Journal of proteome research

Perspective

Cite This: J. Proteome Res. 2018, 17, 4042−4050

pubs.acs.org/jpr

Launching the C-HPP neXt-CP50 Pilot Project for Functional Characterization of Identified Proteins with No Known Function

Young-Ki Paik,*[,†] Lydie Lane,[‡] Takeshi Kawamura,[§] Yu-Ju Chen,[∥] Je-Yoel Cho,[⊥] Joshua LaBaer,[#] Jong Shin Yoo,[▽] Gilberto Domont,[○] Fernando Corrales,[◆] Gilbert S. Omenn,[¶] Alexander Archakov,[+] Sergio Encarnación-Guevara,[□] Siqi Lui,[▲] Ghasem Hosseini Salekdeh,[●,■] Jin-Young Cho,[†] Chae-Yeon Kim,[†] and Christopher M. Overall*[,▼]

[†]Yonsei Proteome Research Center and Department of Integrative Omics, Yonsei University, Sudaemoon-ku, 120-749 Seoul, Korea

[‡]CALIPHO Group, Swiss Institute of Bioinformatics & Department of Microbiology and Molecular Medicine, Faculty of Medicine, University of Geneva, CH-1211 Geneva, Switzerland

[§]Proteomics Laboratory, Isotope Science Center, The University of Tokyo, Bunkyo-Ku, Tokyo 113-0032, Japan

[∥]Institute of Chemistry, Academia Sinica, 128 Academia Road Sec. 2, Nankang, Taipei 115, Taiwan

[⊥]Research Institute for Veterinary Science, College of Veterinary Medicine, Seoul University, 1 Kwanak-ro, Kwanak-gu, 151-742 Seoul, South Korea

[#]Biodesign Institute, Arizona State University, 1001 South McAllister Avenue, Tempe, Arizona 85287-5001, United States

[▽]Division of Mass Spectrometry Research, Korea Basic Science Institute, 28119 Ochang, Korea

[○]Federal University of Rio de Janeiro, Institute of Chemistry, Rio de Janeiro, Rio de Janeiro 21941-909, Brazil

[◆]Functional Proteomics Laboratory, National Center of Biotechnology, CSIC, 28049 Madrid, Spain

[¶]Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan 48109-2218, United States

[+]Institute of Biomedical Chemistry RAS, Moscow 119121, Russia

[□]Center for Genomic Sciences, National Autonomous University of Mexico, Cuernavaca, Morelos 62210, Mexico

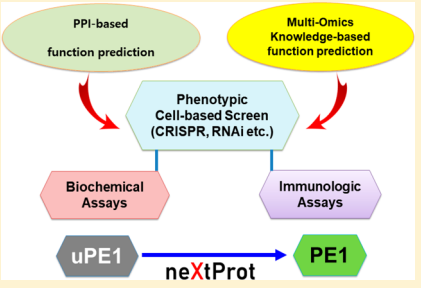
[▲]BGI-Shenzhen, Beishan Industrial Zone, Yantian District, Shenzhen 518083, China

[●]Department of Molecular Systems Biology, Royan Institute for Stem Cell Biology and Technology, 1665659911 Tehran, Iran

[■]Department of Molecular Sciences, Macquarie University, Sydney, New South Wales 2109, Australia

[▼]Centre for Blood Research, Departments of Oral Biological & Medical Sciences and Biochemistry & Molecular Biology, Faculty of Dentistry, University of British Columbia, Vancouver, British Columbia V6T 1Z3, Canada

ABSTRACT: An important goal of the Human Proteome Organization (HUPO) Chromosome-centric Human Proteome Project (C-HPP) is to correctly define the number of canonical proteins encoded by their cognate open reading frames on each chromosome in the human genome. When identified with high confidence of protein evidence (PE), such proteins are termed PE1 proteins in the online database resource, neXtProt. However, proteins that have not been identified unequivocally at the protein level but that have other evidence suggestive of their existence (PE2−4) are termed missing proteins (MPs). The number of MPs has been reduced from 5511 in 2012 to 2186 in 2018 (neXtProt 2018-01-17 release). Although the annotation of the human proteome has made significant progress, the "parts list" alone does not inform function. Indeed, 1937 proteins representing ∼10% of the human proteome have no function either annotated from experimental characterization or predicted by homology to other proteins. Specifically, these 1937 "dark proteins" of the so-called dark proteome are composed of 1260 functionally uncharacterized but identified PE1 proteins, designated as uPE1, plus 677 MPs from categories PE2−PE4, which also have no known or predicted function and are termed uMPs. At the HUPO-2017 Annual Meeting, the C-HPP officially adopted the uPE1 pilot initiative, with 14 participating international teams later committing to demonstrate the feasibility of the functional characterization of large numbers of dark proteins (CP), starting first with 50 uPE1 proteins, in a stepwise chromosome-centric



continued...

Special Issue: Human Proteome Project 2018

Received: May 29, 2018
Published: October 1, 2018

DOI: 10.1021/acs.jproteome.8b00383
J. Proteome Res. 2018, 17, 4042−4050

organizational manner. The second aim of the feasibility phase to characterize protein (CP) functions of 50 uPE1 proteins, termed the neXt-CP50 initiative, is to utilize a variety of approaches and workflows according to individual team expertise, interest, and resources so as to enable the C-HPP to recommend experimentally proven workflows to the proteome community within 3 years. The results from this pilot will not only be the cornerstone of a larger characterization initiative but also enhance understanding of the human proteome and integrated cellular networks for the discovery of new mechanisms of pathology, mechanistically informative biomarkers, and rational drug targets.

KEYWORDS: *C-HPP, dark protein, Human Proteome Project, missing protein, neXt-CP50, protein evidence, proteoform, uncharacterized protein evidence 1 (uPE1), uncharacterized missing protein (uMP)*

INTRODUCTION

Although the completion of the human genome project identified approximately 20 000 protein-coding genes,[1] there have been ongoing updates in defining the credible number of canonical proteins and their probable biological functions. The protein-encoding open reading frame (ORF) numbers from different public databases change from year to year depending on new discoveries in the human genome and identification of their cognate protein products.[2] This unclosed status of human protein numbers continues as a motivator for active research contributions in the Human Proteome Project (HPP) of the Human Proteome Organization (HUPO) for the identification and mapping of all human proteins in a chromosome-centric manner.[3,4]

From the inception of the HPP at HUPO-2010 in Sydney, Australia, an established and mature initiative of the HPP has been the Chromosome-centric HPP (C-HPP), which aims to correctly define the number and identify each of the canonical proteins encoded by their cognate ORFs in the human genome.[4,5] The predicted number of human proteins is now 20 230 (neXtProt 2018-01-17 release) (but this will undoubtedly change over the year), which can be divided into five classes depending on their type for protein existence (PE) (see Table 1 for official HPP definitions): PE1 (17 470, 86.3%) proteins are identified by the highest stringency criteria including data from mass spectrometry (MS) analysis and antibody identification. PE2 (1660, 8.2%) proteins are identified by expressed mRNA transcripts. PE3 (452, 2.2%) proteins are identified by sequence similarity. PE4 (74, 0.4%) proteins are identified by in silico prediction.[6] Hypothetical gene products, pseudogenes, or proteins suggested from other dubious information are designated PE5 (574, 2.8%) (Figure 1). PE5 proteins form part of the pool of potential human proteins but are excluded from counting the total number of predicted canonical proteins.[6,7] A small number of proteins are promoted from PE5 to PE1 each year, but, for the most part, PE5 proteins will contribute little to the final numbers of human proteins.

