pr-2011-00824a.R2 (March 3, 2012)-Accepted March 9, 2012

Standard Guidelines for the Chromosome-Centric Human Proteome Project

Young-Ki Paik[1,], Gilbert S. Omenn[2], Mathias Uhlen[3], Samir Hanash[4], György Marko-Varga[5], Ruedi Aebersold[6], Amos Bairoch[7], Tadashi Yamamoto[8], Pierre Legrain[9], Hyoung-Joo Lee[1], Keun-Na[1], Seul-Ki Jeong[1], Fuchu He[10], Pierre-Alain Binz[7], Toshihide Nishimura[11], Paul Keown[12], Mark S. Baker[13], Jong Shin Yoo[14], Jerome Garin[15], Alexander Archakov[16], John Bergeron[17], Ghasem Hosseini Salekdeh[18], and William S. Hancock[1,19,*]*

[1]Yonsei Proteome Research Center, Yonsei University, Seoul, Korea, [2]Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA, [3]Royal Institute of Technology, Stockholm, Sweden, [4]Fred Hutchinson Cancer Research Center, Seattle, WA, USA, [5]Lund University, Lund, Sweden, [6]Department of Biology, Institute of Molecular Systems Biology, ETH, Zürich, and Faculty of Science, University of Zurich, Switzerland, [7]Swiss Institute of Bioinformatics (SIB) and University of Geneva, Geneva, Switzerland, [8]Graduate School of Medical and Dental Sciences, Niigata University, Niigata, Japan, [9]Ecole Polytechnique, Palaiseau, France, [10]BPRC, Beijing, China, [11]Department of Surgery I, Tokyo Medical University, Tokyo, Japan, [12]Univ. of British Columbia, Vancouver, Canada, [13]Macquarie University, NSW 2109, Australia,[14]Korea Basic Science Institute, Ochang, Korea. [15]CEA, Grenoble, France, [16]Russian Academy of Medical Sciences, Moscow, [17]McGill Univ. Montreal, Canada,[18]Royan Inst, Tehran, Iran, [19]Northeastern University, Boston, MA,

USA,

*To whom correspondence should be addressed:

Young-Ki Paik (paikyk@yonsei.ac.kr) at 82-2-2123-4242 (Tel) and 82-2-393-6589 or William S. Hancock (wi.hancock@neu.edu) at 1-617-869-8458 (Tel) and 617-373-2855 (Fax)

ABSTRACT

The objective of the international Chromosome-Centric Human Proteome Project (C-HPP) is to map and annotate all proteins encoded by the genes on each human chromosome. The C-HPP consortium was established to organize a collaborative network among the research teams responsible for protein mapping of individual chromosomes and to identify compelling biological and genetic mechanisms influencing co-located genes and their protein products. The C-HPP aims to foster the development of proteome analysis and integration of the findings from related molecular omics technologies platforms through collaborations among universities, industries, and private research groups. The C-HPP consortium leadership has elicited broad input for standard guidelines to manage these international efforts more efficiently by mobilizing existing resources and collaborative networks. The C-HPP guidelines set out the collaborative consensus of the C-HPP teams, introduce topics associated with experimental approaches, data production, quality control, treatment and transparency of data, governance of the consortium, and collaborative benefits. A companion approach for the Biology and Disease-Driven HPP (B/D-HPP) component of the Human Proteome Project is currently being organized, building upon the HUPO organ-based and biofluid-based initiatives (www.hupo.org/research). The

common application of these guidelines in the participating laboratories is expected to facilitate the goal of a comprehensive analysis of the human proteome.

Keywords

Antibody, Biology/disease-driven Human Proteome Project, Chromosome-Centric Human Proteome Project, Human Genome Project, Human Proteome Project, Human Proteome Organization, knowledge-base, mass spectrometry,

Abbreviations

Ab, antibody/affinity capture reagent; AST, alternative splicing transcript; C-HPP, chromosome-centric human proteome project; DMG, data management group; EAC, evaluation and assessment committee; EC, executive committee; HGP, human genome project; HUPO, human proteome organization; KB, knowledge-base; MS, mass spectrometry; nsSNP, non-synonymous single nucleotide polymorphism; PIC, principal investigators council; PTM, post-translational modification; RRG, reagent and resources group; SRM, selected reaction monitoring; TCG, technology consulting group.

INTRODUCTION

Since January 2008, the Human Proteome Organization (HUPO) has sponsored Human Proteome Project (HPP) workshops in Barbados, Amsterdam, Seoul, Moscow, Seattle, Montreal, Sydney, Busan, and Geneva and has held discussions at national and regional meetings throughout Asia/Oceania, Europe, and North America (www.hupo.org/research/hpp). HUPO has now officially launched the HPP to map the entire human proteome in a systematic effort using three major pillars: mass spectrometry (MS), antibody/affinity capture reagents (Ab), and bioinformatics-driven knowledge base (KB).[1] The Chromosome-Centric HPP (C-HPP) is one component of the HPP and focuses on constructing the proteomic catalog in a chromosome-by-chromosome fashion.[2,3] With the availability of a well-characterized human genome map, in-depth transcriptome data, and major advances in proteomics technologies and databases, the proteomics community realized that the foundations were already in place to study the full complexity of the human proteome. In the C-HPP, each team will perform quantitative measurement by MS approaches and Ab-based tissue staining of the representative proteins from each protein-coding gene in at least 4 standardized tissue samples of interest (e.g., liver, brain, heart and placenta)[4]. The C-HPP is complementary to a broad array of biology- and disease-driven studies, collectively termed the B/D-HPP, including the existing organ-based and biofluid-based HUPO initiatives (www.hupo.org/research).[5-11] The objective of linking the C-HPP and the B/D-HPP will be expedited by combining the resources of multiple research groups from around the world.

A primary goal of the C-HPP is to identify and characterize proteins that currently lack MS evidence or Ab detection, termed "missing proteins".[3] There are numerous reasons for the lack of quality observations of a given protein, such as incorrect gene annotation, very low abundance, absence of expression in a given tissue, expression only in rare samples, and unfavorable structure (or cleavage sites) for MS studies such as instability or heterogeneity. We will employ a variety of measurement approaches, including RNA sequencing (RNA-Seq) to guide selection of appropriate samples as well as the preparation of recombinantly expressed protein standards and heavy-labeled proteotypic SRM

peptides[11] to ensure that the discovery effect is as complete as possible. This venture represents an organizational innovation to mobilize collaborative efforts on major subsets of proteins and on potential biological features associated with the chromosomal locations of their genes. The C-HPP represents feasible sub-projects that have excited national or international teams of proteomics scientists and their national agencies. A systematic and integrated transcriptome/proteomic measurement will be employed for each previously defined as missing proteins. These studies will be integrated with results from the parallel B/D-HPP program. This work will also enhance the characterization of unexplored transcripts that appear to be protein-coding but so far have no evidence of translation. As with the Human Genome Project (HGP), comparative proteomics studies across species will be encouraged, including recognition of synteny or homology among chromosomes of model organisms (e.g., *C. elegans*, fruit fly, mouse etc).

