first PIC meeting and approved a C-HPP logo, a set of standard guidelines, governance, an Advisory Board, and a plan for Special Issue of JPR and collaboration with other scientific groups. This meeting paved the way for an official launching of the C-HPP in the following year.

On March 8, 2012, the C-HPP group published a landmark paper, which described the concept, plan and deliverables [2] and the standard guidelines [3]. The term, ‘Missing Proteins’ was coined in this paper for the first time [2], followed by setting milestones with two stages (Phase I, 2012–2018; Phase II, 2019–2022) for a 10-year period [3]. During the Beijing Asia Oceania (AO) HUPO Congress on May 5, 2012, the C-HPP leadership outlined the strategy for data management and sharing, which was chronicled in a special issue publication of the JPR. On September 10, 2012, the C-HPP was officially launched in Boston with the establishment of the full membership of the 25 chromosome research groups and key leadership (e.g., Chair-Young-Ki Paik of Korea, Co-Chairs William S. Hancock of the US and Gyorgy Marko-Varga of Sweden). At present, the consortium membership is composed of 19 countries for 25 chromosomes: China, Chromosome (Chr) 1, 8, 20; Switzerland, Chr 2; Japan, Chr 3, X; Taiwan, Chr 4; Netherlands, Chr 5; Canada, Chr 6; Australia/New Zealand (ANZ), Chr 7; Korea, Chr 9, 11, 13; India (formerly Thailand), Chr 12; USA: Chr 10, 17, 22; France: Chr 14; Brazil, Chr 15; Spain: Chr 16; Russia: Chr 18; Mexico (formerly Sweden): Chr 19; Germany (formerly Canada): Chr 21; Iran: Chr Y; Italy: Mitochondria (Mt). The Principal Investigator Council (PIC) members approved a scientific policy governing: (i) data deposition into ProteomeXchange database (PXD), stimulated by the HPP and created by the European Bioinformatics Institute, for publication of any protein identification, (ii) call for papers for a JPR special issue with a target publication date of Jan 1, 2013, (iii) data and sample sharing within the consortium (e.g., testis), (iv) data update on the public websites (e.g., PXD, neXtProt, PeptideAtlas, GPMDB), (v) regular updates of the classification of evidence for predicted proteins (neXtProt) (HPP metrics), and (vi) the use of related resources (e.g., HPA, MS pillar, Biobanks etc.). The C-HPP adapted based on the experiences and lessons from the HGP [6] throughout the project period. Naturally, the various teams varied greatly in resources and capabilities. C-HPP made some efforts to secure a commitment to share raw MS/MS data through PXD and proteome-wide datasets for annotation across the C-HPP teams via neXtProt (led by Lydie Lane), and PeptideAtlas (led by Eric Deutsch), aiming to enhance deliverables team-by-team. However, there remain a lot more challenges to reach this goal due to uneven situations for each team with respect to funding, infrastructure and manpower. Scientifically, many challenges also await for those issues involved in splice variants and PTMs in the course of production of protein parts list [5].

3. Progress

Since the C-HPP was launched in 2012, it has made substantial progress during the past 5 years. The C-HPP teams have played a key role in setting some HPP milestones in six areas of cooperation with the bioinformatics teams and individual investigators. The areas of cooperation are: (i) the “Metrics” system for updating the yearly progress in protein annotation [13], (ii) the PXD data submission rule, which was a first step toward community-wide data sharing, (iii) the MS data interpretation guidelines v2.1 (the Guidelines v2.1) [14, 15], (iv) data managing bioinformatics tools, (v) collaboration for the JPR special issue publications, and (vi) rare sample utilization for MP detection. These accomplishments of the C-HPP would have been impossible without HUPO community-wide support and cooperation.

3.1. The metrics system

Since 2012 and in coordination with PeptideAtlas, neXtProt, HPA, and GPMDB, Omenn and colleagues have established the "Metrics" system as the baseline of progress for the C-HPP and the broader proteomics community. At the beginning of the project, 13,664 proteins were validated (PE1) (neXtProt 2012–12-01), leaving 6,395 proteins to detect and validate [13] (see also **Box 1**). The following year, it was decided to leave aside the proteins annotated as dubious (PE5) because most of them are probably irrelevant non-coding DNA. The MPs were defined as predicted proteins annotated as PE2–4. Currently the PE1 proteins number 17,008 (neXtProt 2017–01-23). Despite notable progress in annotation of the MPs, there still remain 2,579 MPs according to the baseline set in 2017. The C-HPP investigators know the current status of the HPP and so can direct their efforts towards identifying MPs more rationally. The HPP Metrics now serves as a surveillance mechanism of any claims of MP detection.

3.2. PXD data submission rule

In the area of public database utilization, the most important development for the C-HPP success was likely timely utilization of the ProteomeXchange database (PXD) site [16]. The PXD site has facilitated data capture and provides all deposited data sets and meta-data, resulting in enhanced collaborative research within the proteomics community (**Fig. 2**). This site met the needs of investigators and connected all of the individual MS data resources, including PRIDE, PeptideAtlasand the GPMDB. As of June 1, 2017, there were more than 6,000 datasets uploaded into the ProteomeXchange public database. For journal publications, authors are required to provide the PXD identifiers, which also can be used to provide permission for the dataset to be made public on publication. Deposited MS data are used for reanalysis of peptide data in a standardized method and protein matches from those experimental datasets are made by PeptideAtlas [17] and the GPMDB [18]. Note that PeptideAtlas provides FDR calculation based on a single analysis of all public LC-MS/MS data, and therefore offers accurate FDR threshold for the whole available public MS based proteomics data. This allows us to list the genes with identified proteins at 1% FDR level. The incorporation of the PeptideAtlas results into the neXtProt curation of the PE proteins by combining mass spectrometry and multiple other types of protein data [13, 14, 17, 19] also led to a stringent evaluation of the data quality from large-scale experimental data sets (**Fig. 2**). This PXD-based working module may be the first of its kind in the proteome community from which every stakeholder can benefit from high quality research. Data deposition through the PXD demonstrates the degree of scientific progress of the HPP community and serves as a barometer of the project deliverables. Any investigators would benefit from this obligatory data submission, sharing and retrieval process. The accumulated data for individual chromosomes will also facilitate the completion of protein mapping for all chromosomes in the near future. We are confident that the added value of such a coordinated effort on both "Metrics update and utilization of public DBs" will result in the production of a more reliable protein parts list, accurate annotation for MPs, and quantitative proteomics for biomedical applications (e.g., disease biomarkers) (**Fig. 2 & 4**). A notable benefit of PXD is the full availability of the datasets from various papers, which permits standardized reanalysis by PeptideAtlas, GPMDB, and Cox et al with MaxQuant to assess claims for “found” MPs by reviewers before publication or database curators post publication [20].

3.3. The MS data interpretation guidelines v2.1

As the C-HPP moved forward with the mission to identify and characterize MPs, it was necessary to guarantee and exercise the quality of MS data in proteomics [13]. For the C-HPP to succeed, the data must have been deposited in PXD, curated and accepted at either the PeptideAtlas and/or the GPMDB. At present though, only data from PXD is incorporated into the HPP accepted database neXtProt. To facilitate the reuse of individual datasets by other teams, the C-HPP Wiki site, managed by Peter Horvatovich of the Netherlands (Chr. 5 group), provides the list of produced datasets with biological and other details (http://c-hpp.webhosting.rug.nl/tiki-index.php), as well as an excellent communication forum for all consortium members; teams can report their progress and freely obtain important resources for their research. For example, when MS data were available from two draft papers by Kim et al. [22] and Wilhelm et al. [23], there were concerns by some in the community about the gross overmatch of peptides, especially short peptides, to (outdated) predicted protein sequences. However, the availability of all the relevant data in PXD provided an opportunity for the cross-analysis of peptide data produced by different investigators. Several independent re-analyses of major data from these two papers indicated that there were thousands of false-positive protein identifications [13, 24, 25]. Major reason for this higher false positive rate was due to either no calculation of FDR at protein level or combining results from multiple identifications, which leads to the accumulation of errors with respect to the full dataset. Therefore, the HPP leadership established more stringent the guidelines [14]. The Guidelines v2.1 set the requirement for a protein-level false detection rate (FDR) at ≤1% with two uniquely-mapping (proteotypic) peptides of at least nine amino acids in length for claims of finding missing proteins or novel translatable gene products (or proteoforms) produced from lncRNAs or PE5 dubious predicted genes [15]. The Guidelines v2.1 are not only fully implemented by both the neXtProt and PeptideAtlas, but also have been adopted by the C-HPP investigators. Thus, the Guidelines 2.1 raised the bar for the identification of missing proteins and novel peptides, which are peptides that are identified by proteogenomics approaches and cannot be found in the canonical sequences of public databases such as Ensembl and Uniprot. These quality criteria will build a more accurate and more reliable human proteome knowledgebase [14]. There should be additional efforts to scrutinize for the presence of ion ladder features and the absence of anomalies and to carefully check that the results agreed between the multiple search engines used [13] (**Fig. 2**). In addition, it is recommended to validate the identifications using assays with SRM or SWATH and high-quality synthetic peptides [26]. This scrutinized reanalysis should be applied also to those peptides that appear to represent translation products from lncRNAs or from pseudogenes.