In addition to canonical proteins, a vast number of alternative proteoforms produced after splicing, alternate translation initiation sites,[8] proteolytic processing,[9,10,11] and protein post-translational modification (PTM) constitute the human proteome. Proteoforms have diverse and often divergent biological functions in human cells compared with their cognate unmodified parent protein,[12−14] and defining the major protein proteoforms in the human proteome is one of the goals of the C-HPP.[15] Proteins that might eventually be discovered as bona fide human proteins include those candidates encoded by small open reading frame RNAs (smORFs)[16,17] or long noncoding RNA (lncRNA).[18,19] With a few exceptions to date where such products have been discovered and validated as a protein and hence have been classified as PE1 proteins, for example, NBDY (NX_A0A0U1RRE5) and LINC00116 (NX_Q8NCU8), pending credible evidence of widespread smORF and lncRNA protein existence, these are yet to be included in neXtProt.

The 2186 proteins that belong to PE2 to PE4 are designated as missing proteins (MPs) because they lack sufficient experimental evidence of their existence supported by mass spectrometry (MS), antibody detection, or other biological and biochemical measures.[6,7,20,21] Since the C-HPP was officially launched in 2012,[4,5] notable progress has been made in the detection of MPs, with reductions in their numbers from 5511 in 2012 to 2186 in 2018 (neXtProt 2018-01-17 release)[6,7,22,23] reflecting this progress. Although C-HPP investigators played an important role in the annotation, project management, and development of fruitful strategies to achieve MP discovery, this progress was achieved not only by the collaboration of C-HPP teams but also by the contribution of investigators outside the HPP community. Indeed, results from the whole scientific community are regularly curated and integrated by C-HPP team members into the online databases Peptide Atlas[24] and neXtProt,[25] the resources for peptide and protein identifications used as references by the HPP. However, identifying the last MPs is now one of the key rate-limiting factors for the completion of the HPP.

Most MPs have a known function(s) or a predicted function by homology with protein family members. However, using the neXtProt advanced query system (query NXQ_00022 (https://goo.gl/Wf2Qnn)), we retrieved 1937 proteins with no annotated specific function and that account for ∼10% of the total number of human proteins (see below). This remarkable number of proteins implies a vast amount of unidentified new biology. For these proteins that are known to exist (PE1), the functionally uncharacterized proteins were recently termed uPE1[20,21] and there are 1260 uPE1 proteins (neXtProt release 2018-01-17). Moreover, there are also MPs that are both only predicted (PE2−4) and also have no predicted function. These are designated as uMPs, and there are currently 677 uMPs (Figure 1A). Thus the sum of uPE1 (1260) and uMPs (677) accounts for 1937 dark proteins in total. Their distribution across chromosomes is presented in Figure 1B,C. The functional characterization of these proteins is a looming task that must be completed to comprehensively understand the human proteome parts list. At the HPP workshop of the 2017 Annual Meeting of HUPO in Dublin, Ireland, the C-HPP officially adopted uPE1 functionalization as a new pilot project, and this was then embraced by the HPP executive committee. This project aims to characterize the function of up to 50 uPE1 proteins within 3 years in a chromosome-centric manner and to devise a series of experimentally proven workflows and approaches to do so and that can be later recommended to the proteome community as part of a potential new more ambitious initiative of uPE1 characterization. This pilot has already been adopted, and work has commenced by 14 of the national chromosome teams of the C-HPP, funding this work from their own available individual resources.

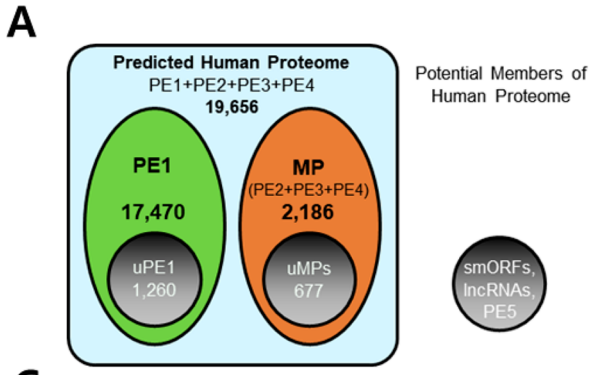
DOI: 10.1021/acs.jproteome.8b00383
J. Proteome Res. 2018, 17, 4042−4050

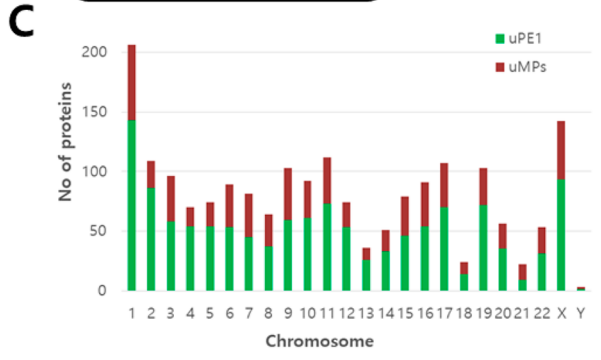
Table 1. Definition of Terms[a]