To initiate the C-HPP, HUPO established a consortium and its initial leadership during the HPP workshop held in Busan, Korea, March 30, 2011. As of January 1, 2012, there are 15 international teams focused on 14 different chromosomes.[2,3] The C-HPP will promote chromosome-based protein mapping world-wide and will capture special biological features of gene variation, gene regulation, and protein expression coordinated at the chromosomal level. Of course, many MS- and Ab-based proteomics studies will continue to generate data at a global (all chromosomes) level. The C-HPP project will organize and report data in a format that is aligned with the DNA sequence of individual chromosomes. In this manner we propose to align the proteomics data set with the output of RNA-Seq and other emerging genomic technologies.

Thus, it is timely to present this proposal for standard C-HPP guidelines for which will be updated by the scientific community to embrace rapid developing technologies and specific demands from the proteomics community as the C-HPP moves forward. The January 2011 issue of JPR[2] announced that this journal and the American Chemical Society will support the establishment of C-HPP as part of HPP, the new scientific initiative of HUPO. One aspect of this support is to publish

articles, which may well include new types of integrative data presentations that support the progress of the HPP and C-HPP.

Mission, Strategy, and Deliverables of the C-HPP Program

The C-HPP consortium will 1) organize and perform chromosome-centric protein identification and characterization using carefully-selected human samples (mission), 2) construct a web of bioinformatics for the C-HPP (including existing platforms such as PRIDE, GPMDB, PeptideAtlas, and UniProt) and emerging initiatives (such as NextProt, ProteomeXchange, Tranche, and HUPO PSI)(strategy), and 3) integrate information derived from the C-HPP with results from the study of cellular and biochemical processes, as well as detailed protein chemistry characterization that will be provided by the B/D-HPP program (deliverables) (Fig. 1 and Supplemental Fig. S1).

To operationalize the C-HPP, we need community-driven standards that can enhance the prospects of successful funding and promote dissemination of the human proteome knowledge base. Examples of interdisciplinary participants in the C-HPP include biologists, chemists, computer scientists, bioinformaticians, biostatisticians, clinicians, pathologists, and epidemiologists who are involved in building community-based standards for proteomic databases and conducting large-scale comparative protein analyses in the context of human disease and diversity.[1,3]

Collaboration within the C-HPP will be based on shared objectives and consensus regarding the best ways to identify the currently missing proteins for each chromosome and characterize all human proteins by the three pillars of MS, Ab, and KB. The consortium will identify proteins from all 24 chromosomes, enabling each C-HPP team to focus on those mapping to its particular chromosome. Collaboration is essential for the success of this project due to the nature and the large scale of work involved in protein mapping, characterization, and tissue localization.[3,4] Following the initial C-HPP team from Korea, 14 additional groups have joined to take a specific chromosome. At least 24 teams are needed to cover the full set of chromosomes, with real possibility that different arms of large

chromosomes may be investigated separately. Different groups may also undertake complementary approaches to the same chromosome, especially if their organ/disease expertise is different (e.g., chromosome 7 for two countries, chromosome 19 for 5 countries). Each team has its own biological interests (e.g., metabolic disease for chromosome 13 by Korea, cancer for chromosome 17 and 7 by US and Australia/New Zealand, liver disease for chromosome 1 by China). Although we organize all those protein mapping data in a chromosome by chromosome approach, we will pursue a chromosome-independent shotgun approach first and then the targeted proteomics using SRM in looking for missing proteins. Once data are collected, they will be shared according to chromosome number in order to ensure complete parts lists. Individual teams must therefore share the full data so that comprehensive protein parts list will eventually be completed.

For potential deliverables, this project takes advantage of MS-based assays for more than two representative proteotypic peptides of an entire coding protein and provides a reference set for the comprehensive quantitative coverage of the human proteome.[4]

A Decade-Long Plan and Short- and Long-Term Challenges

We believe that high-quality, extensive proteome maps are achievable within a planned 10-year period. As outlined in Table 1, during Phase 1 (6 years), the C-HPP group plans to map all proteins lacking good quality MS evidence, three major classes of PTMs, one representative alternative splicing transcription (AST) product[13] and one non-synonymous SNP product, and protein distribution in a major organ/tissue of interest. C-HPP will utilize all high-quality consortium-generated proteomic datasets for focused analysis on individual chromosomes. In phase II (4 years), identified proteins will be further characterized and validated at the genomic/transcriptomic and cellular levels in the 4 selected tissues of interest. C-HPP outputs will also be integrated with all biology/disease-driven HPP research. We will also provide a correlation of C-HPP and B/D HPP study results with recent SNP and haplotypic mapping studies.[14]

As outlined in Table 1, there are a number of short-term challenges including cost-effective technology development, improved cross analysis and integration of different datasets, handling of data variability, procurement of Abs, and harmonization with B/D-HPP. Potential solutions for these challenges may include sharing resources, data, reagents (e.g., Ab) and reference specimens, employing neXtProt, dbSNP, GPMDB and PeptideAtlas, standard data submission system/criteria, close collaborations between Ab providers and C-HPP groups, and sharing data through the C-HPP portal and other public biology/disease DBs. Longer-term challenges are more difficult to predict but will include sample bio-banking and maintenance, inclusion of complex PTM information in different datasets, enhanced detection limit for low abundance (rare) proteins and improved pretreatment of clinical specimens for characterization and SRM analysis. Potential solutions for the longer-term challenges may include collaboration with government agencies for stable bio-banking resources, development of new algorithms for inclusion of PTMs in the biological databases, miniaturization of sample preparation and multiplexed fractionation steps.[14]

Thus, this document sets out the general guidelines and the collaborative consensus of the C-HPP consortium. The C-HPP guidelines introduce topics associated with experimental approaches, data production with quality control, treatment of data, governance of the consortium, and collaborative benefits. An important part of establishing such guidelines will be the review of proposed procedures by a Senior Scientific Advisory Board (SSAB) that has been formed to include many eminent proteomic scientists for both C-HPP and B/D-HPP programs. We expect that over time it will be necessary to update the guidelines to incorporate changes in areas such new gene annotation and transcriptome variations, evolution in proteomic technologies and bioinformatics.