3.4. Data managing bioinformatics tools

Since the C-HPP emphasizes the identification of MPs, it is desirable to have tools for visualizing the datasets, automation of raw MS/MS profiles and the collection of up-to-date information on the activity of individual chromosome teams. We also realized the need for a more comprehensive baseline on the cross-analysis of multi-omics data (genome-proteome, transcriptome-proteins, etc.) in a systematic display form (**Fig. 3**). Some progress should be noted for such effort in several areas during the early phases of the C-HPP. First, neXtProt provides “proteomics” and “peptides” views for each protein, in which all the peptides that were identified are displayed, along with associated metadata. In addition, it provides the sequences of synthetic peptides that can be used for quantitative proteomics analysis according to SRMAtlas [27, 28]. neXtProt also built two important tools for C-HPP: a graphical view of the status of protein validation in each chromosome, and a tool to assess the uniqueness of peptides among human sequences, that takes into account all described single amino-acid mutations [29]. Use of this tool enables unequivocal identification of the peptide as being derived from one protein or not due to consideration of single amino acid variants (SAAVs), an essential step in the identification of MPs according to the current guidelines v2.1.

Second, some bioinformatics databases and tools have been developed by individual groups. For example, the GenomewidePDB 2.0 [30] described new features that integrated transcriptomic information (e.g., alternatively spliced transcripts), annotated peptide information and included an advanced search interface that could find proteins of interest when applying a targeted proteomics strategy. CAPER3.0 is the latest version of the analytical resource for data sets from the Chinese C-HPP Consortium [31]. The PPLine is a Python-based proteogenomic pipeline from the Russian team of Chromosome 18 [32], which has the features of automated discovery of SAAV polymorphisms, indels, and alternatively spliced variants from raw transcriptome and exome sequence data in addition to the prediction of proteotypic peptides. The “dasHPPboard” developed by the Chromosome 16 group in Spain has facilitated the analysis of a variety of proteogenomic data sets. This tool can be used to identify samples with high amount of transcript for MPs of category PE2 [33]. The Michigan Proteome Visualization Tool (MI-PVT), developed by the Chromosome 17 teams of the US, is a web-based tool that displays by chromosome or by protein family [34]. The iterative threading assembly refinement (I-TASSER) and COFACTOR algorithms, which are well-established tools in structural biology for predicting protein folding and protein functions, were applied to the list of 572 PE5 uncertain/dubious predicted proteins to evaluate their prospects for conformational folding and potential functions [35].

As part of the Korean Chr 11 project, Park et al. [36] developed an Integrated Proteomic Pipeline (IPP) using multiple search engines (SEQUEST, MASCOT, MS-GF+) for liquid chromatography (LC) MS/MS analyses of brain tissues with a controlled FDR ≤ 1% at the protein level. They compared the IPP to a conventional proteomic pipeline. In hippocampal tissue, the IPP yielded 5,756 proteins, including 477 alternative splice variants (ASVs) versus 4,453 proteins and 182 ASVs using the IPP. Deutsch et al. (2015) compared the components of the Tiered Human Integrated Sequence Search Databases for Shotgun Proteomics (THISP), with an emphasis on the completeness and efficiency of the alternative search engines [37]. C-HPP investigators have utilized targeted approaches with selected reaction monitoring (SRM) assays for quantitative identifications of the MPs [38–40]. With the availability of the Human SRMAtlas, the accurate detection and quantification of any known or predicted human protein from complex biological specimens is now possible because the SRMAtlas contains 166,000 proteotypic peptides, verified high-resolution spectra, multiplexed SRM assays and a web database with unlimited free access (www.srmatlas.org)[41]. The SRMAtlas has moved further towards the use of sequential window acquisition of all theoretical fragment ion spectra (SWATH) analyses.

Complementary to MS-based protein characterization, HPP investigators now utilize antibody resources provided by the Human Protein Atlas (HPA, http://www.proteinatlas.org) [42] for their deep mining of protein expression in specific tissues of organs [43]. HPA has launched a major quality improvement effort to validate antibody specificity during the same time period that we launched the HPP Guidelines for MS Data Interpretation. Together with PeptideAtlas, HPA provides a rich source for determination of particular proteins of interest with respect to their spatial distribution at a subcellular level, as well as for immunohistochemistry on tissue microarrays for distribution of the protein expression in normal and cancer tissues. The HPA database has been developed in a chromosome-centric manner which enables search for protein expression profiles, protein classes, and chromosome location.

3.5. Collaboration system and annual special issue publications

The C-HPP can be characterized as a two-pronged mission carried out by multinational projects focused on the following: (i) hypothesis-driven study for MPs (seeking their identification, biology, structure, disease involvement and roles in pathways) and (ii) a data gathering cooperative project (constructing a knowledge base). More importantly, the C-HPP has carried out these missions in a cross-community collaboration during the past five years. The C-HPP has created datasets through PXD that are shared with all of the chromosome teams that resulted in the growth of aggregated knowledge and more information on the MPs. When considering such complex missions, there are ample reasons why each team needs to collaborate with each other within the HPP community. The different chromosome teams used different sets of biological or clinical samples (e.g., liver, brain, placenta, testis and sperm, etc.), which often results in a cross-chromosomal protein parts list. It is designed that sharing the results from different samples may offer an opportunity to look into low abundance MPs present across the chromosomes. Each team could pull out those proteins encoded by any chromosome. Of course, those datasets should have been already been vetted for compliance with the Guidelines v2.1 [14].

The terms of cooperation between the chromosome teams were established in Berlin in 2013. The C-HPP PIC members outlined a policy stating that, while each group would enjoy research freedom, they would respect mutual data sharing and study guidelines for individual proteins of interest by consortium members. It was also encouraged that each member would honor the spirit of mutual collaboration between teams (e.g., collaboration between the individual team and the team responsible for the chromosome encoding the protein of interest). This approach would benefit the whole consortium. Many good examples (e.g., Chr 1,8,20; Chr 2 and 14 teams; Chr 11 and 13 teams; Chr 13 and 18 teams) have already been published in the JPR special issues over the past four years [38, 44, 45]. An additional example of collaboration is the Chinese (Chr 20) and Korean (Chr 11) teams exploring the possibility of sharing their special databases and bioinformatic workflow (e.g., RNA Seq and MS profiling of SAAVs).

To expedite the further discovery and annotation of MPs, the C-HPP leadership has proposed two collaborative networks. The first collaborative network was called a “cluster” of C-HPP teams that aimed to accelerate work on a specific target disease of interest (e.g., cancers; Chr 1, 8, 20, 9, 11, 13, 7, 12, 17), reproductive biology (Chr 2, 14, X, Y), and broad application of the In Vitro Transcription/Translation (IVTT) platform (Chr 5, 10, 15, 16, 19). At the 2016 Taipei HUPO Congress, additional groups studying membrane proteins (Chr 4, 18, 21) and neurodegenerative disorders (Chr 1, 3, 6, 11, 12) were organized. A special attribute of the clusters is expected collaboration with the corresponding B/D-HPP groups (e.g. cancers and CPTAC, reproductive biology, and brain). This collaborative strategy will be applied to all areas of the C-HPP, MPs and proteoform discovery, validation of data integration, disease mechanisms and biological studies.

The second collaborative network launched in 2016 was the “neXt-MP50 Challenge”. Whereas progress has been rapid in finding MPs to now, the rate of new identifications is slowing as the low hanging fruits are picked. This is typical in completing projects—it is the last 5% or less that is the hardest to complete. Thus, this new campaign was initiated to galvanize the chromosome teams to focus on identifying a “bite size” of MPs, say 50, by HUPO-2018. Although several chromosomes have fewer than 50 MPs and the Italian mitochondrial DNA team has completed their project (see chromosome-by-chromosome tables and figure in the Metrics paper) [13], some chromosomes have more than 200 missing proteins. Both targeted MP searches and unbiased deep approaches, *e.g.* proteomic analyses of underexplored human cells and tissues, are ongoing by multiple Chromosome Groups. This is the goal of the neXt-MP50 Challenge—itself seen as an interim goal in achieving an accurate first draft of the complete human proteome. Many strategies are possible and follow the interests of the members of the various teams. Leadership pointed out the strategy developed by the Chr 16 analyses and Chr 2/Chr 14 collaboration in order to select the most likely specimens and methods to detect specific missing proteins. In this way, targeted and focused searches can be implemented by each team to specifically seek MPs in the next-MP50 Challenge. Of course, the goal is not set in stone; rather, it is seen as a rally cry to dedicate work to finding recalcitrant MPs as they become harder to find and unequivocally identify.

3.6. Exploring rare samples

Each chromosome team of the C-HPP has been working toward identifying the MPs. One of the major strategies has been to take a deep dive into proteomic analyses of tissues that either have good evidence of transcript expression of the genes of interest (i.e., PE2 in neXtProt), or would be expected to yield MPs based on unusual, hard to access tissue or small numbers of cells that are found either dispersed in other tissue or present as cell layers or sheets lining different tissue compartments for which there is no transcript data. The Swiss Chromosome 2 team, the French Chromosome 14 team [38, 45] and the Chinese team (Chr 1, 8, 20) [40] have exploited this approach with their focus on testis and spermatozoa proteomes. All teams who use the sample resource banks for specific cell lines and clinical material shall apply this deep dive approach. Other rare tissues (i.e., nasal epithelial cells, dental pulp, sensory organ cells, or hair cortex) can also be used to identify MPs. Fetal tissues are also expected to be a source of MPs that may be transcription factors and cytokines that have transient expression temporally or spatially during embryogenesis. Naturally, accessing such tissues is often difficult in many countries, limiting this source of tissue for analysis.