terms	definition
HPP	The Human Proteome Project (HPP) is an international project organized by the Human Proteome Organization (HUPO) that aims to map, annotate, and functionally characterize the entire human proteome in a systematic way using mass spectrometry complemented by antibody and affinity-based techniques. The HPP extends and is a direct counterpart to the Human Genome Project by annotation of the human genome gene products, so adding significant value and insights into human biology. The HPP is composed of two complementary initiatives: the Chromosome-centric HPP (C-HPP) and Biology/Disease HPP (B/D-HPP). The former focuses on the completion of the "parts list" for proteins and their *proteoforms* whereas the latter aims to make proteomics an integral part of multiomics research throughout the life sciences and biomedical research communities. Both initiatives are supported by four resource pillars: (i) mass spectrometry (MS), (ii) affinity reagents (Ab), (iii) knowledgebase (Kb), and (iv) pathology.
C-HPP	The Chromosome-centric HPP (C-HPP) is an international collaborative project of the HPP that aims to map, annotate, and characterize the human proteome on a chromosome-by-chromosome basis. The 25 international teams from 20 countries use various proteomics technologies to study how the proteome is encoded in Chr 1−22, X, Y, and mitochondrial DNA. Currently, major foci of the C-HPP are to map all remaining *missing proteins* (PE2,3,4 proteins in neXtProt 2018-1-17 = 2186) and characterize 1260 uPE1 (uncharacterized PE) proteins in neXtProt 2018-1-17.
B/D-HPP	The Biology/Disease HPP (B/D-HPP) is an international collaborative project of the HPP that deals with mapping, annotating, and characterizing the proteome using proteomics technologies in relation to human biology and/or diseases. The B/D-HPP provides a framework for the coordination of 22 initiatives that integrate about 50 multinational research groups. A *popular proteins* strategy has been developed to stimulate use of targeted proteomics throughout the life sciences and biomedical community.
PE	Protein existence (PE) levels indicate the degree of evidence of the existence of a human protein based on curated information. The levels PE1 to PE5 are assigned by UniProtKB and *neXtProt* as follows. •PE1: evidence at the protein level (identification by mass spectrometry (MS) according to HPP guidelines, validated antibody (Ab)-based detection, or other characterization). •PE2: evidence at the transcript level (detection by RNaseq or presence of expressed sequence tag). •PE3: inferred by gene homology (assigned membership of a defined protein family). •PE4: predicted protein (not yet assigned membership of a defined protein family). •PE5: uncertain (dubious sequences such as erroneous translation products or pseudogenes). Note: The HPP uses the PE levels assigned by neXtProt to monitor progress made collectively by the scientific community toward the complete experimental validation of the human proteome. In 2013, the HPP excluded PE5 entries from the search for missing proteins.
Missing Proteins	Missing proteins (MPs) are defined as those gene-encoded predicted protein entries in neXtProt categories PE2,3,4 that lack any or sufficient experimental from mass spectrometry or other direct protein methods to qualify as PE1. The MS evidence must meet the HPP MS Data Interpretation Guidelines v2.1 (see PE, above).
Proteoforms	Alternative multiple protein products from the same gene resulting from sequence alterations arising from polymorphisms, alternative splicing, RNA editing, post-translational modifications of amino acid side chains, and proteolytic processing events.
uPE1 Proteins	Uncharacterized PE1 proteins (uPE1s) devoid of any functional annotation in neXtProt or only annotated with broad GO MF/BP terms not linked to any specific function[a]
Dark Proteome	The dark proteome is a colloquial term that includes missing proteins (PE2−PE4), uncertain/dubious predicted proteins (PE5), uPE1 proteins, smORF (small proteins), and any proteins translated by long noncoding RNAs or uncharacterized transcripts including those arising from noncoding regions of DNA and/or novel alternative splicing.
neXt-MP50	A specific two-year C-HPP initiative, announced in September 2016, that aims to accelerate the identification and validation of the existence of 50 currently missing proteins per chromosome team while incorporating progress from throughout the international proteomics community.
neXt-CP50	A specific C-HPP initiative, announced in September 2017, that aims to characterize some cellular function(s) of 50 uPE1 proteins within 3 years by >14 C-HPP working groups.
ProteomeXchange	The ProteomeXchange database was stimulated by the HPP and built at the European Bioinformatics Institute to register and coordinate globally the submission of mass spectrometry proteomics data to the main existing proteomics repositories and to facilitate optimal data set dissemination and access. It includes PRIDE, PeptideAtlas, MassIVE, jPOST, and iProX.
PeptideAtlas	The PeptideAtlas is a public data repository developed by the Institute for Systems Biology that accepts submissions of proteomics mass spectrometry data sets from laboratories all over the world, reprocesses them with the Trans-Proteomic Pipeline suite of software tools, collates all data sets into a global view of the human proteome as observed in MS data sets, informs neXtProt of HPP Guidelines-compliant findings, and makes all of the results publicly available to the community.

[a] https://hupo.org/resources/Documents/HPP%20Scientific%20Terms%20Definitions%20and%20Abbreviations_20180830.pdf (Approved by HPP Executive Committee, Sept 1, 2018).

DOI: 10.1021/acs.jproteome.8b00383
J. Proteome Res. 2018, 17, 4042−4050





B

Chr.	uPE1	uMPs		
		uPE2	uPE3	uPE4
1	143	48	8	7
2	86	17	6	0
3	58	30	5	3
4	54	15	0	1
5	54	16	1	3
6	53	26	4	6
7	45	32	1	3
8	37	24	1	2
9	59	27	6	11
10	61	27	3	1
11	73	35	2	2
12	53	17	2	2
13	26	8	1	1
14	33	14	3	1
15	46	18	14	1
16	54	30	6	1
17	70	31	3	3
18	14	8	0	2
19	72	23	6	2
20	35	16	0	5
21	9	11	1	1
22	31	21	1	0
X	93	28	15	6
Y	1	2	0	0
Total	1260	524	89	64

Figure 1. Components of the human proteome grouped by evidence type distributed chromosome by chromosome (see ref 6 for details). PE, protein evidence; MP, missing protein (proteins belonging to PE2, PE3, and PE4); PE5, dubious or uncertain proteins; uPE1, uncharacterized PE1; uMPs, uncharacterized missing proteins. For color designation, green, brown, and shadow represent "well-identified PE1 proteins", "missing proteins (PE2−4)", and "dark proteins (uPE1 plus uMPs)", respectively.

THE DARK PROTEOME

To the best of our knowledge, the term "dark proteins" was first used to indicate the aggregated form of nonfunctional inclusion bodies in "dark" areas when cellular localization was examined using a fluorescent dye.[28] With respect to protein folding patterns, this term represents the general property of complex proteins, which show an amyloid-type shape in abnormal or diseased cells.[29] However, this term was also adopted by the structural proteomics community when a paper by the O'Donoghue group published the dark proteome annotation, which is based mostly on the protein database (PDB: www.pdb.org). In this paper, Perdigao et al.[30] suggested that dark proteins represent regions of proteins that not only are rarely observed by structure determination but also are inaccessible to homology modeling or other in silico analysis tools. Even more, these dark proteins contain intrinsically disordered protein regions with properties of an order-to-disorder transition (adaptability) during binding to other proteins. This feature remains as a structural enigma but has been suggested to be more likely associated with disease, implying that they comprise commercially valuable candidate drug targets.[31] Thus dark proteins are a desirable target for investigation, so much so that the NIH recently funded and now recruits new scientific projects for dark protein investigation, for example, amyloid aggregates in brain cells (https://grants.nih.gov/grants/guide/rfa-files/RFA-AG-18-025.html). By analogy to the term "dark proteins" coined to represent structurally uncharacterized regions, C-HPP investigators have recently adopted the term "dark proteome" to collectively refer to those proteins for which we have insufficient information on either protein expression, structure, function, or all of these: They include, for example, MPs (PE2−4), PE5, uPE1 proteins, and any potential proteins translated from smORF or lncRNAs. This nomenclature is convenient for project management, which usually requires outreach to the public, granting agencies, and other stakeholders. From the point of view of proteome biology, dark proteins may be depicted as two sides of the same coin: one face for the structural enigma and the other for the functional enigma when cataloging the families of uncharacterized human proteins. Thus the term "dark protein" has evolved to become jargon to designate any protein of unknown structure or function or both.

In neXtProt, the status and numbers of dark proteins have been changing every year, which reflects not only improved understanding of the proteome and proteins but also the moving target nature of dark proteins, which requires constant scrutinized monitoring of new annotations. Notwithstanding their difficulty in functional characterization, dark proteins may nonetheless be potential new cellular regulators, drug targets, and biomarkers.[26,27] For the management of HPP completion, it is cautious to begin with a pilot project before launching a full-scale initiative. Thus on March 1, 2018, the uPE1 functionalization pilot project of the C-HPP was termed the neXt-CP50, where CP stands for "characterization of proteins" and aims to characterize the function of up to 50 uPE1 proteins within 3 years. Of the C-HPP consortium international teams, 15 from 11 countries joined this project: Chr 2 (Switzerland), Chr 3 (Japan), Chr 4 (Taiwan), Chr 9, 11, 13 (Korea), Chr 10, 17 (USA), Chr 14 (France), Chr 15 (Brazil), Chr 16 (Spain), Chr 18 (Russia), Chr 19 (Mexico), Chr 20 (China), and Chr Y (Iran).

