The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome

To the Editor:
The Chromosome-Centric Human Proteome Project (C-HPP) aims to define the full set of proteins encoded in each chromosome through development of a standardized approach for analyzing the massive proteomic data sets currently being generated from dedicated efforts of national and international teams. The initial goal of the C-HPP is to identify at least one representative protein encoded by each of the approximately 20,300 human genes[1,2]. The proteins will be characterized for tissue localization and major isoforms, including post-translational modifications (PTMs), using quantitative mass spectrometry and antibody reagents. Our rationale is that effective integration of proteomics data into a genomic framework will lead to improved knowledge of complex biological systems and facilitate access to protein level data. Although the intent to engage in a C-HPP program has been noted[1–3], our objective here is to define the goals and process for its development as a multinational program.

Over the past three years, the Human Proteome Organization (HUPO) has developed a strategy for the first phase of the Human Proteome Project (HPP; http://thehpp.org/; **Supplementary Fig. 1**). HPP1 goals will be achieved through cooperation with the C-HPP to characterize the human proteome on a chromosome-by-chromosome basis and with the biology- and disease-driven projects (B/D-HPP). Human genome studies, such as the 1000 Genomes Project and Encode, and transcriptome sequencing provide a basis for identification of protein isoforms generated by alternative splicing transcripts (ASTs) and by nonsynonymous single-nucleotide polymorphisms (nsSNPs; **Supplementary Fig. 2**). Additional protein forms will be identified through characterization of post-translational modifications. A basic premise of the HPP is that C-HPP data sets will have substantial utility for biological and disease studies. With development of new tools for in-depth characterization of the transcriptome and proteome, the HPP is well positioned to have a strategic role in addressing the complexity of human phenotypes. With this in mind, the HUPO has organized national chromosome teams that will collaborate with well-established laboratories building complementary proteotypic peptides, antibodies and informatics resources.

An important C-HPP goal is to encourage capture and open sharing of proteomic data sets from diverse samples to enhance a gene- and chromosome-centric display This will display several layers of biological information on a common reference platform comparable to a genome browser. Such context will effectively integrate transcriptomics data such as RNA-Seq with proteomic data sets (**Fig. 1**).

Although the C-HPP program has some similarities to the Human Genome Project (HGP)[4] in its quest for complete coverage across the genome, the C-HPP has the added challenge of characterizing protein expression at the tissue, cellular and subcellular levels, as well as PTMs, ASTs and protease-processed protein variants. An example of protein variation is shown for 6 selected genes on chromosome 13 (*BRCA2*, 3 ASTs and 54 SNPs in protein-coding regions (nsSNPs); *RB1*, 2 ASTs and 3 nsSNPs; and *IRS2*, 1 AST and 3 nsSNPs) and chromosome 17 (*BRCA1*, 24 ASTs and 24 nsSNPs; *ERBB2*, 6 ASTs and 13 nsSNPs; and *TP53*, 14 ASTs and 5 nsSNPs; **Table 1** and **Supplementary Table 1**).

The C-HPP will build on the three HPP pillars that provide both technology and resources for mapping the human proteome: mass spectrometry–based SRMAtlas, antibody reagents in the Human Protein Atlas and bioinformatics knowledge linked by ProteomeXchange, specifically the proteomics identification database (PRIDE), Tranche, PeptideAtlas, the global proteome machine database (GPMDB), UniProt and neXtProt (**Supplementary Fig. 3**).

Table 1 Features of salient genes on chromosomes 13 and 17

Gene[a]	AST	nsSNPs
Chromosome 13		
BRCA2	3	54
RB1	2	3
IRS2	1	3
Chromosome 17		
BRCA1	24	24
ERBB2	6	13
TP53	14	5

[a]Ensembl protein and AST information can be found at http://www.ensembl.org/Homo_sapiens/.
AST, alternative splicing transcript; nsSNP, nonsynonymous single-nucleotide polyphorphism assembled from data from the 1000 Genomes Projects.