4. Challenges for the C-HPP

4.1. Missing protein issues

4.1.1. Why is it so difficult to identify missing proteins?—MPs are those proteins that belong to PE2–4, which lack confident MS evidence or any antibody-captured verification, but for which there can be either transcriptomic evidence in human (PE2), or any type of evidence in other species (PE3)(see **Box 1**) [15, 19]. They also include those proteins that have no convincing MS information but do have other types of protein-level evidence (PE1). There has been an emerging 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. Even with high quality mass spectra, peptides confirmed using synthetic peptides and multiple reaction monitoring, it can be challenging to find a match to the missing proteins due to alternative splicing [15]. Contrary to this, there are a few cases for PE1 entries without MS evidence. It seems necessary to place an exception in the Guidelines v2.1 for such 'non-MS MP detection' (e.g., TBL1Y, prestin) in which other cellular and molecular evidences support the protein evidence [47, 48]. This issue has been well discussed in recent publication by Omenn et al., [5]. Briefly, as in case of some of proteins that were driven from lncRNAs or smORFs, HPP group may consider updating the current HPP data interpretation guidelines v2.1, checklist #15 [14] to deal with such extraordinary case in which weaker evidence is offered for an extraordinary protein or coding element detection (e.g., lacking lysine site-unsuitable for trypsin digestion), justify that other peptides cannot be expected. Given the controversial reports on the presence of translatable proteins or peptides produced from lncRNAs and smORFs [49–52], HPP leadership is also currently working on this issue by coordinating for re-analysis of some of claimed translatable proteoforms (e.g., lncRNA) to build a consensus on the update of Guidelines v.2.1.

4.1.2. Potential causes for the status of missing proteins—It is conceivable that MPs may result from one of the following main causes (**Table 1**): (i) A low abundance of proteins can be produced by routine genomic activities e.g., ribosome nascent complex (RNC), alternative splicing and temporal specific gene regulation. These low-abundance proteins are likely to exhibit their transcriptional expression at both tissue and subcellular specific manner (PE2 proteins) and may be involved in gene regulation in poorly understood regions of the genome [53]. Because of the large dynamic range of proteins in biological samples, low-abundance isoforms and proteoforms that may be associated with disease, can be extremely difficult to measure [54]. Proteomic analysis has shown that post-translational modifications can be different in alternative splicing variants and that amino-acid polymorphisms can generate additional variants.(ii) The MPs may result from the limitation of search methods for the MS data or MS analysis e.g., low sensitivity and potential errors in the handling of low-quality experimental spectra [53–56]. To identify MPs, researchers generally use neXtProt or SwissProt, which provide sufficient coverage for protein coding human genes. However, insufficient coverage may also be important for novel peptides or for isoform detection [57].(iii) Highly hydrophobic proteins (such as olfactory receptors, membrane proteins and highly insoluble proteins), which are estimated to represent 20–30% of the total encoded human proteome, are notoriously difficult to identify by MS [58–61]. (iv) The MPs may result from low-resolution mass analysis [59] (v) Rare proteins are produced only at specific times and in specific cells.

The potential solutions for such problems may be: (i) the development of advanced software for spectral search methods, which can result in greater sensitivity and more accuracy in the search for the MPs [54, 55, 62], (ii) a combined solubilization method for the initial sample treatment step [63]; among the reagents, sodium deoxycholate and RapiGest showed good solubility and enzyme compatibility as well as easy removal in later steps prior to the MS analysis; (iii) the use of high-resolution mass spectrometry; and (iv) the use of RNA Seq analysis to better guide tissue selection and timing and to identify tissues or cells not commonly used for MS analyses [64–66] (**Table 1**). In particular for PE2 proteins, C-HPP investigators can approach to perform target proteomics in order to identify PE2 class MPs by utilizing genomics dataset (e.g., RNA-seq) obtained from some specific tissues or cell lines [30]. Duek et al., attempted to detect MPs from testis by analyzing RNA-seq data and proteomic profiling with MRM assays [38, 39].

4.2. Protein variants issue

In recognition of the huge space of protein proteoforms, the characterization of protein products from protein-coding genes should include deep analyses for sequence variants, splice variants and post-translational modifications. Such technical and biological complexity includes additional reference genomes, sequence variants and post-translational modifications that match the *m*/*z* ratios of the features of the reference protein or that explain novel peptides attributed to translation from a lncRNA sequence [17, 66]. We must be alert to such alternative explanations when automated search engines identify a protein that has never before been observed in the same types of specimens. The community also needs to agree on a viable set of next targets, such as alternative splice variants of disease-important proteins or the detection and functional characterization of splice variants [67]. Regarding post-translational modifications, PeptideAtlas has now created a Phospho-PeptideAtlas which has been integrated into neXtProt.

4.3. Olfactory receptors

Olfactory receptors (ORs) are G protein-coupled receptors that detect odorants and are usually expressed in the cell membranes of the cilia and synapses of the olfactory sensory neurons [68] and in the epithelium of the human airway [69]. Sperm cells also express some ORs, which are thought to be involved in chemotaxis to find an egg cell [70]. The ORs form a multigene family consisting of around 1,000 genes in humans, but only ~400 functional genes code for ORs; the remaining 600 candidates are pseudogenes [71,72]. These latter ORs are indeed the best examples of a so-called black box in the proteomics field, although their coding genes are distributed over almost all the chromosomes [73]. For instance, the PeptideAtlas has reevaluated its OR entries and has eliminated the two remaining entries as of 2014 [14]. The GPMDB has reduced its very long list of entries to six with high-quality entries and then recognized that even these may be better matches to other proteins. Ezkurdia et al. [23] examined the spectra made available by Kim et al. [20] of 108 reported olfactory receptor proteins and by Wilhelm et al. [21] of 200 reported olfactory receptor proteins. Ezkurdia and colleagues concluded that not one passed muster for quality spectra or valid protein match [23].

4.4. ENCODE data issue

It was suggested earlier that ENCODE and the C-HPP data should be integrated into modules (metabolic or signaling pathways, gene sets and chromosome regions) and networks to develop system-wide models of biological processes as exemplified in the case of the correlation between *ERBB2* mutation and oncogenic gene expression. There is a good example for the utilization of ENCODE [46] through which both translation of ENCODE data and major proteomic technology pillars improved the identification of the MPs, novel proteoforms and PTMs [74]. The output of the proteomic data curated at the Peptide Atlas and GPMDB sites presented in a chromosome-centric format could be well aligned with ENCODE data [74]. The results from this effort have suggested that some alternative splicing forms detected at the transcript level were in fact translated to proteins, which supports synergistic effort between two projects. This type of collaboration deserves more study in the future.

5. Looking to the future

Given that the highest priority was initially placed on MP annotation, it was necessary for the C-HPP to make changes once this was well under way with substantial progress made. The C-HPP leadership decided to take the immediate action of restructuring the current organization into a more functional and cooperative module (**Fig. 4**). The purposes of this reorganization are to stimulate the activity of each team and the work performance of chromosome-based proteomics research in a cooperative manner, resulting in more credible and reproducible results. This reorganization will improve efficiency and create a closer working relationship between chromosome teams and their resource pillars towards the completion of the scientific goals. With the new clusters of Chr teams, we anticipate creating a strong research focus of chromosome teams on using the advanced proteomics platform technologies and resources, which should enable secure research funding. In this structure, the MP annotation team mainly comprises those involved in the neXt-MP50 challenge campaign led by Chris Overall, whereas the MP bioinformatics team mainly comprises those involved in the public DB maintenance or related bioinformatics services which is led by Lydie Lane. Lastly, in the recent HUPO Congress in Dublin, C-HPP PIC established a new MP functional study team, termed neXt-CP-50 (CP: characterizing unknown function proteins) led by Young-Ki Paik, which would focus more on the characterization of the MPs using various cell lines, rare tissues (e.g., nasal epithelium over the cribriform plate for ORs, IVTT tech, membranes and model organisms [75]. This MP functional study team targets uPE1 proteins (unknown function PE1) which is now 1,232 (neXt-Prot 8–1-2017). These proteins have not revealed any function by any means including, but not limited to, database searching, domain analysis, Intrinsic feature analysis (transmembrane, coiled coil proteins etc.), similarity analysis and metabolic pathway involvement. This indicates a lot more works are waiting for characterization of these proteins. For the short-term, as a pilot phase, each Chr team would prioritize 5–10 uPE1 proteins and carry out protein characterization works according to its own experimental strategy. We anticipate that the reorganized teamwork should be an ideal working model to shorten the time from discovery, annotation and functional characterization to translational proteome biology.

As we begin a new working module (**Fig. 4**), it is necessary to find a better way of enhancing our research capability and the deliverables of the C-HPP teams. To this end, we provide a few action items here. For example, many teams may be interested in organizing studies of families of proteins whose genes often occur in clusters on certain chromosomes; of amplicons, which are quintessential cis-regulated chromosomal features, often with co-expression; and alternative splicing, a key evolutionary development in multicellular organisms with multi-exonic genes to generate greater protein diversity. As PTMs and protein proteoforms often change in response to physical, pathological, nutritional and environmental stress, it is therefore likely that isoforms will show higher selectivity and specificity as diagnostic biomarkers or molecular therapeutic targets than the protein and transcript mixtures from individual genes.