Working Strategy

Stage 1. Experimental procedures for data production (Steps 1-5)

Fig. 1 outlines the overall working strategy of a typical C-HPP which can be divided into two stages,

data production and data management. Once the targets and scope of the C-HPP are determined by the consortium under the aegis of the HPP, the individual teams can determine the biological aims, particular targets, and scope of work for each team's specific chromosome according to consensus built by the consortium.

Step 1 is to make a list of "missing proteins" using the several DBs (e.g., UniProt, Ensembl, GPMDB) by cross checking with an entire list of protein coding genes. At the same time, another effort will be made to improve the quality of mass spectrometric identifications in all but the highest probability category. "Missing proteins" are defined as those proteins which have only transcriptomic evidence and a predicted sequence (or are inferred by homology), or those partially-identified proteins, one in which there is transcript evidence for the existence of the corresponding protein is available without convincing MS information. The development of a list of "missing proteins" and follow-up work pose a question as to whether it is a technical issue (unsuitable protein physical properties or enzyme cleavage sites) or the fact that the gene identification is problematic or has not been deposited in the appropriate databases. Each of the Ensembl accession numbers is linked to the evidence available for that particular protein. A potential target protein list can be made by selecting a specific functional group of interest (e.g., onco-proteins) from those compiled protein pool including the missing proteins.

Step 2 will be to obtain specific mRNA expression pattern by RNA-seq and reverse transcription-polymerase chain reaction, based on public databases (GeneCards [www.genecards.org] and dbEST [www.ncbi.nlm.nig.gov/dbEST]) with defined expression thresholds. For these transcriptomic analyses, we will work with a group of cell biologists who have various specific cell lines and stem cells which may provide very unusual, rarely expressed proteins that are hard to detect under normal culture conditions. At the consortium level, a pool of these unique cell cultures will be established and then each team would acquire additional cell lines of interest. Using these cell lines, we will analyze those target protein data obtained from the high resolution mass spectrometry (e.g., Thermo Orbitrap, ABSIEX Triple TOF etc) with the aid of proteome informatics and validate these

proteins with appropriate antibody and SRM approach (SRMAtlas). In collaboration with genomics group, each team will perform RNA-seq analysis using the given samples and then match with the RNA-seq data with newly identified proteins as well as known proteins for the given specimens. This approach will provide considerable reference information on each missing protein in a given sample (or reference sample). Analyses of these "missing proteins" can then be performed using recombinant proteins and mass spectrometry to produce peptide signatures that can be used for cross-checking with SRM library.[5] This can be further verified by antibodies at the cellular or molecular level using Western blotting or immune-cytological analyses using the Abs that were raised against the synthetic peptides of the proteins of interest. No matter what sample will be used, eventually all missing proteins will be identified by various chromosome groups and cataloged in a concerted way. The target samples selected by each chromosome group will be determined by the focus of their biological/disease studies as well as the patterns of tissue expression of hard-to-characterize proteins.

Step 3 will be to characterize at least one representative isoform and three major translational modifications (PTMs) (i.e., phosphoryl-, glycosyl-, and acetyl-) for each protein. While this goal will not define all modified residues in any given protein due to expected technical and stability issues or other PTMs we expect that such data will be useful to guide focused follow-up studies by protein chemists. The C-HPP project will also examine the occurrence of these three major PTMs in splice isoforms, and variants with non-synonymous single nucleotide polymorphisms.

Step 4 will be to explore the annotation and disease related context of newly identified proteins in collaboration with the B/D project. To facilitate these studies, C-HPP will coordinate with the government-funded clinical specimen banks (e.g., NIH-funded tissue reference bank) for a repository of paired samples (normal vs. disease) which will be made available to HPP participants.

Step 5 will be to perform the validation works including proteomic profiling with re-identification, quantifications and cross-validation (SRM and Ab captured methods). Characterization of identified missing proteins and their representative isoforms, PTMs, nsSNPs and splice variants will

be verified by MS and Ab. Knowledge of RNA expression, biological function, isoform identification, and disease relation must be included, which accompanies a stringent validation process. To this end, first, confirmation and replication must be performed by the same team; independent validation should be performed with entirely fresh samples and preferably a different laboratory. These procedures will be standardized and supervised.

While Ensembl will be used as our standard genomic database, there is considerable diversity in the proteomic profiling methods used by different laboratories. Varying methods exist for enzymatic digestion (e.g., Trypsin, GluC, LysC, and AspN)[14], predigestion fractionation (e.g., OFFGEL, IPG-IEF, hpRP, SCX, 1D gel, and extensive fractionation of intact proteins)[16], identification (high resolution MS, Orbitrap or Triple TOF)[17], quantification using labeling methods (e.g., iTRAQ, mTRAQ, and TMT),[18] and selected reaction monitoring (SRM) (e.g., Triple-Quadrupole type MS/MS: QTRAP).[19] For both protein profiling and re-identification, a selected reaction monitoring (SRM) peptide database will be established, consisting of experimentally-observed peptides from liquid chromatography tandem mass spectrometry (MS/MS) experiments with SRM peptide standards.[19] Alternative detection methods will be required to cross-validate SRM and Ab captured methods.[20] With its goals of identifying proteotypic peptides, the C-HPP group will also employ SRM assays and AB affinity reagents in collaboration with several other ongoing proteomics initiatives such as the SRMAtlas (www.srmatlas.org) and the Human Protein Atlas (www.proteinatlas.org) (Supplemental Table S1).[9,21] The existing databases will be assembled in the context of missing or poorly-characterized proteins, as well as for the further characterization of already well-identified proteins. In particular, SRMAtlas would enhance this process. These will be studied within the category of the chromosome parts list (Table 1), which can be used as a format for the C-HPP to interact in new ways with HUPO initiatives.

For the quantitation of proteins by tandem mass spectrometry, C-HPP teams will employ a typical peptide-centric analysis approach using heavy isotope-labeled synthetic peptides for each proteotypic peptide.[22] For quality control of the dataset produced by each C-HPP group, we can apply a

general principle for data acquisition. For example, for MS/MS results searched by Mascot,[23] only top-ranked results would be retained, and the same top-ranked results would be merged for the same query as a peptide group with Mascot identity threshold ($p<0.05$). Any peptides with a mass tolerance within 10 ppm will be further curated in proteomic databases such as the GPMDB for inclusion or not as parts for improving the quality of MS identification of known proteins or beginning to assemble a parts list for missing proteins.