The C-HPP does not propose any alteration in the work flow of a typical proteomics laboratory; instead, it seeks more effective use of data encompassed in existing bioinformatics resources, which will be combined with targeted studies to generate a robust list of observed protein isoforms (**Supplementary Fig. 3**). A potential challenge to data collection from different laboratories is the diversity of instrument and bioinformatics platforms and quality criteria. The C-HPP will work closely with proteomics journals, and use existing data (GPMDB and PeptideAtlas), literature curation (Uniprot and neXtProt) and standardization programs (PSI, CPTAC, Unimod, ABRF and ASMS) to ensure that the data collection is efficient, with consistent quality assurance and quality control. Journal mandates for deposition of raw data upon publication will reinforce this process[5]. The C-HPP has already encouraged formation of chromosome-formatted databases (http://www.nextprot.org/; http://www.gpm.org/) in which new data sets are integrated with existing ones. In this manner the C-HPP will capture the protein evidence emerging from the hundreds of laboratories worldwide engaged in hypothesis-driven

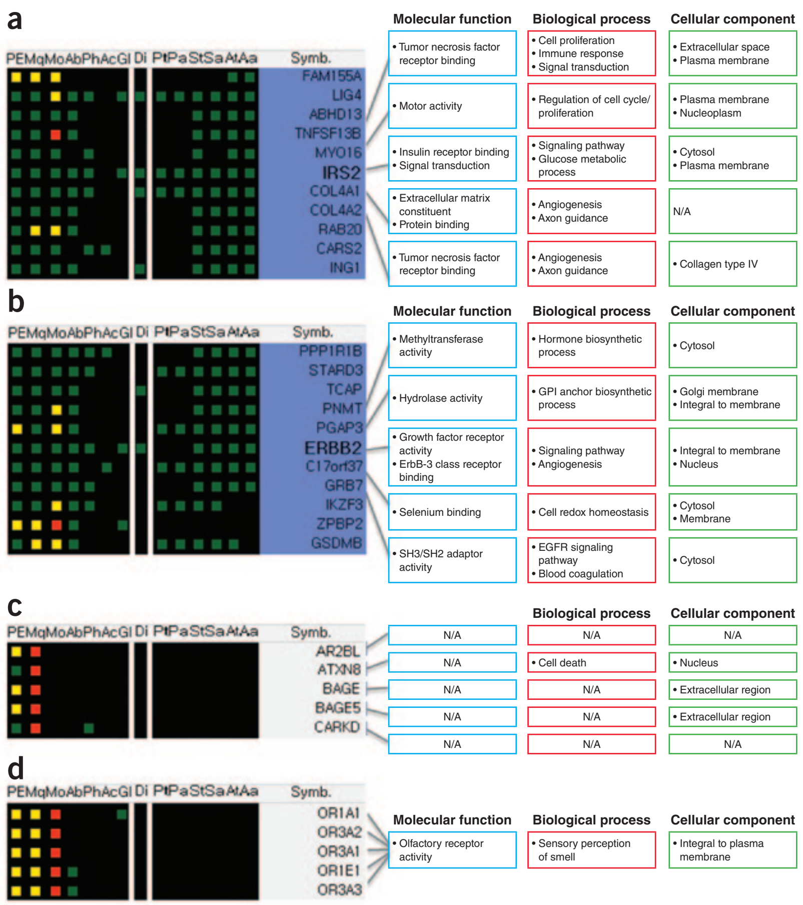

Figure 1 Genomic, transcriptomic and protein information for the set of genes present in selected regions of chromosomes 13 and 17. (**a**,**b**) The information provides a comprehensive landscape with respect to protein evidence, quality of mass spectrometry–based protein identification, availability of antibody, disease relationship, and phosphorylation, acetylation, glycosylation and transcriptomic information. It shows the degree of protein annotation on two important regions on chromosomes 13 (**a**) and 17 (**b**) and regions with little annotated protein information on chromosomes 13 (**c**) and 17 (**d**). PE, protein evidence from UniProt; Mq, mass quality from GPMDB; Mo, number of mass spectrometry data sets in GPMDB; Ab, antibody availability; Di, disease information; Ph, Ac and Gl, phosphoryl, acetyl and glyco, respectively; Pt, placenta transcript; Pa, placenta AST; St, SKBR3 breast cancer cell line transcript; Sa, SKBR3 breast cancer cell line AST; At, A431 transcript; Aa, A431 AST. Green denotes presence and black denotes lack of information in the following data sets: transcript, disease, PTMs and antibodies. For protein evidence, green, yellow and red represent high (protein evidence), medium (transcriptomic evidence) and low (neither) evidence, respectively. Number of individual data sets and quality of mass spectrometry evidence was established according to GPMDB scores: green, >20 observations, log(*e*) < –5; yellow, 6–19 observations, –3 ≤ log(*e*) < –5; red, 1–5 observations, –1 ≤ log(*e*) < –3. For the relationship of each protein to disease, we used both Online Mendelian Inheritance in Man (OMIM; NCBI, confirmed Mendelian phenotype) and Cancer Gene Census (CGC, Sanger Center; cancer gene information). For the PTM information, we used UniProt/UniPep containing experimental PTM site information and GPMDB providing mass spectrometry information for the PTMs.

research or high-throughput proteome-wide studies.