6. Conclusions

Here we described that much has been achieved over the past few years since the C-HPP initiative has successfully promoted global collaborations while pursuing the full protein parts list, a moving target, based on the known protein coding genes. It has also established a system that enables the deposition of proteomics datasets and provides standardized reanalysis of those data sets. Several issues have emerged in regard to the potential functional aspects and detection of lncRNAs and small proteins. Having integrated the transcriptomic and proteomic information, we now look forward to the next phase of better understanding the complexity of human biology. The progress to date encourages us to reset the goals and scope of our research by adding other targets e.g., alternate splice forms, PTMs, and SNP-derived protein characterization. The smORF proteins (or mini-proteins) may also be good targets. We will have more effort on characterizing sequence variants, amplicons, and splice isoforms of disease-related proteins. We hope that newly initiated neXt-CP50 campaign will motivate all members in pursuit of new functions of uPE1 in parallel with neXt-MP50 campaign, which needs more interactions with the B/D-HPP teams. National team-based long-term funding to cover the entire period of project will also be enhanced by functional analysis of MPs, which may lead to useful disease biomarker discovery. Considering the current speed of progress and advances in detection techniques, C-HPP has recently extended its term from 2022 to 2027 [76]. These additional years will provide more time to focus on the functional characterization of the MPs (neXt-CP50) [76].

7. Expert commentary

Although it is quite embryonic stage to include the translational products from lncRNA and smORFs due to lack of sufficient evidence, in a long-run, it is plausible to predict their potential assignment of novel proteins to human protein parts list upon identification and validation. Given that there has been not much progress, with the exception of detailed publications from Chr 12 and Chr 17 on interesting amplicons, more active investigation of cis-regulatory phenomena and amplicons should be pursued at the protein level. More effort should be done in order to expand the use of rare biological samples with the clues that the HPA transcript results show relatively few tissue-specific transcripts/proteins outside of testis/sperm and brain. We suggest that the C-HPP initiative will support the ongoing evolution of the proteomics field by adopting information flowing from molecular biology advances in integrated transcriptomics/proteomic measurements, progression of research organizations from individual laboratories to international research alliances, deep dive proteomic discovery experiments of protein variants generated from alternative splicing transcripts. The development of informatics systems and interfaces to allow the integration of genomic, proteomic and individual protein variation information will expedite this process. At the juncture of imitation of new campaign, neXt-CP50, which targets uPE1 proteins, it is time to establish an integrated SOP for elucidating protein function and big data for new biology.

8. Five-year view

In overall, much has been achieved over the past few years since the C-HPP initiative has successfully promoted global collaborations. It is anticipated that there will be the early draft version of accurate full parts list of MS-and Ab-based protein annotation of proteins in accordance with function and cellular localization. We shall have more clear and robust criteria for annotation of proteins that are compatible with both MS and non-MS validation of protein existence by upgrading the current guidelines v2.1. In the area of analytical method development, it may be necessary to combine the genomic resources with the proteome DB to identify rare or low abundance proteoforms as components of protein parts list. It has also established a system that enables the deposition of proteomics datasets and provides standardized reanalysis of those data sets. Having integrated the transcriptomic and proteomic information, we now look forward to the next phase of better understanding the complexity of human biology. The draft of an accurate proteome map will become available with the functional map. The progress to date encourages us to reset the goals and scope of our research by adding other targets e.g., number of alternate splice forms, PTMs, and SNP-derived protein characterization. The smORF proteins (or mini-proteins) can also be good targets. Considering the current speed of progress and advances in detection techniques, we are in a position to think about extending the end point from 2022 to 2027 [77]. These additional years will provide more time to focus on the functional characterization of the MPs.

Several issues have emerged in regard to the functional aspects and detection of lncRNAs and small proteins. These peptides or small proteins seem to be important emerging targets for cancer biomarker studies. We suggest that the C-HPP initiative will support the ongoing evolution of the proteomics field. Such changes include: the earlier adoption of information flowing from molecular biology advances such as ENCODE; integrated transcriptomics/proteomic measurements; progression of research organizations from individual laboratories to international research alliances; deep dive proteomic discovery experiments; top-down analyses of protein variants generated from alternative splicing variants and alternative splicing transcripts; better statistical tools for assessing extremely large data sets; and the development of informatics systems and interfaces to allow the integration of genomic, proteomic and individual protein variation information.

Acknowledgements

We thank proteomics scientists, leaders of public DBs, PIC members, and funding agencies. We thank Dr. W. Hancock for his great contribution to the early phase of the C-HPP. Special thanks go to Dr. Jin-Young Cho for his help on drawing the figures and JW Kim for their help preparing this manuscript. This paper is dedicated to all the C-HPP members as well as all other related investigators who contributed their efforts and data to move this global project forward in various ways.

Funding

This work was supported by grants from the Korean Ministry of Health and Welfare: HI13C2098-International Consortium Project and HI16C0257 (awarded to Y.-K. Paik), from the U.S. National Institutes of Health: P30ES017885–01A1 and U24CA210967 (awarded to G.S.Omenn), and from Canadian Institutes of Health Research, 7-year Foundation Grant and Canada Research Chair in Protease Proteomics and Systems Biology: FDN-148408 (awarded to C.M. Overal).

References

Papers of special note have been highlighted as:

* of interest

** of considerable interest

1**. Legrain P, Aebersold R, Archakov A, et al. The human proteome project: current state and future direction. Mol. Cell. Proteomics 2011;10:M111.009993

Describe an integrative HPP with a general experimental strategy that use the three working pillars for HPP: mass spectrometry, antibody capture, and bioinformatics tools and knowledge bases.

2**. Paik YK, Jeong SK, Omenn GS, et al. The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. Nat. Biotechnol 2012;30:221–223.

Describe the concept, goals, scope and deliverables of C-HPP in the context of cooperation of 25-international teams. Term ‘missing protein’ was first coined in this paper.

[PubMed: 22398612]

3**. Paik YK, Omenn GS, Uhlen M, et al. Standard Guidelines for the Chromosome-Centric Human Proteome Project. J. Proteome Res 2012;11:2005–2013.

Describe the first standard guidelines and stepwise experimental approaches including data production and govenance of the project.

[PubMed: 22443261]

4. Paik YK, Overall CM, Deutsch EW, et al. Progress in the Chromosome-Centric Human Proteome Project as Highlighted in the Annual Special Issue IV. J. Proteome Res 2016;15:3945–3950. [PubMed: 27809547]

5. Omenn GS, Lane L, Lundberg EK, Overall CM, Deutsch EW. Progress on the HUPO Draft Human Proteome: 2017 Metrics of the Human Proteome Project. J Proteome Res.2017 8 30. doi: 10.1021/acs.jproteome.7b00375. [Epub ahead of print] PubMed PMID: . [PubMed: 28853897]

6. Collins FS, Morgan M, Patrinos A. The Human Genome Project: lessons from large-scale biology. Science. 2003;300:286–290. [PubMed: 12690187]

7. Kim HS. Genomic impact, chromosomal distribution and transcriptional regulation of HERV elements. Mol. Cells 2012;33:539–544. [PubMed: 22562360]

8. A Gene Centric Human Proteome Project" - HUPO Views. Mol Cell Proteomics. 2010; 9: 427. [PubMed: 20124355]

9. Omenn GS. Report from the 2nd Annual US HUPO Meeting on the HUPO Human Plasma Proteome Project. Expert Rev. Proteomics. 2006;3:165–168. [PubMed: 16608429]

10. The HPP White Paper, https://www.hupo.org/publications, Vancouver, Canada (cited 2017 10 14) available from www.hupo.org,

11. Services RF, Proteomics ponders prime time, Scienc, 2009; 321: 1758–1761

12. Hancock WS, Omenn GS, Legrain P and Paik Y-K Proteomics, Human Proteome Project, and Chromosomes. J. Proteome Res, 2011;10: 1

13. Omenn GS, Lane L, Lundberg EK, et al. Metrics for the Human Proteome Project 2016: Progress on Identifying and Characterizing the Human Proteome, Including Post-Translational Modifications. J. Proteome Res 2016;15:3951–3960. [PubMed: 27487407]

14**. Deutsch EW, Overall CM, Van Eyk JE, et al. Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 2.1. J. Proteome Res. 2016;15:3961–3970.

Describes mass spectrometry data interpretation guidelines that should be applied to all HPP data contributions.

[PubMed: 27490519]

15. Omenn GS, Lane L, Lundberg EK, et al. Metrics for the Human Proteome Project 2015: Progress on the Human Proteome and Guidelines for High-Confidence Protein Identification. J. Proteome Res. 2015;14:3452–3460. [PubMed: 26155816]

16**. Vizcaíno JA, Deutsch EW, Wang R, et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nat. Biotechnol 2014;32:223–226.