SELECTION OF TARGETS AND A STEPWISE APPROACH

To test the feasibility of the functional characterization of large numbers of dark proteins—1937 at present—the 14 teams are focusing on specific tractable targets that can be investigated

DOI: 10.1021/acs.jproteome.8b00383
J. Proteome Res. 2018, 17, 4042−4050

within the 3 year term. Among the dark proteins, we have chosen the uPE1 proteins over uMPs and proteoforms (e.g., novel smORF[15] and lncRNA candidate proteins[17]) as the most promising targets because they are so far the best annotated proteins in accordance with the HPP guidelines.[32] Although these proteoforms have become attractive, they remain on the periphery of acceptance by the proteomics community as bona fide proteins.[33] Therefore, only 50 uPE1 proteins were chosen as the first targets for the neXt-CP50 challenge. Proteoform characterization and the new roles found for the newly promoted uPE1 to PE1 proteins in pathology will likely be integral in deciphering the function for the uPE1 targets during this pilot project or for others that may be done subsequently, thus also meshing with the other goals of the C-HPP, the Biology/Disease-driven Human Proteome Project (B/D-HPP), and the pathology pillar of the HPP.

The neXt-CP50 challenge consists of a limited number of targets (50 uPE1 proteins) to be investigated over a 3 year period using various experimental platforms (Figure 2). The idea behind the installation of the neXt-CP50 challenge prior to a long-term initiative that would eventually target all 1937 dark proteins was the necessity to carefully devise and test stepwise various strategies and workflows according to the individual team's expertise, interest, and collaborations that can be later recommended to the wider community as experimentally validated successful strategies for uPE1 functionalization. In so doing, potential pitfalls for uPE1 characterization and new collaborations of complementary expertise could be established. The success of the challenge will serve as a barometer for the milestones of a potential full-scale project aiming to shed light on all ~2000 dark proteins (Figure 1). In the planning of the future large-scale project, the gathered expertise and experimental success stories can be used to organize the capable teams and to seek public funds for this higher profile potential HPP and HUPO project.

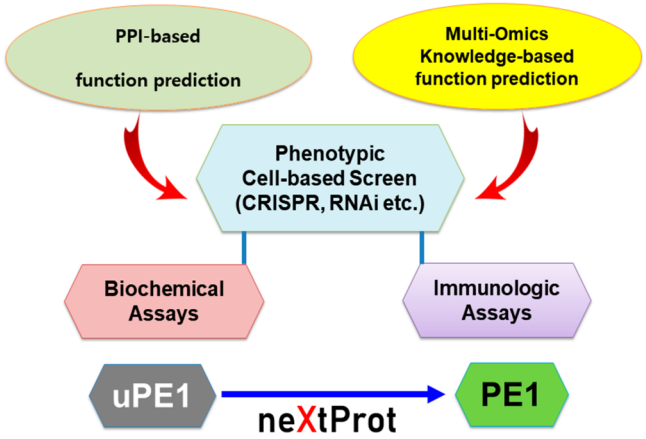


Figure 2. Potential experimental approaches for functional characterization of uPE1 proteins. This can be divided into two workflows, PPI-based function prediction and Multi-Omics Knowledge-based function prediction. The prediction results can further be tested by using phenotypic cell-based screens that can utilize biochemical or immunologic assays for the verification of the prediction results. These experiments can be used in a flexible manner as appropriate to verify and validate the function(s) of target dark proteins. Functionally validated dark uPE1 proteins will be promoted to PE1 and curated by neXtProt.

For the initial neXt-CP50 challenge, the C-HPP is proceeding on a chromosome-team centric basis with the participating national chromosome teams selecting uPE1 target proteins encoded by their respective chromosome. This has proven to be an efficient approach where the currently active national teams can divide the annotation and experimental workload according to chromosome. The 14 participating C-HPP teams each selected three to five chromosome-specific uPE1 proteins in March 2018. These targeted uPE1 proteins are now being subjected to full functional screening by a variety of strategies for detailed characterization of biological function (Figure 2). Feedback from these teams revealed that their C-HPP investigators are scientifically motivated by the reward of the discovery of novel and potentially pathologically important functions for these proteins and hence many have suggested that these are very fundable opportunities with their National Granting agencies.

EXPERIMENTAL STRATEGIES

The functional characterization of dark proteins needs integrative in vitro and in vivo experimental techniques, reagents, knockdown and rescue cells, and mutant model animals lacking specific gene function, with human clinical samples for validation.[34] Initially, the neXt-CP50 projects are expected to be mainly screening exercises to find function, with validation and then characterization of function occurring by various multitiered experiments. Hence, collaborations with B/D−C-HPP based-HPP and pathology pillar investigators are logical to contribute to the validation of the new functions of the uPE1 proteins in human disease.

Potential experimental schemes that could be adopted by individual teams according to in-house expertise, collaborations, and research interest are shown in Figure 2. These are divided into two workflow start points: (1) PPI-based function prediction by inference with known functions of the interacting partners of the uPE1 proteins and (2) Multi-Omics Knowledge-based function prediction. To advance knowledge of protein function using uPE1 protein overexpression or knockdown cells of target uPE1s by some teams, phenotypic cell-based screens (3) can be used to confirm these hypotheses or used as an alternate start point to screen for functions. For validation and further characterization of predicted functions, (4) biochemical assays and (5) immunological cell and tissue localization in healthy and diseased tissues can be employed. Thus these serial stepwise approaches ought to produce new biological entry points to predict or screen for function, followed by validation and characterization studies to promote the uPE1 targets to PE1 proteins (Figure 2).

First, to predict candidate functions for the uPE1 targets from PPI-based function prediction, teams may choose to mine the well-established BioGRID,[35] IntAct which is produced by IMEx consortium,[36] and Meta DB (STRING) databases (Figure 3). In addition, teams may choose to use PPI information from the BioPlex (biophysical interactions of ORFeome-based complexes) network, which is the result of creating thousands of human cell lines (e.g., HEK293T), with each expressing a tagged version of a protein from the ORFeome collection.[37] In the elucidation of PPI for uPE1 proteins, teams can use either immunoprecipitation pulldowns (small scale) or AP−MS or yeast two-hybrid (medium to large scale) methods.[38] These methods will inform potential uPE1 protein (bait) function from the known functions of their biologically relevant interactors. The combination of these methods with phylogenetic and domain analysis may further inform on the type of protein family or pathway, which can then be further characterized biochemically and immunolocalized to further build a unique PPI network where the uPE1 target is physically or functionally involved.