The Parsimony method will be used for obtaining a non-redundant protein list and the criteria set for selection of the representative protein and better annotation record, for example, according to the source database for the IPI record.[24] In fact, Parsimony and marked reduction of redundant protein matches is the basis of the new Cedar scheme for the PeptideAtlas.[25] The "canonical" list (and "covering" list) of proteins from any dataset or from a whole PeptideAtlas is greatly reduced from the "sequence-specific," "peptide-unique," and, "possibly-distinguished" lists. For the peptide identification on a large scale, we would select highly confident peptides (e.g., FDR < 1 %) through intensive statistical analyses. Since Peptide Atlas Cedar scheme uses 1% FDR for proteins, hence we want to set the similar FDR for peptides in our experiment. In particular, for identification of PTM peptides within each peptide sequence, all modifications must be clearly located (unless ambiguous). For the case of quantitative analysis, using the technical replicates and statistical methods, we can use only validated results for data set production.

neXtProt will be one of tools of choice used by the C-HPP to integrate data from the Human Protein Atlas and three other resources (PeptideAtlas, SRMAtlas, and UniProt) and to provide subcellular information and the number of PTMs from good quality, high-throughput studies. This system should be able to provide some advanced tools such as an advanced search system, export options to extract proteomics-relevant data, and a tool to analyze and "cluster" lists of proteins (for example, those originating from MS identification runs).

Stage 2: Quality control of data and submission procedures (Steps 6-8)

To make the deliverables from the C-HPP more usable, the C-HPP group will employ high level quality control procedures that include robust, sensitive, accurate, portable, and quantitative assays, as described in the guidelines. Quality control of data produced by each group shall follow steps outlined in Fig. 1.

Step 6 will be to normalize the dataset to select the highest degree of confidence. Given the presence of two types of proteotypic peptides and ambiguously mapped peptides (www.proteomecenter.org), a more sensitive unambiguous, and quantitative assay should be developed for a protein. We will use SRMs as main engines for quantitation since the SRMAtlas is already available and provides suggested transitions and collision energy settings, observed retention times, calculated hydrophobicity, and information concerning peptide fragments. The consortium will utilize the new Cedar scheme and make updates of the datasets as better SRM methods become available. As a starting point for setting data quality control standards and for transparency of data production, we propose to adopt the Molecular and Cellular Proteomics Guidelines for transparency of data reports.[26] To produce results with confidence, we will use high accuracy (<10 ppm), high resolution (R>30,000) MS. It would also be better to define the false discovery rate range and mass tolerance accuracy for both HPP and MCP.[26] To manage data quality control, a solid guideline for the quality of PTM peptides is necessary. These guidelines should include the appropriate tools necessary to guarantee the confidence of PTM peptides.

Step 7 will be to set the standard operational procedures for data handling. As we learned from the exploratory phase of the Human Plasma Proteome Project[6] to consistently characterize proteins by MS methods, we should also put significant efforts into quality control and ensure commitment to deep analysis of specimens. To this end, the C-HPP group can have options for specimen selection. Category 1 deals with reference samples (e.g., non-disease or non-perturbed biological or clinical samples) without any unique test samples (e.g., normal cells, tissues, and organs) from which data can

be made available to the public without restrictions. Category 2 deals with unique test samples (e.g., disease, mutants, screening targets, etc.) from which data can be handled by an internal group with a view of sharing important biological and pathological findings. C-HPP needs efficient linking of MS data submitted to data repositories and then to data curation and meta-data analysis sites (Fig. 1). An important step will be the development of a fully-functioning ProteomeXchange system (www.proteomexchange.org), with explicit numbers of large datasets uploaded and automatically transferred among the parties. For quality control of datasets and standard data formatting, emphasis will be placed on the need for measurement standards, shared database systems, and a link between HPP and other on-going large-scale scientific initiatives such as ENCODE[27] and the 1000-genome project.[28]

With the development of C-HPP it will be necessary to propose standards for quantitative measurments. As such standardization can become a severe problem for measuring the ultra-low abundance proteins, it will be reasonable to compare amounts as number of protein copies per unit of biosample (e.g. per cell), as currently is widely adopted for RNA seq field.

Step 8 will be to build the C-HPP databases and utilize the data for biology and disease research in collaboration with B/D-HPP group. Findings and results generated by an individual group will belong to the corresponding investigators for publication. However, the results from each group should be published or deposited in the central C-HPP databases. It is also necessary to encourage all PIs to offer an opportunity of contribution and even re-analysis to the B/D and C-HPP teams.[1] It is intended that the data and meta-data will be promptly and openly available. The C-HPP web portal (www.c-hpp.org) that is also linked to the main HPP web site (http://thehpp.org) will be used to facilitate the overall progress and management of the C-HPP and will become a central focal point for the HPP in terms of publicizing the project, its goals, and the results. The KB committee within the HPP consortium will be in charge of data collection, annotation, deposition, retrieval, and management according to standardized formats. Appropriate documents, timelines, and links to the participating laboratories, funding agencies, and major proteomic resources and initiatives will populate the portal.

The potential deliverables of each group's research would be the protein information, including its name, primary function and major cellular localization. However, it would be beneficial to study comparative proteomics on suitable paired samples (normal *vs.* disease) to identify some differentially expressed proteins in certain physiological settings in collaboration with clinicians, pathologists and epidemiologist.

Overall Policies and Communications

Recruitment of additional teams. As opposed to the current consortium teams which have been recruited based upon the voluntary participation, we want to set out a general application procedure for additional teams as described in 'Conditions for Collaborations' section in this document.

Industrial partnership. The C-HPP consortium will need, at all stages of this project suitable for industrial partnerships (e.g., reagents, data, specimens, technology platforms) to provide prompt access to technical developments (MS, Ab, and KB) and high quality data processing. While the development and maintenance of tissue banking facilities is an important resource for our initiative, it is not central to the mission of C-HPP and thus the consortium members will collaborate with appropriate governmental programs (e.g., NIH-funded Tissue Research Banks, etc). Within the initiatives, each team of the consortium will select biological samples that express proteins of interest or are suitable for tissue expression studies. This work accompanies quality control on the samples using both MS (proteotypic peptide) and immunohistochemistry with a selected panel of antibodies.

Communications. The C-HPP consortium works toward its goals through ongoing series of workshops, teleconferences, and email communications. As a group, we aim to publish on all major activities and are particularly interested in recruiting new members motivated to contribute to issues surrounding the C-HPP datasets. For the C-HPP collaboration to succeed, all decisions and activities must be conducted openly, and decisions especially must be conducted in a transparent manner. This is best accomplished through clear, frequent, and open methods of communication among all of the

participating groups, either by phone, email, or through virtual private networks. Each PI is responsible for forwarding all communications to members of his or her own group. The larger proteomics and scientific communities will be kept informed through the HPP portal on the HUPO website. Regular open workshops and meetings grouping all C-HPP initiatives will be organized to guarantee a common set of rules for sharing/exchanging data and addressing protein annotation criteria.