Describes the implementation of the ProteomeXchange database for submission and dissemination of MS-based proteomics data in the proteomics community

. [PubMed: 24727771]

17. Deutsch EW, Albar JP, Binz PA, et al. Development of data representation standards by the human proteome organization proteomics standards initiative. J. Am. Med. Informatics Assoc 2015;

18. Fenyö D, Beavis RC. The GPMDB REST interface. Bioinformatics. 2015;31:2056–2058. [PubMed: 25697819]

19. Lane L, Bairoch A, Beavis RC, et al. Metrics for the Human Proteome Project 2013–2014 and Strategies for Finding Missing Proteins. J. Proteome Res 2014;13:15–20. [PubMed: 24364385]

20. Cox JT, Marginean I, Smith RD, et al. On the ionization and ion transmission efficiencies of different ESI-MS interfaces. J. Am. Soc. Mass Spectrom 2015;26:55–62. [PubMed: 25267087]

21. C-HPP Wiki Site, http://c-hpp.webhosting.rug.nl/tiki-index.php, Groningen The Netherlands, cited Oct 14, 2017 Available from c-hpp.org

22**. Kim MS, Pinto SM, Getnet D, et al. A draft map of the human proteome. Nature. 2014;509:575–581.

Describes a draft version of human proteome map which contains various human tissues and cells [in accordance with the C-HPP data frame—what does this mean?].

[PubMed: 24870542]

23**. Wilhelm M, Schlegl J, Hahne H, et al. Mass-spectrometry-based draft of the human proteome. Nature. 2014;509:582–587.

Describe the first draft version of human proteome map which contains various human tissues cells, and biofluids [in accordance with the C-HPP data frame—what does this mean?].

[PubMed: 24870543]

24. Ezkurdia I, Vázquez J, Valencia A, et al. Analyzing the First Drafts of the Human Proteome. J. Proteome Res. 2014;13:3854–3855. [PubMed: 25014353]

25**. Savitski MM, Wilhelm M, Hahne H, et al. A Scalable Approach for Protein False Discovery Rate Estimation in Large Proteomic Data Sets. Mol. Cell. Proteomics 2015;14:2394–2404.

Greatly reduces the claims of proteins detected in Wilhelm et al, ref 19.

[PubMed: 25987413]

26. Schubert OT, Gillet LC, Collins BC, et al. Building high-quality assay libraries for targeted analysis of SWATH MS data. Nat. Protoc 2015;10:426–441. [PubMed: 25675208]
27. Gaudet P, Argoud-Puy G, Cusin I, et al. NeXtProt: Organizing protein knowledge in the context of human proteome projects. J. Proteome Res 2013;12:293–298. [PubMed: 23205526]
28. Gaudet P, Michel PA, Zahn-Zabal M, et al. The neXtProt knowledgebase on human proteins: 2017 update. Nucleic Acids Res. 2017;45:D177–D182. [PubMed: 27899619]
29. Schaeffer M, Gateau A, Teixeira D, et al. The neXtProt peptide uniqueness checker: a tool for the proteomics community. Bioinformatics. 2017;10.1093/bioinformatics/btx318. [Epub ahead of print]
30. Jeong SK, Hancock WS, Paik YK. GenomewidePDB 2.0: A Newly Upgraded Versatile Proteogenomic Database for the Chromosome-Centric Human Proteome Project. J. Proteome Res 2015;14:3710–3719. [PubMed: 26272709]
31. Yang S, Zhang X, Diao L, et al. CAPER 3.0: A Scalable Cloud-Based System for Data-Intensive Analysis of Chromosome-Centric Human Proteome Project Data Sets. J. Proteome Res 2015;14:3720–3728. [PubMed: 25794139]
32. Krasnov GS, Dmitriev AA, Kudryavtseva AV, et al. PPLine: An Automated Pipeline for SNP, SAP, and Splice Variant Detection in the Context of Proteogenomics. J. Proteome Res 2015;14:3729–3737. [PubMed: 26147802]
33. Tabas-Madrid D, Alves-Cruzeiro J, Segura V, et al. Proteogenomics Dashboard for the Human Proteome Project. J. Proteome Res 2015;14:3738–3749. [PubMed: 26144527]
34. Panwar B, Menon R, Eksi R, et al. MI-PVT: A Tool for Visualizing the Chromosome-Centric Human Proteome. J. Proteome Res 2015;14:3762–3767. [PubMed: 26204236]
35. Dong Q, Menon R, Omenn GS, et al. Structural Bioinformatics Inspection of neXtProt PE5 Proteins in the Human Proteome. J. Proteome Res 2015;14:3750–3761. [PubMed: 26193931]
36. Park GW, Hwang H, Kim KH, et al. Integrated Proteomic Pipeline Using Multiple Search Engines for a Proteogenomic Study with a Controlled Protein False Discovery Rate. J. Proteome Res 2016;15:4082–4090. [PubMed: 27537616]
37. Deutsch EW, Sun Z, Campbell DS, Binz PA, Farrah T, Shteynberg D, Mendoza L, Omenn GS, Moritz RL. Tiered Human Integrated Sequence Search Databases for Shotgun Proteomics. J Proteome Res 2016;15:4091–4100. [PubMed: 27577934]
38. Vandenbrouck Y, Lane L, Carapito C, et al. Looking for Missing Proteins in the Proteome of Human Spermatozoa: An Update. J. Proteome Res 2016;15:3998–4019. [PubMed: 27444420]
39. Duek P, Bairoch A, Gateau A, et al. Missing Protein Landscape of Human Chromosomes 2 and 14: Progress and Current Status. J. Proteome Res 2016;15:3971–3978. [PubMed: 27487287]
40. Zhao M, Wei W, Cheng L, et al. Searching Missing Proteins Based on the Optimization of Membrane Protein Enrichment and Digestion Process. J. Proteome Res. 2016;15:4020–4029. [PubMed: 27485413]
41. SRM Atlas, www.srmatlas.org, Systems Biology, Seattle (cited 2017 Oct 14), available from http://www.peptideatlas.org/
42. Human Protein Atlas, http://www.proteinatlas.org, Stockholm, Sweden (cited 2017 Oct 14) available from http://www.scilifelab.se/

43**. Uhlen M, Fagerberg L, Hallstrom BM, et al. Tissue-based map of the human proteome. Science. 2015;347:1260419

Describes the first map of the human tissue proteome based on transcriptomics and antibody-profiing.

[PubMed: 25613900]

44. Carapito C, Lane L, Benama M, et al. 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;14:3621–3634. [PubMed: 26132440]
45. Jumeau F, Com E, Lane L, et al. Human Spermatozoa as a Model for Detecting Missing Proteins in the Context of the Chromosome-Centric Human Proteome Project. J. Proteome Res 2015;14:3606–3620. [PubMed: 26168773]

46. Fan Y, Zhang Y, Xu S, et al. Insights from ENCODE on Missing Proteins: Why β-Defensin Expression Is Scarcely Detected. J. Proteome Res 2015;14:3635–3644. [PubMed: 26258396]
47. Meyfour A, Ansari H, Pahlavan S, Mirshahvaladi S, et al., Y Chromosome Missing Protein, TBL1Y, May Play an Important Role in Cardiac Differentiation. J Proteome Res 2017 9 13. doi: 10.1021/acs.jproteome.7b00391. [Epub ahead of print] PubMed PMID: . [PubMed: 28853286]
48. Mohamedali A, Ahn SB, Sreenivasan VKA, Ranganathan S, Baker MS. Human Prestin: A Candidate PE1 Protein Lacking Stringent Mass Spectrometric Evidence? J Proteome Res 2017 9 21. doi: 10.1021/acs.jproteome.7b00354. [Epub ahead of print] PubMed PMID: . [PubMed: 28895742]
49. Calviello L; Mukherjee N; Wyler E; Zauber H; Hirsekorn A; Selbach M; Landthaler M; Obermayer B; Ohler U, Detecting actively translated open reading frames in ribosome profiling data. Nat Methods 2016, 13, (2), 165–70.
50. Wang T; Cui Y; Jin J; Guo J; Wang G; Yin X; He QY; Zhang G, Translating mRNAs strongly correlate to proteins in a multivariate manner and their translation ratios are phenotype specific. Nucleic Acids Res 2013, 41, (9), 4743–54.
51. Verheggen K; Volders PJ; Mestdagh P; Menschaert G; Van Damme P; Gevaert K; Martens L; Vandesompele J, Non-coding after all: Biases in proteomics data do not explain observed absence of lncRNA translation products. J Proteome Res 2017.
52. Slavoff SA; Mitchell AJ; Schwaid AG; Cabili MN; Ma J; Levin JZ; Karger AD; Budnik BA; Rinn JL; Saghatelian A, Peptidomic discovery of short open reading frame-encoded peptides in human cells. Nat Chem Biol 2013, 9, (1), 59–64.
53. Chang C, Li L, Zhang C, et al. Systematic analyses of the transcriptome, translatome, an proteome provide a global view and potential strategy for the C-HPP. J Proteome Res. 2014;13(1):38–49. [PubMed: 24256510]
54. Vakilian H, Mirzaei M, Sharifi Tabar M, et al. DDX3Y, a Male-Specific Region of Y Chromosome Gene, May Modulate Neuronal Differentiation. J Proteome Res. 2015 9 4;14(9):3474–8341. [PubMed: 26144214]
55. Yen CY, Houel S, Ahn NG, et al. Spectrum-to-spectrum searching using a proteome-wide spectral library. Mol. Cell. Proteomics. 2011;10:M111.007666.
56. Cho JY, Lee HJ, Jeong SK, et al. Combination of Multiple Spectral Libraries Improves the Current Search Methods Used to Identify Missing Proteins in the Chromosome-Centric Human Proteome Project. J. Proteome Res. 2015;14:4959–4966. [PubMed: 26330117]
57. Wang BH, Reisman S, Bailey M, et al. Peptidomic profiles of post myocardial infarction rats affinity depleted plasma using matrix-assisted laser desorption/ionization time of flight (MALDI-ToF) mass spectrometry. Clin Transl Med. 2012 6 15;1(1):11. doi: 10.1186/2001-1326-1-11. [PubMed: 23369288]
58. Paulo JA, Gaun A, Kadiyala V, et al. Subcellular fractionation enhances proteome coverage of pancreatic duct cells. Biochim Biophys Acta. 2013 4;1834(4):791–7. [PubMed: 23352835]
59. Cox B, Emili A. Tissue subcellular fractionation and protein extraction for use in mass-spectrometry-based proteomics. Nat Protoc. 2006;1(4):1872–8. [PubMed: 17487171]
60. Moreda-Piñeiro A, García-Otero N, Bermejo-Barrera P. A review on preparative and semi-preparative offgel electrophoresis for multidimensional protein/peptide assessment. Anal Chim Acta. 2014 7 11;836:1–17. doi: 10.1016/j.aca.2014.04.053. [PubMed: 24974865]
61. Segura V, Garin-Muga A, Guruceaga E, Corrales FJ. Progress and pitfalls in finding the ‘missing proteins’ from the human proteome map. Expert Rev Proteomics. 2017 1;14(1):9–14. doi: 10.1080/14789450.2017.1265450. Epub 2016 Dec 2. PubMed PMID: . [PubMed: 27885863]
62. Cho JY, Lee HJ, Jeong SK, et al. Epsilon-Q: An Automated Analyzer Interface for Mass Spectral Library Search and Label-Free Protein Quantification. J. Proteome Res 2017(in press)
63. Chen Y, Li Y, Zhong J, et al. Identification of Missing Proteins Defined by Chromosome-Centric Proteome Project in the Cytoplasmic Detergent-Insoluble Proteins. J. Proteome Res 2015;14:3693–3709. [PubMed: 26108252]
64. Eckhard U, Marino G, Abbey SR, Tharmarajah G, Matthew I, Overall CM, 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, 2 22;7:299–310.