DOI: 10.1021/acs.jproteome.8b00383
J. Proteome Res. 2018, 17, 4042−4050

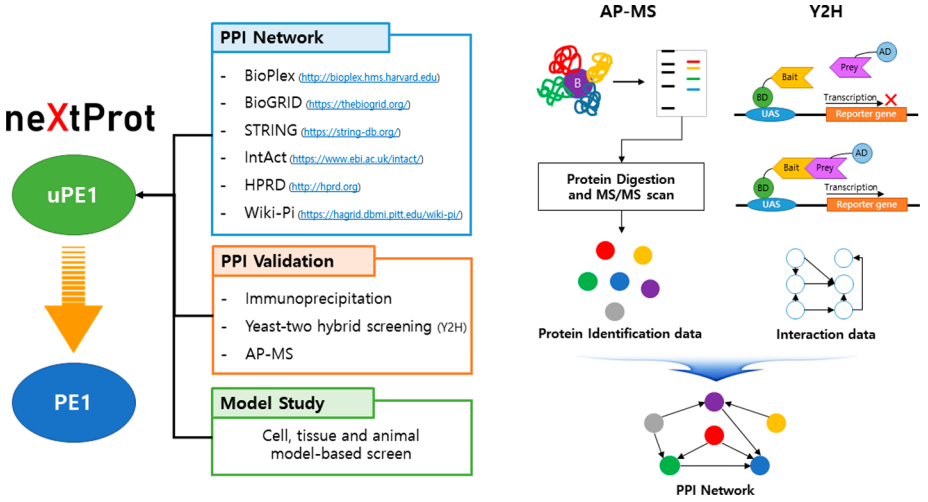


Figure 3. Workflow of PPI-based Function Prediction of uPE1. This workflow utilizes the currently well-established PPI-network DBs in order to predict probable functions of uPE1 based on those protein−protein interaction data. Several components of this workflow are shown, which are interchangeably exercised singly or combined where appropriate. AP-MS, Affinity-purification mass spectrometry.

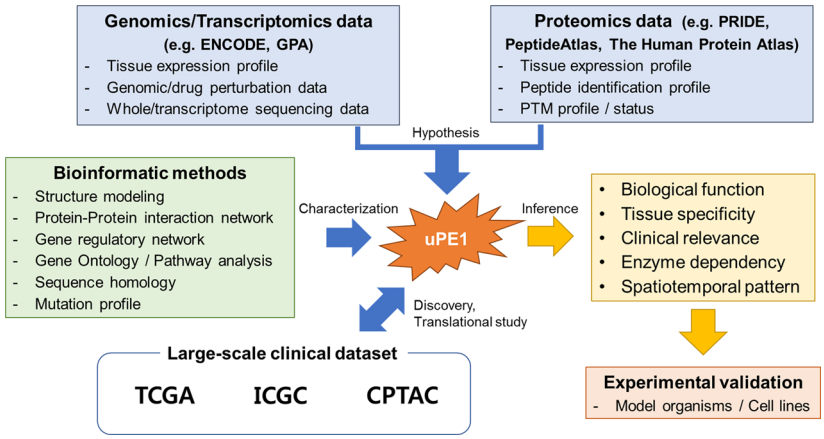


Figure 4. Multi-omics knowledge-based function prediction of uPE1 protein targets. This workflow utilizes the well-established publicly available large-scale or global profiling knowledge base, which may well cover diverse functions of proteins or genes at the molecular levels. These multiomics data are more likely data- and algorithm-dependent.

Second, for the multiomics data integration-based strategy, teams can also access the well-established publicly available large-scale or global profiling knowledge base (Figure 4). This is because various omics data that may cover diverse functions of proteins or genes at the molecular levels are already well constructed that can then be used to mine, integrate, and analyze these well-established DBs to obtain clues for the function of uPE1 proteins.

Teams may also obtain information on protein function using previously published research data or community resources or public databases. For example, with the use of genomic expression profiles deposited in ENCODE[39] and protein expression data in tissues provided by Human Protein Atlas,[40] teams may be able to add supportive evidence of protein function based on where and how uPE1 targets are expressed in specific tissues and cells. Spatiotemporal expression patterns and PTMs of uPE1 available in the PRIDE Archive or PeptideAtlas[41,42] under specific perturbation stimuli[43] (Gene Perturbation Atlas) are also suggested resources for the teams. Such data mining may lead to the prediction of potential regulators or pathways in which uPE1 targets are embedded and hence function. Clinical multi-omics data provided by The Cancer Genome Atlas (TCGA),[44] International Cancer Genome Consortium (ICGC),[45] and Clinical Proteomics Tumor Analysis Consortium (CPTAC)[46] are now accessible. uPE1 teams using these resources may detect clues and further evidence to support hypotheses on the potential structure (e.g., PTM association) and function (mutation, expression change) of uPE1 proteins in disease. To validate these predicted functions experimentally, teams may utilize cell models or model organisms (Figure 2).

Third, supporting the above-described strategies, teams might employ any of the currently available technology or platforms according to in-house expertise and interest where appropriate. Examples are (i) phenotypic screening by pathway inhibition/modulation and sequential assays through various pathway tests after protein/gene knockdown, (ii) Bio-ID or other interactome analyses to provide clues as to binding partners and hence candidate functions based on binders, (iii) genetics-based functional complementation assays, and (iv) the development of novel tools and algorithms for protein function and pathway analysis. On a positive result in the first pass screens, the next experimental tier needs to be invoked for validation experiments by secondary assays using (v) cell-based assays and (vi) biochemical assays to test the function of enzymes, receptors, transporters, and other protein types. On confirmation of function by

DOI: 10.1021/acs.jproteome.8b00383
J. Proteome Res. 2018, 17, 4042−4050

secondary assays, the next step is deeper characterization of the protein function(s) using (vii) model organisms and perhaps orthologous gene knockouts. Finally, (viii) the studies should culminate using human tissue and disease validation including immunolocalization and expression analyses for clinical translation relevance. GWAS, SNP, and expression correlation analyses can also be used to establish disease relevance. Approaches v−viii will not necessarily be exercised in sequence because each uPE1 has a different degree of information in the literature and different challenges. The C-HPP teams will require the acquisition of good resources for samples, recombinant protein, plasmids, antibodies and aptamers, and bioinformatics databases. This latter phase will be facilitated by forging collaborations with the B/D-HPP and the new pathology pillar of the HPP. The following are successful examples of the use of some of these strategies.

The most common methods for human protein functional screens will be the removal or mutation of the target gene or proteins and then functional analysis of the consequences. Two popular methods are available for this purpose, CRISPR/Cas and RNA interference (RNAi), of which the former became the most efficient and fast and had less error in determining protein function[47] (Figure 2). For example, Clift et al.[48] investigated the function of a protein by deleting the target gene with CRISPR/Cas. For a direct example of a uPE1 study using a phenotypic screen after RNAi knockdown, Desmurs et al.[47] characterized the function of C11orf83 (now called UQCC3) as a new assembly factor for the *bc*1 complex and also a stabilizing factor for the III2/IV supercomplex, which is required for proper mitochondrial morphology and function. Screening mutant phenotypic readouts was recently performed in bacteria.[49] Here genome-wide mutant fitness data were used to identify mutant phenotypes for 11 779 protein-coding genes that had not been annotated with a specific function.[50] Unfortunately, only a few of these bacterial genes have homologues in human and could be used to accelerate the neXt-CP50 project. Thus teams can check for the relevance of these approaches by checking for conservation in bacteria.