Collaborative Benefits

General rules. If this work has been organized (or funded) by the C-HPP consortium, the consortium should be recognized in any publication. However, these data can also be used together by individual investigators for grant applications and presentations of the C-HPP. Use of the consortium-produced data after the primary publication will be governed by the principles outlined below. All scientists involved in the collaboration will declare possible conflicts of interest and will sign a document promising confidentiality. Analyses should emphasize reproducibility and transparency.[23]

The first and foremost objective of the C-HPP will be to find the missing proteins that have no protein evidence (or no high-quality evidence) when checked in the GPMDB which has already provided very comprehensive information, including spreadsheets of observed proteins for all chromosomes (http://www.thegpm.org/lists/). Such information and the further characterization of all known proteins will accelerate a wide range of studies by individual investigators and groups worldwide.

Investigators may also benefit through increased or stable funding, further learning opportunities, potential for career advancement, and learning more about the most appropriate strategies to uncover the function of proteins in complex diseases. There is a clear recognition that for the C-HPP collaboration to move forward, any decision made must attempt to provide mutual benefit for all those involved in the C-HPP. It will be difficult to ensure that the benefit is equally distributed, or that it is equal in kind among all partners, nevertheless there must be an assurance of mutual benefit. Each participant will have a mutual responsibility, including the necessity of giving up the claim on a chromosome if the

team does not feel it has the capabilities to deliver results at a level commensurate with the rest of the consortium, or if the EAC determines progress is unsatisfactory according to agreed criteria.

Intellectual property (IP). Each member should understand that an initial discovery of missing proteins cannot be the subject of IP protection and should share this information by releasing their datasets into the public domain within 30 days after initial identification. This will give an opportunity for confirmation of the results by other C-HPP participants. This policy is based on the notion that all involved entities should share a common purpose and should strive to be equal in terms of membership in the consortium. However, as a general practice principle, investigators, their institutions, and funding agencies may register and retain the IP from certain commercial applications based on that were made on the extensive characterization of missing proteins discovered from any of the human chromosomes.

Conditions for Collaborations

The C-HPP teams are currently seeking additional teams interested in mapping human proteins on a long-term basis, specifically to complete the roster of chromosome-based teams from the present 14 chromosomes (assuming all proceed) to all 24 chromosomes. This information will be updated from time to time as more teams join the consortium. There is no rush, however, to secure commitments for all 24 chromosomes, since the early players may have a very useful role in demonstrating feasible ways to use and visualize already available data and may stimulate bolder proposals from those who see the possibilities demonstrated and national funding emerge. Current teams may also request to undertake one or more additional chromosomes.

1) Current Teams (as of January 1, 2012)

Each team has an obligation to submit an annual activity report to EC.

2) New Entries

For those institutions or research groups interested in participating in the C-HPP, a proposal containing

the following information should be submitted:

(1) Name of applicant's organization

(2) Proposed initiative on a specific chromosome: Please provide information about the experimental strategy under proposal, including the target chromosome of initiative envisioned, preliminary results, data sources, experimental plans, compelling biological features or questions, proposed use of HUPO HPP reagents and reference specimens, and scientific and logistical advisers.

(3) Information about the memers of chromosome team (s)

- Name of group leader (PI) and institution
- Information about recent research publications in the field of proteomics
- Proposed contribution of each sub-group and research facilities

(4) Existing Resources: Please provide information about the existing resources available to the proposed program, including granting agencies (if any) and duration of support.

- research scientists and staff personnel, and their institutional affiliation
- research facilities
- existing and anticipated funding

(5) Vision for the Collaboration: Please indicate the reasons for participating collaboration through the C-HPP consortium including:

- potential role/contribution
- expected resources, reagents, and guidance from the C-HPP

Proposals should be sent to the chair and EC of the C-HPP consortium:

CONCLUSIONS AND PERSPECTIVES

This document (version 1.0) sets forth the general guidelines and the collaborative consensus of the C-HPP consortium. We anticipate that this version will be updated to embrace rapid development of the omics technologies and research environment. Collaboration within the C-HPP will be based on

shared objectives and consensus on the best ways to identify and characterize missing proteins that lack MS peptide evidence for each chromosome. This guideline establishes recommended experimental procedures and data production through which a list of missing proteins is expected to be compiled as well as the quality of mass spectrometric identifications improved in all but the highest probability category. Using the multi-step procedures for data quality control, this guideline led to a guarantee of the confidentiality of PTM peptides by filtering out any potential incompatible items between MCP guidelines and HPP data formats. By employing options for data submission, both reference sample (e.g., normal cells, tissues, and organs) and unique test samples (e.g., disease, mutants, screening targets, etc.) from which data can be handled by an internal group since it will be like proprietary or important biological findings. The collaboration needs strong leadership to reinforce the overall policies and run the everyday tasks of the project. However, it can only be successful if it makes maximum use of the creativity and energy of the participating research teams. We expect that benefits from collaborative efforts within the frame of this guideline with respect to sustained funding, shared resources, and reduced costs will produce synergistic outputs, aiding the generation of novel hypotheses relevant to the better understanding of human biology and disease mechanisms.

ACKNOWLEDGEMENTS

This work was supported by the World Class University program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (R31-2008-000-10086-0, WSH & YKP), the National Research Lab Program 2011-0028112 (to YKP) from the Ministry of Education, Science and Technology, and The National Project for the Personalized Genomic Medicine A111218-11-CP01 (to YKP) from the Ministry of Health and Welfare.

Supporting Information Available: This material is available free of charge via the Internet at

http://pubs.acs.org.

REFERENCES

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

(2) Hancock, W.; Omenn, G.; Legrain, P.; Paik, Y. K. Proteomics, human proteome project, and chromosomes. *J Proteome Res* **2011**, *10*, 210

(3) Paik, Y.K.; Jeong, S.K.; Omenn, G.S.; Uhlen, M.; Hanash, S. *et al.,* The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. *Nat. Biotechnol*., **2012** (*in press*)

(4) Uhlén, M.; Oksvold, P.; Algenas, C.; Hamsten, C.; *et al*., Antibody-based protein profiling

of the human chromosome 21. *Mol Cell Proteomics*. **2011, *10*:** M111.013458.