65. Wisniewski ES, Rees DK, Chege EW. Proteolytic-based method for the identification of human growth hormone. J Forensic Sci. 2009 1;54(1):122–7. [PubMed: 19120827]

66. Menon R, Panwar B, Eksi R, et al. Computational Inferences of the Functions of Alternative/Noncanonical Splice Isoforms Specific to HER2+/ER–/PR–Breast Cancers, a Chromosome 17 C-HPP Study. J. Proteome Res 2015;14:3519–3529. [PubMed: 26147891]

67. Nesvizhskii AI. Proteogenomics: concepts, applications and computational strategies. Nat. Methods 2014;11:1114–1125. [PubMed: 25357241]

68. Gu X, Karp PH, Brody SL, et al. Chemosensory functions for pulmonary neuroendocrine cells. A. J. Respir. Cell Mol. Biol 2014;50:637–646.

69. Hallem EA, Dahanukar A, Carlson JR. Insect odor and taste receptors. Annu. Rev. Entomol 2006;51:113–135. [PubMed: 16332206]

70. Niimura Y Evolutionary dynamics of olfactory receptor genes in chordates: interaction between environments and genomic contents. Hum. Genomics 2009;4:107–118. [PubMed: 20038498]

71. Gilad Y, Lancet D. Population differences in the human functional olfactory repertoire. Mol. Biol. Evol 2003;20:307–314. [PubMed: 12644552]

72. Gaillard I, Rouquier S, Giorgi D. Olfactory receptors. Cell. Mol. Life Sci 2004;61:456–469. [PubMed: 14999405]

73. Baker MS, Ahn SB, Mohamedali A, et al. Accelerating the search for the missing proteins in the human proteome. Nat. Commun 2017;8:14271. [PubMed: 28117396]

74**. Paik YK, Hancock WS. Uniting ENCODE with genome-wide proteomics. Nat. Biotechnol 2012;30:1065–1067.

Describes how ENCODE data and C-HPP can be integrated to define the full complexity of proteogenomic information in different biological states.

[PubMed: 23138303]

75. Na K, Shin H, Cho JY, et al. A systematic proteogenomic approach to exploring a novel function of NHERF1 involved in human reproductive disorder: lessons for exploring missing proteins. J. Proteome Res (in press).

76. Paik YK, Overall CM, Deutsch E, et al., Progress and future direction of chromosome-centric human proteome project, J. Proteome Res. (being submitted; out on 12 1, 2017)

Box 1

Start with HPP:

C-HPP: An international project which deals with mapping and characterizing the human proteome in a chromosome-by-chromosome manner using various omics technologies.

B/D-HPP: An international project which studies the human proteome with respect to the biology and disease using various organ tissues, bio-fluids and cell lines.

Biomarker: A measurable indicator (gene, protein) of some biological state/ condition.

FDR: False discovery rate: an estimate of how likely a particular protein ID is to be the result of random matching, rather than a “true” ID (<23451.0% at both peptide and protein level is being employed by HPP)

Isoform: Any of different forms of the same protein that may be produced from very closely related gene duplicates or by alternative splicing

Missing Proteins: Those gene-encoded proteins with less confident MS evidence or inadequately annotated

PE Levels (PE-1 to PE5): a degree of evidence for protein existence used in UniProtKB and neXtProt

PE1: evidence at protein level (e.g. clear identification by mass spectrometry)

PE2: evidence at transcript level (e.g. the existence of cDNA)

PE3: inferred by homology (assigned membership of a defined protein family)

PE4: predicted (not yet been assigned membership of a defined protein family)

PE5: uncertain (e.g. dubious sequences by such as erroneous translation products)

Key issues

- How best can we improve methodological approaches to find low abundance MPs?
- More important challenge than ever would be to establish an integrated tool box and workflow which illustrate all innovative methods from enrichment, detection, quantification and functional characterization of the remaining 2570 MPs and 1232 uPE1.
- What would be an elegant way to distinguish authentic unique peptides of the MPs from their potential SSAVs or other isoforms?
- How can an amplicon contribute to increase of human protein types and numbers? If so how can we detect and capture those amplicon products efficiently?
- How can obtain rarely used biological samples for discovery and characterization of MPs?
- How best we can integrate those datasets from characterizing MPs such as alternative splice variants (ASVs), lncRNA, RNC RNAs into the current proteome DB?
- As we approach to a new phase of the long-term C-HPP program, how can we enhance the capabilities and deliverables of the C-HPP teams?
- What would be an efficient way of cross-chromosome analysis for annotating entire MPs.
- What would be most challenging problems in integrating whole human proteome?

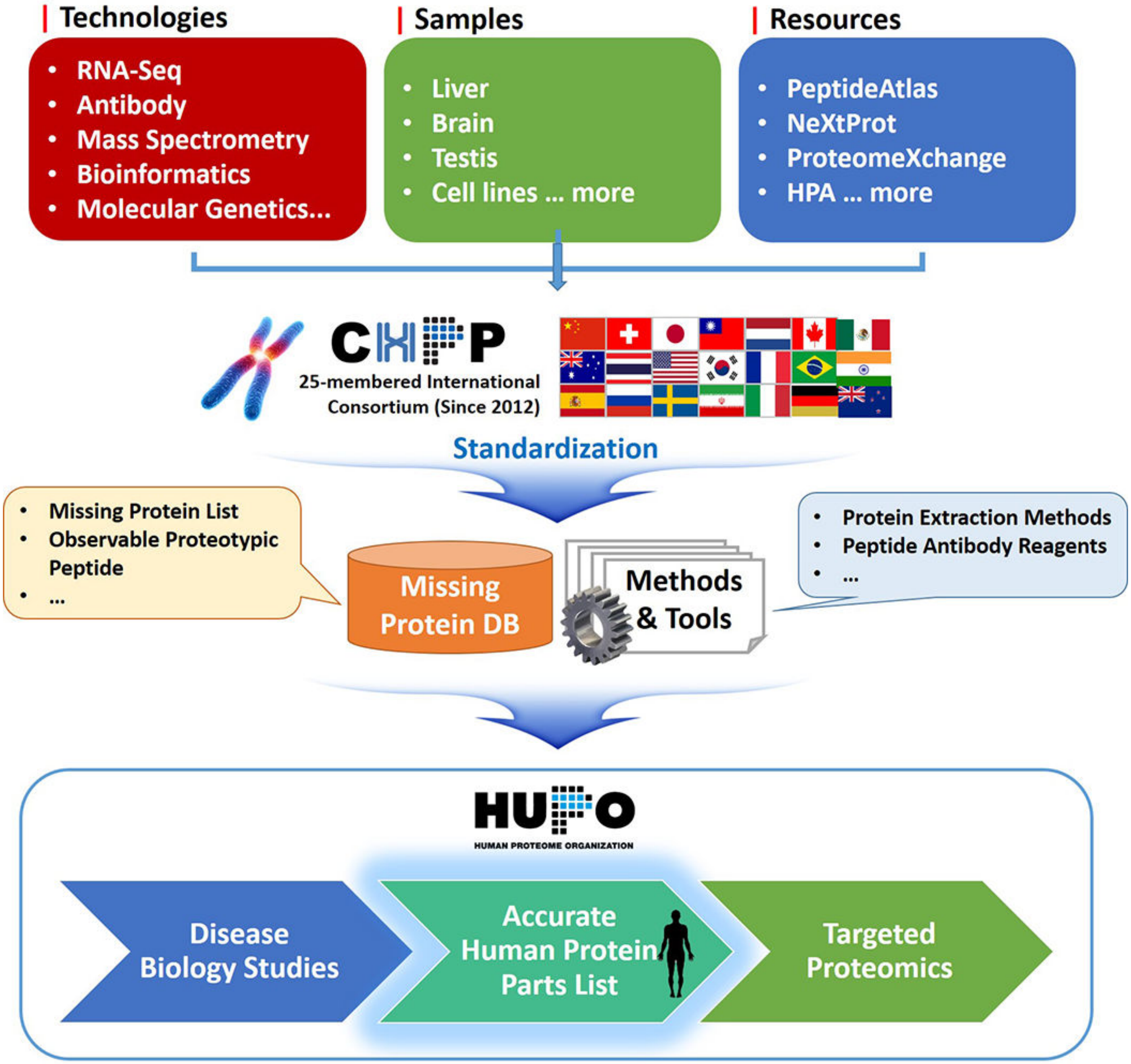

Figure 1.