Although chromosome-based protein data curation is a relatively new concept in proteomics[2], our justification is based in part on compatibility of this data format with the output of RNA-Seq. We think the search for yet-to-be-discovered protein products of genes can be informed by transcriptomics measurements of selected tissues and cell lines. The C-HPP will also prioritize specific tasks to laboratories with expertise in particular protein subsets (for example, membrane proteins), specific protein variations (PTMs, alternative splicing and protease-processed variants), deep profiling for low-abundance proteins and targeted subcellular localization studies. We recognize the popularity of other current bioinformatic methods used to organize complex data sets by functional classes; we will incorporate this information into the C-HPP browser. An example of such a global view for selected regions of chromosomes 13 and 17 (**Fig. 1**) summarizes the following extensive data sets on the basis of existing data compilations: protein evidence, mass spectrometry data, antibody availability, major PTMs, disease information and transcript level, including ASTs from three different samples in a format viewable for associations between data sets and information gaps in specific chromosome regions.

In phase 1 (~6 years), the C-HPP plans to map all proteins currently lacking high-quality mass spectrometry evidence, three major classes of PTMs, many representative AST products[6] and many nsSNP sequence variants. The characterizations will be followed by antibody-based detection in selected tissues and cell lines. In phase 2 (~4 years), identified proteins will be characterized and validated with additional proteomic and antibody measurements. Throughout this 10-year project, the C-HPP aims to generate information useful in the search for new biomarkers and drug targets and also in the study of disease gene families clustered in each chromosome (for example, the cytokeratin gene family in chromosome 17). C-HPP outputs will be integrated with output from the parallel B/D-HPP project. The C-HPP has selected the UniProt protein list (based on Ensembl genome builds) as the starting point for identified proteins. Individual chromosome teams will use information collected in well-annotated databases (for example, GPMDB, PeptideAtlas and neXtProt) to develop a list of missing or poorly identified proteins for a particular chromosome. A plot of such data (for example, **Fig. 1**) can identify chromosomal regions with low amounts of data. For example, there is protein paucity for regions on chromosome 17 that contain olfactory receptors and keratin-binding proteins; this may be expressed in limited proteomic data sets for nasal epithelium and bone and hair samples, respectively (**Fig. 1**). The missing data can be obtained through collaborations with laboratories with expertise in such samples or by selection of new sample sets for protein identifications guided by transcriptomics measurements. To facilitate selection of samples suitable for mass spectrometry discovery of an individual missing protein, the C-HPP will collaborate with RNA-Seq laboratories to take advantage of specimens and transcriptomics data (**Supplementary Fig. 2**). We recognize that some proteins may not be suited for mass spectrometry measurements owing to their physical properties or lack of appropriate biological samples; other approaches such as generation of ribosomal DNA standards, antibody localization approaches and molecular biology tools will be used.