(5) Orchard, S.; Taylor, C.; Hermjakob, H.; Zhu, W. et al., Current status of proteomic standards development. *Expert Rev Proteomics* **2005**, *1*, 179-83

(6) Omenn, G. S.; States, D. J.; Adamski, M.; Blackwell, T. W. *et al*., Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. *Proteomics* **2005**, *5*, 3226-45

(7) He, F. Human liver proteome project: plan, progress, and perspectives. *Mol Cell Proteomics* **2005**, 4, 1841-8

(8) Uhlen, M.; Ponten, F. Antibody-based proteomics for human tissue profiling. *Mol Cell Proteomics* **2005**, *4*, 384-93

(9) Hamacher, M.; Marcus, K.; Stephan, C.; Klose, J.; Park, Y. M.; Meyer, H. E. HUPO Brain Proteome Project: toward a code of conduct. *Mol Cell Proteomics* **2008**, *7*, 457

(10) Yamamoto, T.; Langham, R. G.; Ronco, P.; Knepper, M. A.; Thongboonkerd, V. Towards standard protocols and guidelines for urine proteomics: a report on the Human Kidney and Urine Proteome Project (HKUPP) symposium and workshop, 6 October 2007, Seoul, Korea and 1 November 2007, San Francisco, CA, USA. *Proteomics* **2008**, *8*, 2156-9

(11) Fingar, V. H.; Wieman, T. J.; Doak, K. W. Changes in tumor interstitial pressure induced by photodynamic therapy. *Photochem Photobiol* **1991**, *53*, 763-8

(12) Picotti, P.; Rinner, O.; Stallmach, R.; Dautel, F. *et al*., High-throughput generation of selected reaction-monitoring assays for proteins and proteomes. *Nat Methods,* **2010,** *7*:43-6.

(13) Menon, R.; Omenn, G.S. Identification of alternatively spliced transcripts using a proteomic informatics approach. *Methods Mol Biol.* 2011, ***696***, 319-26

(14) Im KM, Kirchhoff T, Wang X, Green T, Chow CY, *et al*., Haplotype structure in Ashkenazi Jewish BRCA1 and BRCA2 mutation carriers. *Hum Genet*. 2011, ***130***, 685-99.

(15) Swaney, D. L.; Wenger, C. D.; Coon, J. J.; Value of using multiple proteases for large-scale mass spectrometry-based proteomics. *J Proteome Res*. **2010**,*9*,1323-9.

(16) Horvatovich, P.; Hoekman, B.; Govorukhina, N.; Bischoff, R.; Multidimensional chromatography coupled to mass spectrometry in analysing complex proteomics samples. *J. Sep. Sci.* **2010**, *33*,1421-37.

(17) Makarov, A.; Denisov, E.; Kholomeev, A.; Balschun, W.; Lange, O.; Strupat, K.; Horning, S.; Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. *Anal. Chem*. **2006**, *78*, 2113-20.

(18) Pichler, P.; Köcher, T.; Holzmann, J.; Mazanek, M.; Taus, T.; Ammerer, G.; Mechtler, K.; Peptide labeling with isobaric tags yields higher identification rates using iTRAQ 4-plex compared to TMT 6-plex and iTRAQ 8-plex on LTQ Orbitrap. *Anal Chem.* 2010, *82*, 6549-58.

(19) Kitteringham, N. R.; Jenkins, R. E.; Lane, C. S.; Elliott, V. L.; Park, B. K.; Multiple reaction monitoring for quantitative biomarker analysis in proteomics and metabolomics.*J Chromatogr B Analyt Technol Biomed Life Sci*. **2009,** *877*,1229-39

(20) Schmidt, A.; Beck, M.; Malmström, J.; Lam, H.; Claassen, M.; Campbell, D.; Aebersold, R. Absolute quantification of microbial proteomes at different states by directed mass spectrometry. *Mol Syst Biol* **2011**, *7*, 510

(21) Berglund, L.; Björling, E.; Oksvold, P.; Fagerberg, L.; Asplund, A.; Szigyarto, C. A.; Persson, A.; Ottosson, J.; Wernérus, H.; Nilsson, P.; Lundberg, E.; Sivertsson, A.; Navani, S.; Wester, K.; Kampf, C.; Hober, S.; Pontén, F.; Uhlén, M. A genecentric Human Protein Atlas for expression profiles based on antibodies. *Mol Cell Proteomics* **2008**, *7*, 2019-27

(22) Duncan, M. W.; Aebersold, R.; Caprioli, R. M. The pros and cons of peptide-centric proteomics. *Nat Biotechnol* **2010**, *28*, 659-64

(23) Kapp, E.A.; Schütz, F.' Connolly, L.M.; Chakel, J. A.; Meza, J. E. *et al*., An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis.*Proteomics*. **2005,** *5*,3475-90.

(24) Zhang, B.; Chambers, M. C.; Tabb, D. L. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *J Proteome Res* **2007**, *6*, 3549-57

(25) Farrah, T.; Deutsch, E. W.; Omenn, G. S.; Campbell, D. S.; Sun, Z. *et al*., A high-confidence human plasma proteome reference set with estimated concentrations in PeptideAtlas. *Mol Cell*

Proteomics **2011**, *10*, M110.006353

(26) Bradshaw, R. A.; Burlingame, A. L.; Carr, S.; Aebersold, R. Reporting protein identification data: the next generation of guidelines. *Mol Cell Proteomics* **2006**, *5*, 787-8

(27) ENCODE Project Consortium; Myers, R. M.; Stamatoyannopoulos, J.; Snyder, M.; Dunham, I.; Hardison, R. C. *et al*., User's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* **2011**, *9*, e1001046

(28) Conrad, D. F.; Keebler, J. E.; DePristo, M. A.; Lindsay, S. J.; Zhang, Y. *et al.*, Variation in genome-wide mutation rates within and between human families. *Nat Genet* **2011**, *43*, 712-4