Overall concept of C-HPP. Shown here are the key components of C-HPP endeavors in the areas of technology, samples, resources, team works, standard guidelines and potential deliverables. National flags represent all those hosting countries participated in the early times as well as in the current stage.

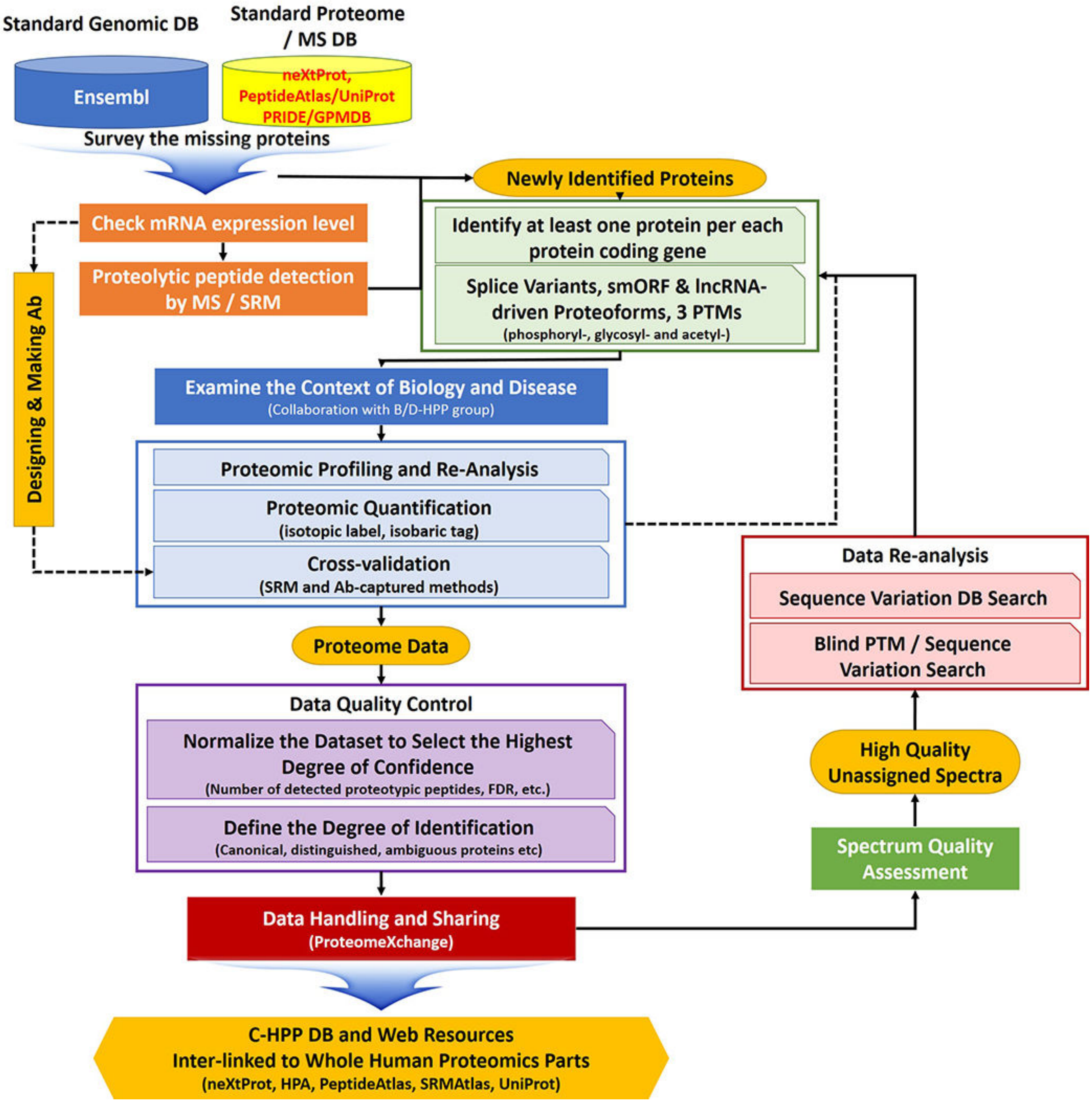
Standard Genomic DB
Standard Proteome / MS DB
Ensembl
neXtProt, PeptideAtlas/UniProt PRIDE/GPMDB
Survey the missing proteins
Newly Identified Proteins
Identify at least one protein per each protein coding gene
Splice Variants, smORF & lncRNA-driven Proteoforms, 3 PTMs
(phosphoryl-, glycosyl- and acetyl-)
Check mRNA expression level
Proteolytic peptide detection by MS / SRM
Designing & Making Ab
Examine the Context of Biology and Disease
(Collaboration with B/D-HPP group)
Proteomic Profiling and Re-Analysis
Proteomic Quantification
(isotopic label, isobaric tag)
Cross-validation
(SRM and Ab-captured methods)
Proteome Data
Data Quality Control
Normalize the Dataset to Select the Highest Degree of Confidence
(Number of detected proteotypic peptides, FDR, etc.)
Define the Degree of Identification
(Canonical, distinguished, ambiguous proteins etc)
Data Handling and Sharing
(ProteomeXchange)
Data Re-analysis
Sequence Variation DB Search
Blind PTM / Sequence Variation Search
High Quality Unassigned Spectra
Spectrum Quality Assessment
C-HPP DB and Web Resources
Inter-linked to Whole Human Proteomics Parts
(neXtProt, HPA, PeptideAtlas, SRMAtlas, UniProt)

Figure 2.

The C-HPP Workflow. The flow of proposed standard working procedures, coordination between data management and potential outputs from the community-based research efforts (modified from Fig. 1 in ref. [3]).

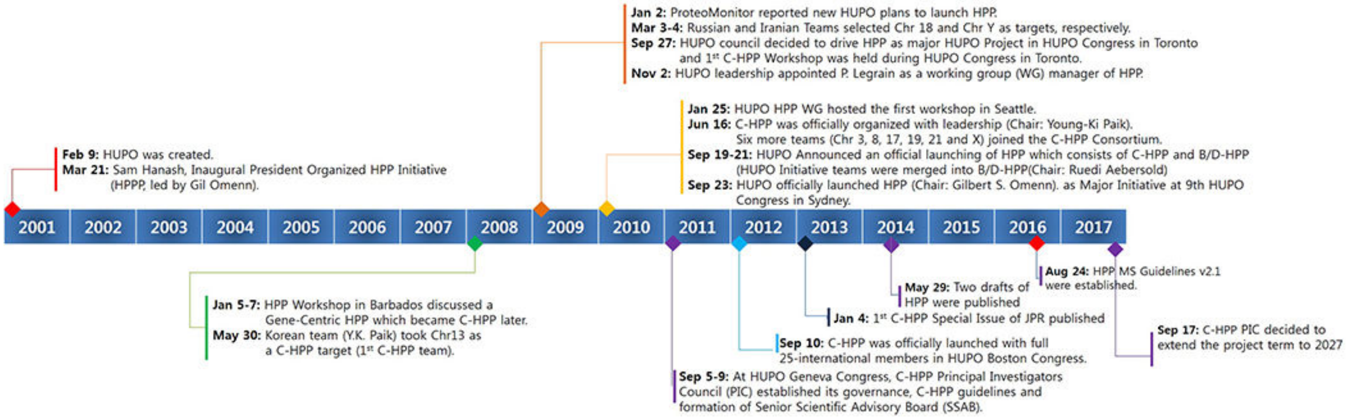
Jan 2: ProteoMonitor reported new HUPO plans to launch HPP.
Mar 3-4: Russian and Iranian Teams selected Chr 18 and Chr Y as targets, respectively.
Sep 27: HUPO council decided to drive HPP as major HUPO Project in HUPO Congress in Toronto and 1st C-HPP Workshop was held during HUPO Congress in Toronto.
Nov 2: HUPO leadership appointed P. Legrain as a working group (WG) manager of HPP.
Jan 25: HUPO HPP WG hosted the first workshop in Seattle.
Jun 16: C-HPP was officially organized with leadership (Chair: Young-Ki Paik). Six more teams (Chr 3, 8, 17, 19, 21 and X) joined the C-HPP Consortium.
Sep 19-21: HUPO Announced an official launching of HPP which consists of C-HPP and B/D-HPP (HUPO Initiative teams were merged into B/D-HPP(Chair: Ruedi Aebersold)
Sep 23: HUPO officially launched HPP (Chair: Gilbert S. Omenn). as Major Initiative at 9th HUPO Congress in Sydney.
Feb 9: HUPO was created.
Mar 21: Sam Hanash, Inaugural President Organized HPP Initiative (HPPP, led by Gil Omenn).
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
Aug 24: HPP MS Guidelines v2.1 were established.
May 29: Two drafts of HPP were published
Jan 4: 1st C-HPP Special Issue of JPR published
Jan 5-7: HPP Workshop in Barbados discussed a Gene-Centric HPP which became C-HPP later.
May 30: Korean team (Y.K. Paik) took Chr13 as a C-HPP target (1st C-HPP team).
Sep 17: C-HPP PIC decided to extend the project term to 2027
Sep 10: C-HPP was officially launched with full 25-international members in HUPO Boston Congress.
Sep 5-9: At HUPO Geneva Congress, C-HPP Principal Investigators Council (PIC) established its governance, C-HPP guidelines and formation of Senior Scientific Advisory Board (SSAB).