A previous example of a functional study of uPE1 proteins performed after bioinformatics-based in silico functional prediction along with in vitro assays was reported by Mary et al.[51] With this approach, they first proposed that APIP might have a role in the methionine salvage pathway and then verified its role in HeLa cells by cell-based gene knockdown and biochemical assays.[38] Teams can also design a novel integrative bioinformatics-based tool to predict the function of uPE1 proteins by combining a motif search with structural similarity, surface comparison, and active site template matching. For example, McKay et al.[52] showed that the combination of an in-house bioinformatics tool called ProMol with already existing in silico analytical tools (e.g., Blast, Pfam, and Dali) could be applied to predict the hypothetical function of 65 proteins of unknown function.[52]

For genetics-based functional complementation assays, Lane's group in the Ch 2 team performed cross-species functional prediction analysis (e.g., zebrafish, human), gene knockdown (or overexpression), and gene complementation assays to assign C2orf62 and its interacting partner as key proteins involved in primary ciliogenesis in human cells and the modulation of actin polymerization.[53] Paik's group also employed a genetic complementation assay (e.g., mutant of *C. elegans nrf-2*$^{-/-}$), manipulation of cellular expression (e.g., RNAi knockdown), and biochemical verification in vitro and in vivo to characterize a new function of NHERF1, which is involved in human reproduction.[34] Thus a combination of two or three screens is a promising starting point to characterize the function of the dark uPE1 proteins.

As a novel computational approach that employed PPI-based function prediction, Zhang et al.[54] (in this issue) recently introduced structure and protein interaction-based gene ontology annotations for predicting the functions of uPE1 proteins. The Chromosome 17 team developed a hybrid pipeline that creates protein structure prediction using I-TASSER and infers functional insights for the target protein from the functional templates recognized by COFACTOR. As a case study, they applied the pipeline to all 66 uPE1 encoded by human chromosome 17 (as of neXtProt 2017-07-01). Benchmark testing on a control set of 100 well-characterized proteins randomly selected from the same chromosome showed high Gene Ontology (GO) term prediction accuracies of 0.69, 0.57, and 0.67 for molecular function (MF), biological process (BP), and cellular component (CC), respectively. Three pipelines of function annotations (homology detection, protein−protein interaction network inference, and structure template identification) are exploited by COFACTOR. Detailed analyses show that structure template detection based on low-resolution protein structure prediction made the major contribution to enhancement of the sensitivity and precision of the annotation predictions, especially for cases that do not have sequence-level homologous templates. For the 66 chromosome-17 uPE1 proteins, the I-TASSER/COFACTOR pipeline confidently assigned MF, BP, and CC for 13, 33, and 49 proteins, respectively, with predicted functions ranging from sphingosine *N*-acyltransferase activity and sugar transmembrane transporter to cytoskeleton constitution. The predictions for each of these proteins are tabulated. 13 proteins with confident MF prediction are highlighted; 11 of these 13 are among the 33 with confident BP predictions and 12 are among the 49 with confident CC predictions. This novel computational approach to systematically annotate protein function in the human proteome can be extended to all of the chromosomes and provides useful insights to guide experimental design and follow-up validation studies of these uncharacterized proteins.

■ PERSPECTIVES AND CONCLUSIONS

This C-HPP neXt-CP50 challenge to uncover both the exact count and functions of dark proteins is timely. We anticipate that this pilot project will also mobilize some of the less active groups within the C-HPP or B/D-HPP to increase their involvement in a potential neXt-CP2000 (Figure 5) with this functional discovery effort. Joint efforts by the C-HPP and B/D-HPP to investigate dark proteins is anticipated to add value to the HPP. As for MP identification, we also anticipate that investigators outside HPP will greatly contribute to this endeavor. From the pilot project, we aim to learn much about both the efficiency and bottlenecks for the characterization of protein function that can be performed in a university setting. This is important because many of these approaches are already in use in a pharmaceutical company setting, where greater resources can be brought to bear upon their drug target and drug development plans. These experiences now in this pilot project will inform the neXt-CP2000 work plan if adopted in the future (Figure 5). To make this plan move forward, some immediate action items must be considered: (i) setting criteria for sufficient evidence of claims of dark protein function and (ii) finding the most suitable biological samples and mutant strains of model animals for the

DOI: 10.1021/acs.jproteome.8b00383
J. Proteome Res. 2018, 17, 4042−4050

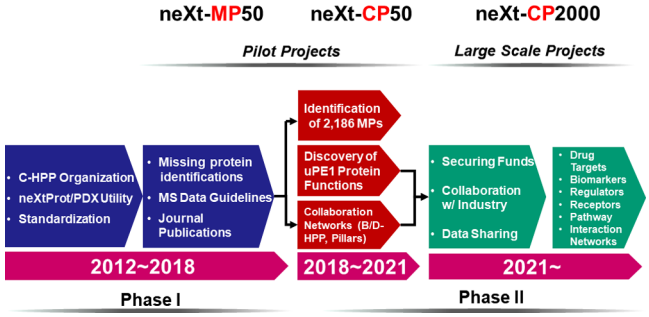


Figure 5. Proposed short- and long-term plan to characterize human dark proteins of unknown functions. The terms neXt-CP50 and neXt-CP2000 stand for characterizing 50 uPE1, a small set of dark proteins, in 3 years (a pilot) and then ~2000 dark proteins (1260 uPE1 plus 677 uMPs) over a longer period of time. The term "CP" stands for characterization of proteins with unknown functions.

mutant phenotypic screening of dark protein function both in vitro and in vivo.

In the long term, we expect that structural biologists and the HPP investigators will work together toward understanding their common targets—the dark proteins, with respect to structure *and* function, as illustrated by the use of I-TASSER and COFACTOR algorithms by Chromosome 17. This pilot project is not just a simple "stamp collection" task. There is no doubt that the functional characterization of dark proteins is much more than annotation and will be beneficial to molecular biology research and biomedical sciences. This new knowledge will enhance understanding of the flow of information from the gene to the phenome, where the proteome is positioned in the middle of this information flow and executes functions essential for life. Thus the results of the neXt-CP50 challenge will enhance the understanding of integrated cellular networks and communication between molecules in cells and tissues in health and disease, identify new drug targets, and generate biomarkers of disease. We predict that the neXt-CP50 challenge will motivate the HPP community in a new mission of understanding human proteome biology and human health.

AUTHOR INFORMATION

Corresponding Authors

*Y.-K.P.: E-mail: paikyk@yonsei.ac.kr.
*C.M.O.: E-mail: chris.overall@ubc.ca.