Table 1. The proposed two phases of C-HPP and short-term/long-term challenges

Phase	Phase I	Phase II
Years	***6***	***4***
Milestones	•Organization, SOP and Guidelines •Mapping and characterization of 6000 missing proteins having no MS evidence[a] •Mapping the predicted three major PTMs (phosphoryl-, glycosyl-, acetyl-) of 10,000 well-known proteins[b] •One representative AST and nsSNP of predicted c.a. 14,300 well-known proteins •Genomic/transcriptomic/proteomic validation of predicted c.a. 14,300 well-known proteins •Cellular localization and quantitation of newly identified proteins in at least three tissues	•Further characterization of whole human proteins (~20,300) with respect to gene location on each chromosome, cellular distribution and quantitation •Validation of three major PTMs present in 20,300 human proteins •Validation of one representative AST and nsSNP of 20,300 well-known proteins •Genomic/transcriptomic validation of whole proteins in at least three representative tissues •Development of drug targets and biomarker candidates of interest. •Functional studies of gene families/clusters in each chromosome
Coping with short-term and longer-term challenges	**Short-Term**	**Solutions**
	•Sustainable funding/cost savings •Effective cross analysis and integration of different datasets •Handling of data variability •Harmonization with biology-driven projects •Procurement of affinity captured reagents	•Sharing resources, data, and reagents (AB), ref specimens •Employing neXtProt, dbSNP, GPMDB and PeptideAtlas •Standard data submission system/criteria •Sharing data through C-HPP portal and other public DBs •Close collaborations between providers and C-HPP groups
	Longer-Term	**Solutions**
	•Enhanced detection limit for low abundance (rare) proteins •Improved pretreatment of clinical specimens for characterization and SRM analysis •Inclusion of PTM information in different datasets •Sample bio-banking and maintenance	•Miniaturization of sample preparation/efficient fractionations •Continued refinement of non-redundant protein list •Development of new algorithms for inclusion of PTMs •Collaboration with government agencies

[a]Less well-known: proteins that have only transcriptomic evidence, but not proteomic MS data (constitutes about 6,000 proteins).

[b]Well-known: proteins that have both transcriptomic and proteomics MS data. The data for proteins under investigation will be integrated into one common C-HPP portal by contributions from each chromosome team.

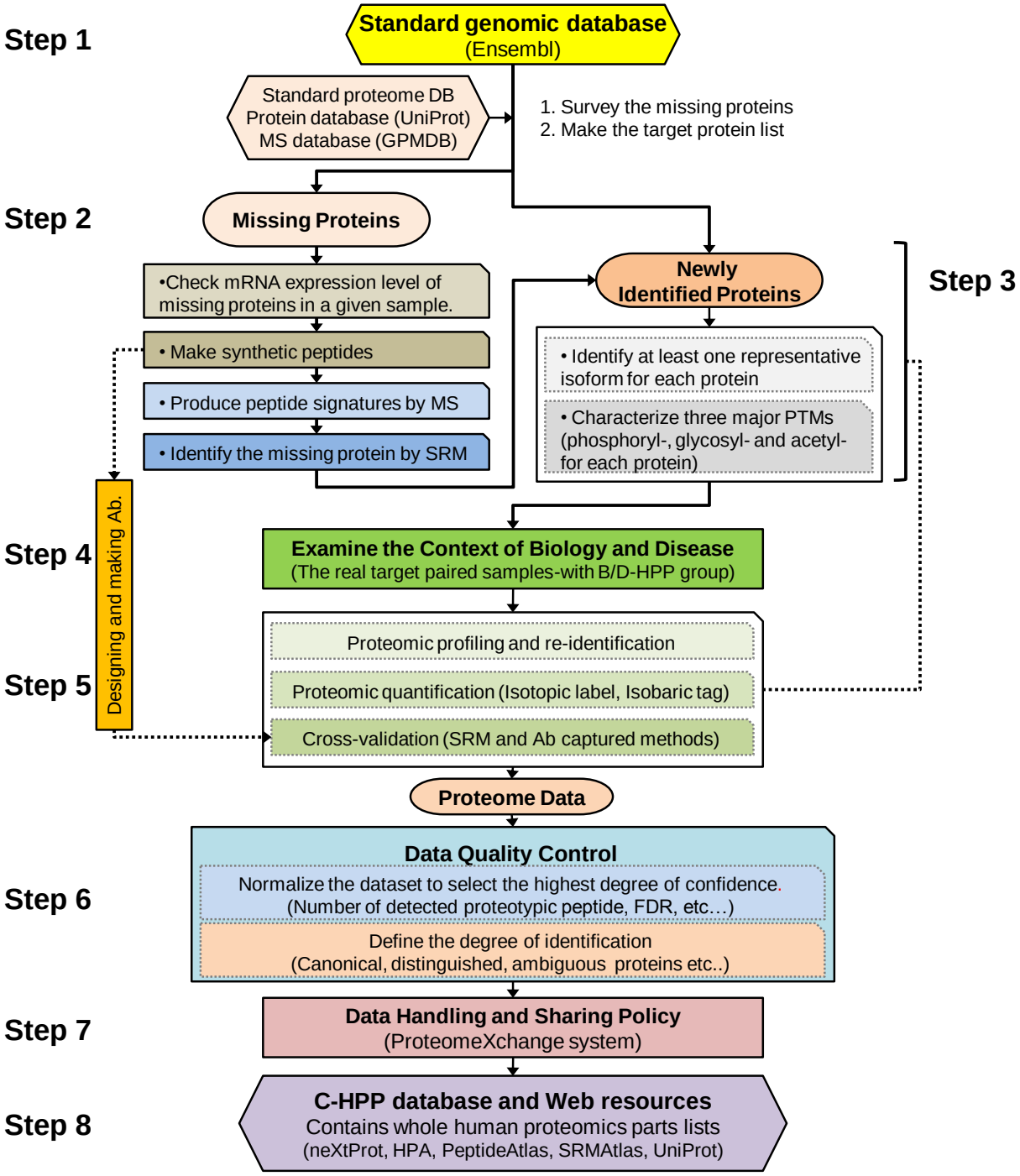
[c]In the case where a protein cannot be detected by MS measurement of the enzymatic digest, the CHPP will use alternative approaches, such as targeted SRM measurement or Ab based assays.

Figure Legend

Fig. 1. Flow of the overall procedures of the C-HPP work.