Figure 3.

Chronicle of C-HPP Development. Only key events that are relevant to the C-HPP are marked in the timeline.

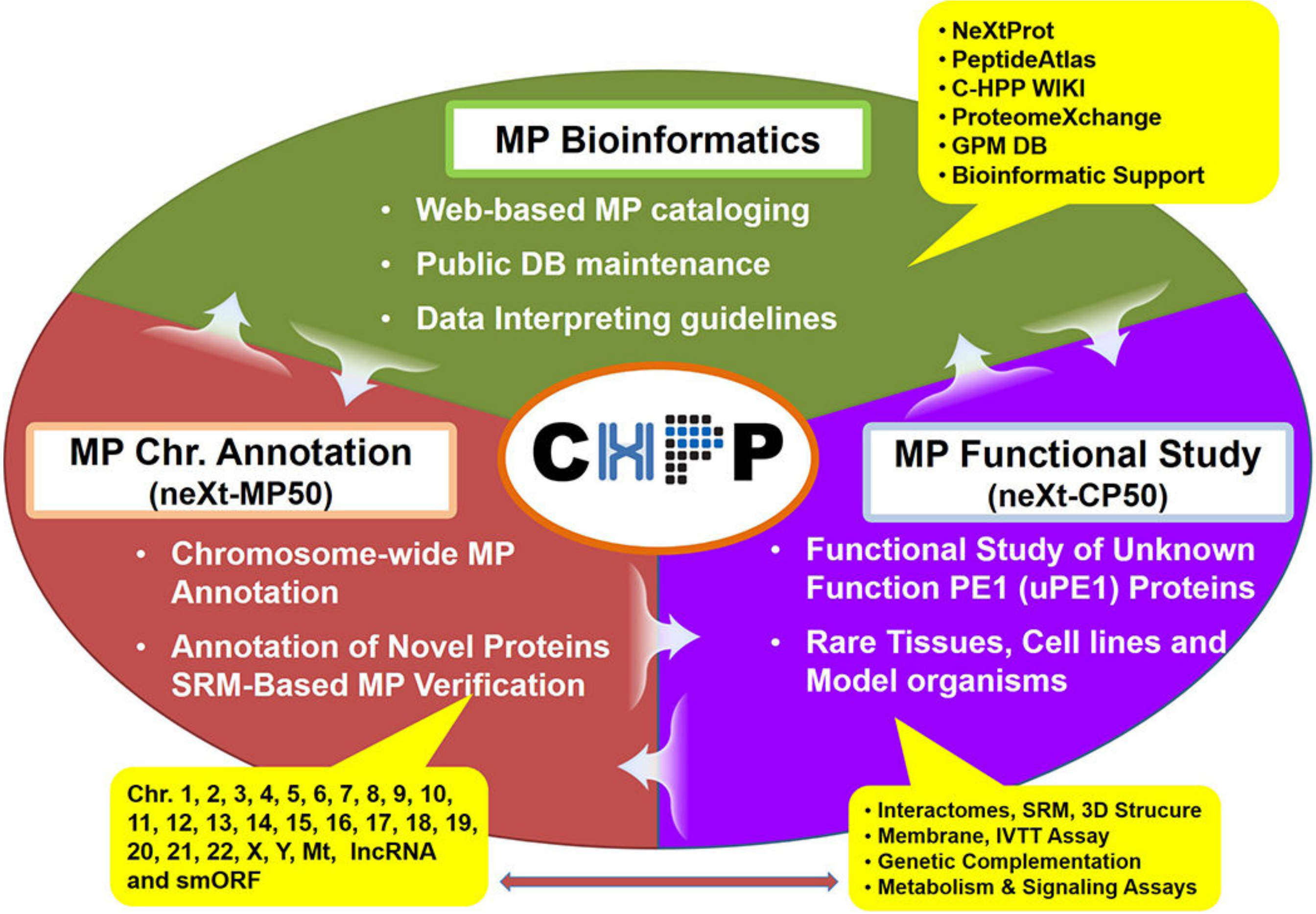
• NeXtProt
• PeptideAtlas
• C-HPP WIKI
• ProteomeXchange
• GPM DB
• Bioinformatic Support
MP Bioinformatics
• Web-based MP cataloging
• Public DB maintenance
• Data Interpreting guidelines
C-HPP
MP Chr. Annotation
(neXt-MP50)
• Chromosome-wide MP Annotation
• Annotation of Novel Proteins SRM-Based MP Verification
MP Functional Study
(neXt-CP50)
• Functional Study of Unknown Function PE1 (uPE1) Proteins
• Rare Tissues, Cell lines and Model organisms
Chr. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, X, Y, Mt, lncRNA and smORF
• Interactomes, SRM, 3D Strucure
• Membrane, IVTT Assay
• Genetic Complementation
• Metabolism & Signaling Assays

Figure 4.

Triad of the C-HPP working group with new interactive corporation module. Shown are the three teams with different goals and approaches toward draft of accurate proteome map.

Table 1.

Possible Causes of Missing Proteins Absence and Solutions This information was collected from the literatures. Difficulties in detection of missing proteins would not be due to a single problem but rather be caused by multiple reasons. Listed are only some representative cases for difficulties in detection of missing proteins.

Possible causes	Strategy for Detection and Analysis	References
Low abundance proteins produced by routine genomic activities (e.g., RNC, alternative splicing, cell-type specific gene regulations, SNPs) in normal or disease state.	·Improved tissue/cell selection guided by RNA Seq analysis, Proteogenomic analysis and proof. (splice variants). Use PGNexus tool (Tay et al.,)	Chang C. et al., 2014 [42], Vakilian et al., 2015 [43]
Limitation of search method for mass spectral data (e.g., low sensitivity and potential errors in the handling of low-quality experimental spectra; absence of proper DBs (e.g., SNP) and others.	·Development of advanced software for spectral search methods. ·Use the High-Capacity Sequence DBs ·Match with Theoretical Peptide Spectra (Anatomy of a six-frame search) ·Protein identification using customized protein sequence databases derived from RNA-Seq Data ·Reanalysis of those unassigned MS spectra that might have been unidentified in one search engine ·Combined public spectral library (NIST, ISB) and in-house spectral library with non-redundancy. ·Use of ENCODE data ·Epigenetic manipulation Amplicons	Yen et al., 2011 [44]; Cho et al., 2016 [45,51] Wang et al., 2012 [46] Fan et al., 2015 [41] Yang et al., 2015 [28]
Very low abundance of proteins and loss of proteins due to incomplete separation.	·Systematic Subcellular Fractionation and Diverse Separation Methods ·Off-gel separation ·High/Low pH Reverse Phase LC separation and Diverse column (e.g, SCX, HILIC) separation ·Use of a single shot analysis	Paulo et al. 2013 [47] Cox and Emili, 2006 [48] Pineiro et al., 2014 [49]
Rare proteins produced only at specific times or in specific cells.	·Use of detergent insoluble cytoplasmic proteins fractions ·Use of rare samples (e.g., stem cells, specific type of tissues-human dental pulp, spermatozoa), guided by RNA-Seq	Chen et al., 2015 [52] Jumeau et al., 2015 [40] Eckhard et al., 2015 [53]
Hydrophobicity of proteins (e.g., membrane proteins, highly insoluble proteins).	·Enrichment and clean-up for MS analysis ·Sample clean-up for Label Free Quantitation MS analysis ·Optimized and improved by filter aided sample preparation ·Improved analytical system (LC-SRM; High pH RP Stage Tip Fractionation etc.)	Wiśniewski et al., 2009 [54] Horvatovich et al., 2015 [55]
Experimental error occurs during sample preparation: Sample loss Desalting/Enrichment, Contamination Trypsin autolysis peptides, hair, skin keratins, bacterial contamination Matrix molecules, clusters, unknown contaminants Enzyme Missed/non-specific cleavage	·Automatic sample control ·Optimized sample preparing system; Reproducibility	