Given expected refinements in the human gene list, the C-HPP protein list will reflect updates in Uniprot that are captured in proteomic databases. To ensure consistent data quality across chromosome groups, the C-HPP will encourage prompt deposition of data. For antibody-based studies, the C-HPP will promote the use of cultured primary or transformed cells, including induced pluripotent stem cells, which can be maintained in perpetuity for reanalysis and for subcellular fractionation. Such cell-based studies will be augmented with tissue profiling, as in the Human Protein Atlas project. Enrichment for nuclear, mitochondrial and other subcellular organelles may be especially informative[7,8]. The C-HPP will integrate antibody- and mass spectrometry–based measurements.

Another goal of the C-HPP is to procure high-quality reagents. To augment commercially available sources, the national teams will establish centralized antibody banks. This will be achieved through a close collaboration between each chromosome group and antibody resource groups or suppliers. In a similar manner, selected reaction monitoring peptide banks will be developed for quantitative mass spectrometry measurements.

The project will meet its aims when the comprehensive C-HPP database is 100% matched with the 20,300 protein-coding genes annotated on the human genome sequence, including at least one representative AST and nsSNP, tissue localization and three classes of PTMs in whole-chromosome sets (22 autosomal, X and Y; **Supplementary Table 2**).

The C-HPP is led by cochairs Young-Ki Paik (Korea), Bill Hancock (USA) and Gyorgy Marko-Vargas (Sweden), an executive committee and a council of principal investigators of each of the chromosome teams (thus far, 15 investigators for 14 chromosomes; **Supplementary Fig. 2**). The initial C-HPP team emerged from an exploratory group in Korea that selected chromosome 13; it has several key metabolic disease genes (for example, *IRS2*, which is associated with diabetes, and *CLF*, which is associated with cholesterol metabolism). Diverse approaches have been pursued by other countries and teams. A US team has focused on breast cancer, selecting chromosome 17, which contains the oncogenes *ERBB2* and *BRCA1*.

Similarly, the Australia-New Zealand team selected chromosome 7, with a focus on colon cancer and epidermal growth factor receptor. C-HPP guidelines now have been set for the assignment and progress review of chromosome-based teams and standardization of outputs (Y.-K. Paik, G.S. Omenn, M. Uhlen, S. Hanash, G. Marko-Varga *et al.*, unpublished data). As of December 2011, based on their interests in a specific disease (for example, male infertility in Iran) or gene cluster (for example, liver-origin proteins in China), other international teams have chosen chromosomes 1 (China), 2 (Switzerland), 3 (Japan), 6 (Canada), 11 (Korea), 14 (France), 18 (Russia), 19 (Sweden and Germany, Norway, India, China and Spain), 21 (Canada), X (Japan) and Y (Iran). A Swedish team has published extensive findings for chromosome 21 (ref. 9). The C-HPP guidelines specify management of the project, data quality and data sharing metrics, reporting formats, and processes and criteria by which countries or researchers are designated to take the lead for a specific chromosome (Y.-K. Paik *et al.*, unpublished data).

In conclusion, we envision that effective integration of transcriptomics and proteomics data will provide insights through a more complete 'parts list' and enhance a comprehensive understanding of human biology. The HPP and the C-HPP represent an even larger endeavor than the HGP. This challenge has led the HUPO to promote an efficient approach of recruiting national teams with clear areas of responsibility and effective collaborations among leading proteomic laboratories in the HPP consortium. Recognizing the complexity of the human proteome, we have set 10-year goals for characterizing the major forms of the complete set of proteins. The C-HPP will provide a global open Web interface for data collection, curation and presentation of the proteome parts list and will stimulate availability of high-quality protein capture and signature peptide reagents (**Supplementary Table 2**). Importantly, the C-HPP will work with governmental funding bodies to address major gaps in proteomics infrastructure, such as secure archiving of large data sets.