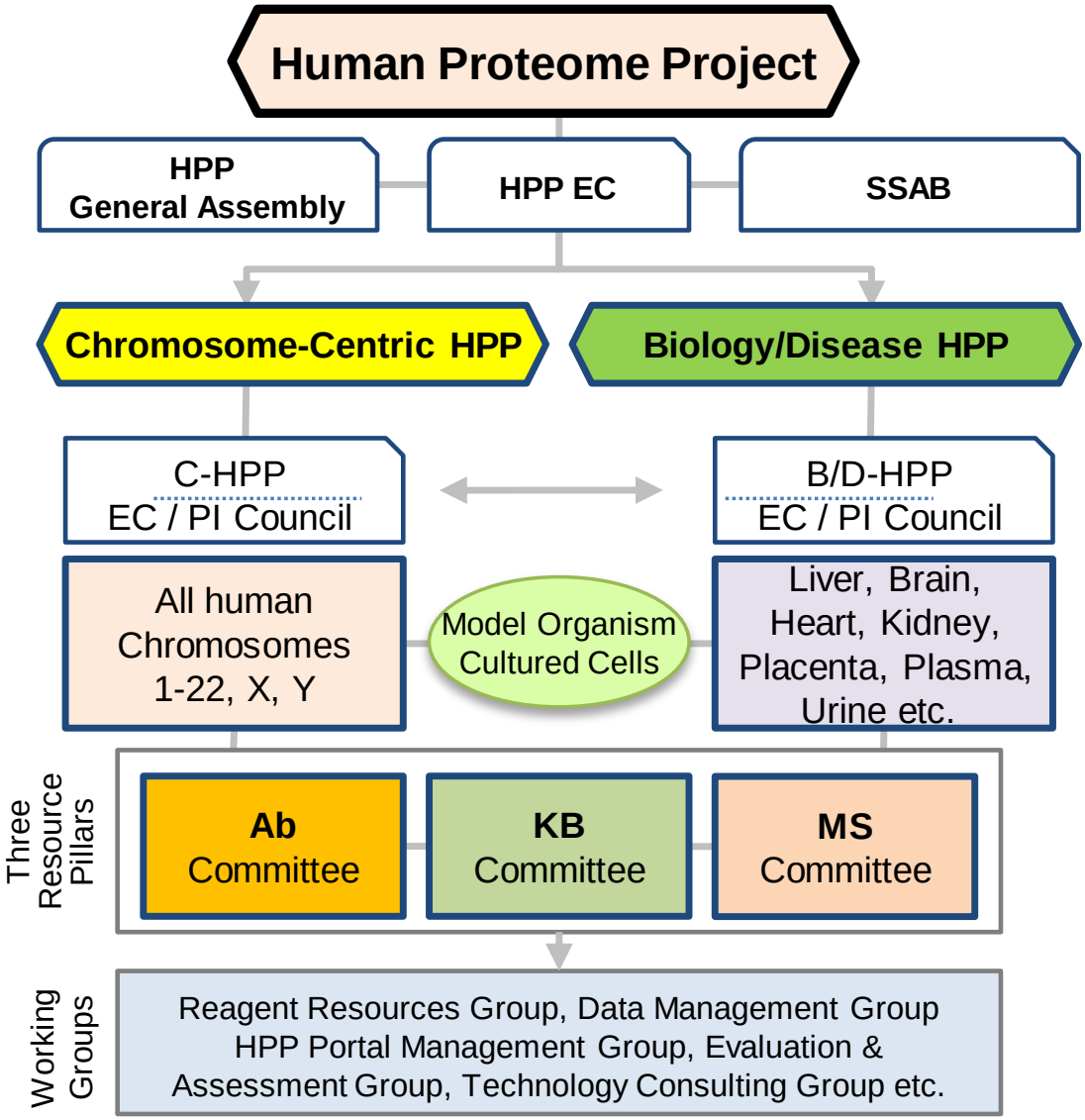
Shown here is a recommended procedure for data production and management in the C-HPP. Alternative procedure can also be applied depending on the sample type or progress of each experiment. The strategy used for the C-HPP starts with the definition of the proteins coded for by the human genome (based on UniProt), a list of missing/poorly-characterized proteins, and proteomic profiling studies to identify those missing proteins.

Fig. 1


Step 1
Standard genomic database
(Ensembl)
Standard proteome DB
Protein database (UniProt)
MS database (GPMDB)
1. Survey the missing proteins
2. Make the target protein list
Step 2
Missing Proteins
•Check mRNA expression level of missing proteins in a given sample.
• Make synthetic peptides
• Produce peptide signatures by MS
• Identify the missing protein by SRM
Newly Identified Proteins
Step 3
• Identify at least one representative isoform for each protein
• Characterize three major PTMs (phosphoryl-, glycosyl- and acetyl- for each protein)
Designing and making Ab.
Step 4
Examine the Context of Biology and Disease
(The real target paired samples-with B/D-HPP group)
Step 5
Proteomic profiling and re-identification
Proteomic quantification (Isotopic label, Isobaric tag)
Cross-validation (SRM and Ab captured methods)
Proteome Data
Step 6
Data Quality Control
Normalize the dataset to select the highest degree of confidence.
(Number of detected proteotypic peptide, FDR, etc…)
Define the degree of identification
(Canonical, distinguished, ambiguous proteins etc..)
Step 7
Data Handling and Sharing Policy
(ProteomeXchange system)
Step 8
C-HPP database and Web resources
Contains whole human proteomics parts lists
(neXtProt, HPA, PeptideAtlas, SRMAtlas, UniProt)

Supplemental Table S1. Major web sites of C-HPP related public resources

Web page/DB name	URL (Web address)
HUPO	
HUPO homepage	http://www.hupo.org.
HUPO initiatives	http://www.hupo.org/research/
Human Proteome Project homepage	Http://thehpp.org
C-HPP homepage	http://www.c-hpp.org
Genome/Proteome database	
UniProt	http://www.uniprot.org/
Ensembl	http://www.ensembl.org/
NCBI : National Center for Biotechnology Information	http://www.ncbi.nlm.nih.gov/
KEGG : Kyoto Encyclopedia of Genes and Genomes	http://www.genome.jp/kegg/
Gene Ontology	http://www.geneontology.org/
AntibodypediA	http://www.antibodypedia.org/
Proteomic data repository	
ProteomeXchange	http://www.proteomexchange.org/
Tranche	https://proteomecommons.org/tranche/
PRIDE	http://www.ebi.ac.uk/pride/
The GPMDB	http://gpmdb.thegpm.org/
PeptideAtlas	http://www.peptideatlas.org/
SRM/MRM atlas	http://www.srmatlas.org/
Human specific databases	
neXtProt	http://www.nextprot.org/
The Human Protein Atlas	http://www.proteinatlas.org/
Human Protein Reference Database	http://www.hprd.org/



Supplemental Fig. S1. Organization of C-HPP governance as part of HPP

In the C-HPP, multiple working groups are formed to analyze and manage data and produce the protein partslist encoded by each chromosome with respect to both functional context and disease relation in collaboration with B/D-HPP (adopted and modified from Paik et al. [unpublished publication]). Abbreviations used here are: Ab, antibody/affinity capture reagent; EC, executive committee; KB, bioinformatics knowledge base; MS, mass spectrometry; PI, principal investigator; SSAB, senior scientific advisory board.

TABLE OF CONTENTS SYNOPSIS

With establishment of standard guidelines of the C-HPP, we envision that most of the datasets obtained independently from individual investigators will have the stringent quality control system that can be stored in C-HPP data repositories for public use.

Synopsis Figure