ORCID

Young-Ki Paik: 0000-0002-8146-1751
Lydie Lane: 0000-0002-9818-3030
Jong Shin Yoo: 0000-0002-8588-3310
Gilberto Domont: 0000-0002-1329-6483
Fernando Corrales: 0000-0002-0231-5159
Gilbert S. Omenn: 0000-0002-8976-6074
Siqi Lui: 0000-0001-9744-3681
Christopher M. Overall: 0000-0001-5844-2731

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank proteomics scientists, leaders of public DBs, C-HPP investigators and associate members, and funding agencies. This paper is dedicated to all of 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. This work was supported by grants from the Korean Ministry of Health and Welfare: [HI13C2098]-International Consortium Project and [HI16C0257] (awarded to Y.-K.P.); from SIB Swiss Institute of Bioinformatics; from the Canadian Institutes of Health Research, 7-year Foundation Grant, and a Canada Research Chair in Protease Proteomics and Systems Biology: [FDN-148408] (awarded to C.M.O.); and National Institutes of Health P30 ES017885 and U24CA210967 (G.S.O.).

ABBREVIATIONS

B/D-HPP, Biology/Disease-driven Human Proteome Project; C-HPP, Chromosome-centric Human Proteome Project; GWAS, genome-wide association study; lncRNA, long non-coding RNA; MP, missing protein; PE, protein evidence or existence; smORF, small open reading frame; uMP, uncharacterized missing protein; uPE1, uncharacterized PE1.

REFERENCES

(1) Collins, F. S.; Morgan, M.; Patrinos, A. The Human Genome Project: lessons from large-scale biology. *Science* **2003**, *300* (5617), 286−90.
(2) Southan, C. Last rolls of the yoyo: Assessing the human canonical protein count. *F1000Research* **2017**, *6*, 448.
(3) Legrain, P.; Aebersold, R.; Archakov, A.; Bairoch, A.; Bala, K. The human proteome project: Current state and future direction. *Mol. Cell. Proteomics* **2011**, *10*, M111.009993.
(4) 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**, *30* (3), 221−3.
(5) Paik, Y. K.; Omenn, G. S.; Uhlen, M.; Hanash, S.; Marko-Varga, G.; et al. Standard guidelines for the chromosome-centric human proteome project. *J. Proteome Res.* **2012**, *11* (4), 2005−13.
(6) Omenn, G. S.; Lane, L.; Overall, C. M.; Corrales, F. J.; Schwenk, J. M.; et al. Progress on Identifying and Characterizing the Human Proteome: 2018 Metrics from the HUPO Human Proteome Project. *J. Proteome Res.* **2018**, DOI: 10.1021/acs.jproteome.8b00441.
(7) Lane, L.; Bairoch, A.; Beavis, R. C.; Deutsch, E. W.; Gaudet, P.; et al. Metrics for the Human Proteome Project 2013−2014 and strategies for finding missing proteins. *J. Proteome Res.* **2014**, *13* (1), 15−20.
(8) Na, C. H.; Barbhuiya, M. A.; Kim, M. S.; Verbruggen, S.; Eacker, S. M.; et al. Discovery of noncanonical translation initiation sites through mass spectrometric analysis of protein N termini. *Genome Res.* **2018**, *28* (1), 25−36.
(9) Klein, T.; Eckhard, U.; Dufour, A.; Solis, N.; Overall, C. M. Proteolytic Cleavage-Mechanisms, Function, and "Omic" Approaches for a Near-Ubiquitous Posttranslational Modification. *Chem. Rev.* **2018**, *118* (3), 1137−1168.
(10) Fortelny, N.; Pavlidis, P.; Overall, C. M. The path of no return–Truncated protein N-termini and current ignorance of their genesis. *Proteomics* **2015**, *15* (14), 2547−52.
(11) Smith, L. M.; Kelleher, N. L. Proteoforms as the next proteomics currency. *Science* **2018**, *359* (6380), 1106−1107.
(12) LeDuc, R. D.; Schwammle, V.; Shortreed, M. R.; Cesnik, A. J.; Solntsev, S. K.; et al. ProForma: A Standard Proteoform Notation. *J. Proteome Res.* **2018**, *17* (3), 1321−1325.
(13) Schaffer, L. V.; Shortreed, M. R.; Cesnik, A. J.; Frey, B. L.; Solntsev, S. K.; et al. Expanding Proteoform Identifications in Top-Down Proteomic Analyses by Constructing Proteoform Families. *Anal. Chem.* **2018**, *90* (2), 1325−1333.
(14) Aebersold, R.; Agar, J. N.; Amster, I. J.; Baker, M. S.; Bertozzi, C. R.; et al. How many human proteoforms are there? *Nat. Chem. Biol.* **2018**, *14* (3), 206−214.
(15) Paik, Y. K.; Omenn, G. S.; Overall, C. M.; Deutsch, E. W.; Hancock, W. S. Recent Advances in the Chromosome-Centric Human