Note: Supplementary information is available on the Nature Biotechnology website.

AUTHOR CONTRIBUTIONS

Y-K.P. and W.S.H. conceived strategies in coordination with G.S.O., M.U., S.H., G.M.-V., E.W.D., R.A., A.B., D.H. and P.L.; S.-K.J. carried out profile analysis with programs developed. J.Y.K and H.K. provided clinical samples and information; H.-J.L., F.Y., F.Z., Y.Z., S.Y.C., K.N., K.Y.K., E.-Y.L., E.-Y.C., Y.C., R.C. and A.D.T. carried out various experiments including sample preparation, proteomic analysis and RNA sequencing of cell lines.

ACKNOWLEDGMENTS

We thank R. Beavis for his critical comments and support for this work. Y.-K.P. thanks the Korean Human Proteome Organization HPP planning committee members for their contribution to the project in the early phase. Work involving chromosomes 13 and 17 in this paper was supported in part by the World Class University program funded by the Korean Ministry of Education, Science and Technology (to Y.-K.P. and W.S.H.), a grant from the Korean Ministry of Health and Welfare (to Y.-K.P.) and grants from the US National Cancer Institute (to M.S. and W.S.H.).

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Young-Ki Paik[1], Seul-Ki Jeong[1], Gilbert S Omenn[2,3], Mathias Uhlen[4], Samir Hanash[5], Sang Yun Cho[1,13], Hyoung-Joo Lee[1], Keun Na[1], Eun-Young Choi[1], Fangfei Yan[6], Fan Zhang[6], Yue Zhang[6], Michael Snyder[7], Yong Cheng[7], Rui Chen[7], György Marko-Varga[8], Eric W Deutsch[3], Hoguen Kim[9], Ja-Young Kwon[9], Ruedi Aebersold[10], Amos Bairoch[11], Allen D Taylor[4], Kwang Youl Kim[1], Eun-Young Lee[1], Denis Hochstrasser[11], Pierre Legrain[12] & William S. Hancock[1,5]

[1]Yonsei Proteome Research Center, Yonsei University, Seoul, Korea. [2]Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, USA. [3]Institute for Systems Biology, Seattle, Washington, USA. [4]Royal Institute of Technology, Stockholm, Sweden. [5]Fred Hutchinson Cancer Research Center, Seattle, Washington, USA. [6]Northeastern University, Boston, Massachusetts, USA. [7]Stanford University, Palo Alto, California, USA. [8]Lund University, Lund, Sweden. [9]Yonsei University College of Medicine, Seoul, Korea. [10]Department of Biology, Institute of Molecular Systems Biology, Eidgenössische Technische Hochschule, Zürich, Switzerland, and Faculty of Science, University of Zurich, Zurich, Switzerland. [11]Swiss Institute of Bioinformatics and University of Geneva, Geneva, Switzerland. [12]Ecole Polytechnique, Palaiseau, France. [13]Present address: Korean National Institute of Health, Osong, Korea. e-mail: Y.-K.P. (paikyk@yonsei.ac.kr) or W.S.H. (wi.hancock@neu.edu)

1. Legrain, P. *et al. Mol. Cell Proteomics* **10**, M111.009993 (2011).
2. Hancock, W. *et al. J. Proteome Res.* **10**, 210 (2011).
3. Service, R.F. *Science* **321**, 1758–1761 (2008).
4. Lander, E.S. *et al. Nature* **409**, 860–921 (2001)
5. Farrah, T. *et al. Mol. Cell. Proteomics* **10**, M110.006353 (2011).
6. Menon, R. & Omenn, G.S. *Methods Mol. Biol.* **696**, 319–326 (2011).
7. Gnad, F. *et al. Mol. Cell. Proteomics* **9**, 2642–2653 (2011).
8. Walther, T.C. & Mann, M. *J. Cell Biol.* **190**, 491–500 (2010).
9. Uhlén, M. *et al. Mol. Cell. Proteomics* published online, doi:10.1074/mcp.M111.013458 (31 October 2011).