DOI: 10.1021/acs.jproteome.8b00383
J. Proteome Res. 2018, 17, 4042−4050

Proteome Project: Missing Proteins in the Spot Light. *J. Proteome Res.* **2015**, *14* (9), 3409−14.
(16) Saghatelian, A.; Couso, J. P. Discovery and characterization of smORF-encoded bioactive polypeptides. *Nat. Chem. Biol.* **2015**, *11* (12), 909−16.
(17) McFedries, A.; Schwaid, A.; Saghatelian, A. Methods for the elucidation of protein-small molecule interactions. *Chem. Biol.* **2013**, *20* (5), 667−73.
(18) Banfai, B.; Jia, H.; Khatun, J.; Wood, E.; Risk, B.; et al. Long noncoding RNAs are rarely translated in two human cell lines. *Genome Res.* **2012**, *22* (9), 1646−57.
(19) Ruiz-Orera, J.; Messeguer, X.; Subirana, J. A.; Alba, M. M. Long non-coding RNAs as a source of new peptides. *eLife* **2014**, *3*, No. e03523.
(20) Paik, Y. K.; Omenn, G. S.; Hancock, W. S.; Lane, L.; Overall, C. M. Advances in the Chromosome-Centric Human Proteome Project: looking to the future. *Expert Rev. Proteomics* **2017**, *14* (12), 1059−1071.
(21) Paik, Y. K.; Overall, C. M.; Deutsch, E. W.; Van Eyk, J. E.; Omenn, G. S. Progress and Future Direction of Chromosome-Centric Human Proteome Project. *J. Proteome Res.* **2017**, *16* (12), 4253−4258.
(22) Marko-Varga, G.; Omenn, G. S.; Paik, Y. K.; Hancock, W. S. A first step toward completion of a genome-wide characterization of the human proteome. *J. Proteome Res.* **2013**, *12* (1), 1−5.
(23) Omenn, G. S.; Lane, L.; Lundberg, E. K.; Overall, C. M.; Deutsch, E. W. Progress on the HUPO Draft Human Proteome: 2017 Metrics of the Human Proteome Project. *J. Proteome Res.* **2017**, *16* (12), 4281−4287.
(24) Farrah, T.; Deutsch, E. W.; Hoopmann, M. R.; Hallows, J. L.; Sun, Z.; et al. The state of the human proteome in 2012 as viewed through PeptideAtlas. *J. Proteome Res.* **2013**, *12* (1), 162−71.
(25) Gaudet, P.; Michel, P. A.; Zahn-Zabal, M.; Britan, A.; Cusin, I.; et al. The neXtProt knowledgebase on human proteins: 2017 update. *Nucleic Acids Res.* **2017**, *45* (D1), D177−D182.
(26) Roth, B. L.; Kroeze, W. K. Integrated Approaches for Genome-wide Interrogation of the Druggable Non-olfactory G Protein-coupled Receptor Superfamily. *J. Biol. Chem.* **2015**, *290* (32), 19471−7.
(27) Delgado, A. P.; Brandao, P.; Chapado, M. J.; Hamid, S.; Narayanan, R. Open reading frames associated with cancer in the dark matter of the human genome. *Cancer Res.* **2014**, *11* (4), 201−13.
(28) Iafolla, M. A.; Mazumder, M.; Sardana, V.; Velauthapillai, T.; Pannu, K.; et al. Dark proteins: effect of inclusion body formation on quantification of protein expression. *Proteins: Struct., Funct., Genet.* **2008**, *72* (4), 1233−42.
(29) Schnabel, J. Protein folding: The dark side of proteins. *Nature* **2010**, *464* (7290), 828−9.
(30) Perdigao, N.; Heinrich, J.; Stolte, C.; Sabir, K. S.; Buckley, M. J.; et al. Unexpected features of the dark proteome. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112* (52), 15898−903.
(31) Perdigao, N.; Rosa, A. C.; O'Donoghue, S. I. The Dark Proteome Database. *BioData Min.* **2017**, *10*, 24.
(32) Deutsch, E. W.; Overall, C. M.; Van Eyk, J. E.; Baker, M. S.; Paik, Y. K.; et al. Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 2.1. *J. Proteome Res.* **2016**, *15* (11), 3961−3970.
(33) Verheggen, K.; Volders, P. J.; Mestdagh, P.; Menschaert, G.; Van Damme, P.; et al. Noncoding after All: Biases in Proteomics Data Do Not Explain Observed Absence of lncRNA Translation Products. *J. Proteome Res.* **2017**, *16* (7), 2508−2515.
(34) Na, K.; Shin, H.; Cho, J. Y.; Jung, S. H.; Lim, J.; et al. Systematic Proteogenomic Approach To Exploring a Novel Function for NHERF1 in Human Reproductive Disorder: Lessons for Exploring Missing Proteins. *J. Proteome Res.* **2017**, *16* (12), 4455−4467.
(35) Stark, C.; Breitkreutz, B. J.; Reguly, T.; Boucher, L.; Breitkreutz, A.; Tyers, M. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* **2006**, *34*, D535−9.
(36) Hermjakob, H.; Montecchi-Palazzi, L.; Lewington, C.; Mudali, S.; Kerrien, S.; et al. IntAct: an open source molecular interaction database. *Nucleic Acids Res.* **2004**, *32*, D452−5.
(37) Huttlin, E. L.; Ting, L.; Bruckner, R. J.; Gebreab, F.; Gygi, M. P.; et al. The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* **2015**, *162* (2), 425−440.
(38) Sardiu, M. E.; Washburn, M. P. Building protein-protein interaction networks with proteomics and informatics tools. *J. Biol. Chem.* **2011**, *286* (27), 23645−51.
(39) The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **2012**, *489* (7414), 57−74.
(40) Uhlen, M.; Fagerberg, L.; Hallstrom, B. M.; Lindskog, C.; Oksvold, P.; et al. Tissue-based map of the human proteome. *Science* **2015**, *347* (6220), 1260419.
(41) Vizcaino, J. A.; Csordas, A.; Del-Toro, N.; Dianes, J. A.; Griss, J.; et al. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **2016**, *44* (22), 11033.
(42) Deutsch, E. W. The PeptideAtlas Project. *Methods Mol. Biol.* **2010**, *604*, 285−96.
(43) Xiao, Y.; Gong, Y.; Lv, Y.; Lan, Y.; Hu, J.; et al. Gene Perturbation Atlas (GPA): a single-gene perturbation repository for characterizing functional mechanisms of coding and non-coding genes. *Sci. Rep.* **2015**, *5*, 10889.
(44) Weinstein, J. N.; Collisson, E. A.; Mills, G. B.; Shaw, K. R.; Ozenberger, B. A.; Ellrott, K.; Shmulevich, I.; Sander, C.; Stuart, J. M. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **2013**, *45* (10), 1113−20.
(45) Zhang, J.; Baran, J.; Cros, A.; Guberman, J. M.; Haider, S.; Hsu, J.; Liang, Y.; Rivkin, E.; Wang, J.; Whitty, B.; Wong-Erasmus, M.; Yao, L.; Kasprzyk, A. International Cancer Genome Consortium Data Portal–a one-stop shop for cancer genomics data. *Database* **2011**, *2011*, bar026.
(46) Ellis, M. J.; Gillette, M.; Carr, S. A.; Paulovich, A. G.; Smith, R. D.; Rodland, K. K.; Townsend, R. R.; Kinsinger, C.; Mesri, M.; Rodriguez, H.; Liebler, D. C. Clinical Proteomic Tumor Analysis, C., Connecting genomic alterations to cancer biology with proteomics: the NCI Clinical Proteomic Tumor Analysis Consortium. *Cancer Discovery* **2013**, *3* (10), 1108−12.
(47) Desmurs, M.; Foti, M.; Raemy, E.; Vaz, F. M.; Martinou, J. C.; et al. C11orf83, a mitochondrial cardiolipin-binding protein involved in bc1 complex assembly and supercomplex stabilization. *Mol. Cell. Biol.* **2015**, *35* (7), 1139−56.
(48) Clift, D.; McEwan, W. A.; Labzin, L. I.; Konieczny, V.; Mogessie, B.; et al. A Method for the Acute and Rapid Degradation of Endogenous Proteins. *Cell* **2017**, *171* (7), 1692−1706.e18.
(49) Price, M. N.; Wetmore, K. M.; Waters, R. J.; Callaghan, M.; Ray, J.; et al. Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature* **2018**, *557* (7706), 503−509.
(50) Jinek, M.; Chylinski, K.; Fonfara, I.; Hauer, M.; Doudna, J. A.; et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **2012**, *337* (6096), 816−21.
(51) Mary, C.; Duek, P.; Salleron, L.; Tienz, P.; Bumann, D.; et al. Functional identification of APIP as human mtnB, a key enzyme in the methionine salvage pathway. *PLoS One* **2012**, *7* (12), e52877.
(52) McKay, T.; Hart, K.; Horn, A.; Kessler, H.; Dodge, G.; et al. Annotation of proteins of unknown function: initial enzyme results. *J. Struct. Funct. Genomics* **2015**, *16* (1), 43−54.
(53) Bontems, F.; Fish, R. J.; Borlat, I.; Lembo, F.; Chocu, S.; et al. C2orf62 and TTC17 are involved in actin organization and ciliogenesis in zebrafish and human. *PLoS One* **2014**, *9* (1), e86476.
(54) Zhang, C.; Wei, X.; Omenn, G. S.; Zhang, Y. Structure and Protein Interaction-Based Gene Ontology Annotations Reveal Likely Functions of Uncharacterized Proteins on Human Chromosome 17. *J. Proteome Res.* **2018**, DOI: 10.1021/acs.jproteome.8b00453.

DOI: 10.1021/acs.jproteome.8b00383
J. Proteome Res. 2018, 17, 4042−4050