A Quest for Missing Proteins: update 2015 on Chromosome-Centric Human Proteome Project

Péter Horvatovich[1] (p.l.horvatovich@rug.nl), Emma K Lundberg[2] (emma.lundberg@scilifelab.se), Yu-Ju Chen[3] (yujuchen@gate.sinica.edu.tw), Ting-Yi Sung[4] (tsung@iis.sinica.edu.tw), Fuchu He[5] (hefc@bmi.ac.cn), Edouard C. Nice[6] (ed.nice@monash.edu), Robert J Goode[6] (robert.goode@monash.edu), Simon Yu[6] (xiaoming.yu@monash.edu), Shoba Ranganathan[7] (shoba.ranganathan@mq.edu.au), Mark S. Baker[8] (mark.baker@mq.edu.au), Gilberto B Domont[9] (gilberto@iq.ufrj.br), Erika Velasquez[9] (erika_velasquez2706@hotmail.com), Dong Li[6] (lidong.bprc@foxmail.com), Siqi Liu[10] (siqiliu@genomics.cn), Quanhui Wang[10] (wangqh@genomics.cn), Qing-Yu He[11] (tqyhe@email.jnu.edu.cn), Rajasree Menon[12] (rajmenon@umich.edu), Yuanfang Guan[13] (yuanfang.guan@gmail.com), Fernando J. Corrales[14, 15] (fjcorrales@unav.es), Victor Segura[14, 15] (vsegura@unav.es), J. Ignacio Casal[14, 15] (icasal@cib.csic.es), Alberto Pascual-Montano[14, 15] (pascual@cnb.csic.es), Juan P. Albar[14, 15] (jpalbar@proteored.org)[**], Manuel Fuentes[16] (mfuentes@usal.es), Maria Gonzalez-Gonzalez[16] (mariagg@usal.es), Paula Diez[16] (pauladg@usal.es), Nieves Ibarrola[16] (nibarrola@usal.es), Rosa M Degano[16] (romade@usal.es), Yassene Mohammed[17, 18] (Yassene@proteincentre.com), Christoph H. Borchers[17] (christoph@proteincentre.com), Andrea Urbani[19, 20] (andrea.urbani@uniroma2.it), Alessio Soggiu[21] (alessio.soggiu@unimi.it), Tadashi Yamamoto[22] (tdsymmt@med.niigata-u.ac.jp), Alexander Archakov[23] (alexander.archakov@ibmc.msk.ru), Elena Ponomarenko[23] (2463731@gmail.com), Andrey Lisitsa[23] (lisitsa.av@gmail.com), Cheryl F. Lichti[24] (cflichti@utmb.edu), Ekaterina Mostovenko[24] (ekamosto@utmb.edu), Roger A. Kroes[25] (r-kroes@northwestern.edu), Melinda Rezeli[26] (melinda.rezeli@bme.lth.se), Ákos Végvári[26] (akos.vegvari@bme.lth.se), Thomas E. Fehniger[26] (thomas.fehniger@bme.lth.se), Rainer Bischoff[1] (r.p.h.bischoff@rug.nl), Juan Antonio Vizcaíno[27] (juan@ebi.ac.uk), Eric W Deutsch[28] (edeutsch@systemsbiology.org), Lydie Lane[29, 30] (lydie.lane@isb-sib.ch), Carol L. Nilsson[24](carol.nilsson@utmb.edu), György Marko-Varga[26] (Gyorgy.Marko-Varga@bme.lth.se), Gilbert S. Omenn[31] (gomenn@med.umich.edu), Seul-Ki Jeong[32] (jeongsk@proteomix.org), Jin-Young Cho[32] (chojy@proteomix.org), Young-Ki Paik[32] (paikyk@yonsei.ac.kr), William S Hancock[33] (wi.hancock@neu.edu)*

[1]Analytical Biochemistry, Department of Pharmacy, University of Groningen, A. Deusinglaan 1, 9713 AV Groningen, The Netherlands

[2]Science for Life Laboratory, KTH - Royal Institute of Technology, SE-171 21 Stockholm, Sweden

[3]Institute of Chemistry, Academia Sinica, Taipei, Taiwan

[4]Institute of Information Science, Academia Sinica, Taipei, Taiwan

[5]Beijing Proteome Research Center, Beijing, China

[6]Department of Biochemistry and Molecular Biology, Monash University, Clayton, Victoria 3800, Australia

[7]Department of Chemistry and Biomolecular Sciences and ARC Centre of Excellence in Bioinformatics, Macquarie University, Sydney, NSW 2109, Australia

[8]Australian School of Advanced Medicine, Macquarie University, Sydney, NSW 2109, Australia

[9]Federal University of Rio de Janeiro, Proteomics Unit, Department of Biochemistry, Institute of Chemistry[10]Beijing Institute of Genomics and BGI Shenzhen

[11]Key Laboratory of Functional Protein Research of Guangdong Higher Education Institutes, College of Life Science and Technology, Jinan University, Guangzhou 510632, China

[12]Department of Computational Medicine & Bioinformatics, University of Michigan, Ann Arbor, MI, 48109-2218, USA

[13]Departments of Computational Medicine & Bioinformatics and Computer Sciences, University of Michigan, Ann Arbor, MI, 48109-2218, USA

[14]ProteoRed-ISCIII. Biomolecular and Bioinformatics Resources Platform (PRB2), Spanish Consortium of C-HPP (Chr-16), CIMA, Spain.

[15]Chr16 SpHPP consortium

[16]Department of Cellular and Molecular Medicine. Centro de Investigaciones Biológicas (CIB-CSIC), Madrid, Spain

[17]Centro Nacional de Biotecnologia (CNB-CSIC). Cantoblanco. Madrid. Spain

[16]Cancer Research Center. Proteomics Unit and General Service of Cytometry. Department of Medicine. University of Salmanca-CSIC. IBSAL. Campus Miguel de Unamuno s/n. 37007, Salamanca. Spain

[17]University of Victoria - Genome British Columbia Proteomics Centre, Vancouver Island Technology Park, #3101 – 4464 Markham St., Victoria, BC V8Z 7X8, Canada

[18]Center for Proteomics and Metabolomics, Leiden University Medical Center, 2333 ZA, Leiden, The Netherlands

[19]Proteomics and Metabonomic, Laboratory, Fondazione Santa Lucia, Rome, Italy
[20]Department of Experimental Medicine and Surgery, University of Rome "Tor Vergata", Rome, Italy
[21]Department of Veterinary Science and Public Health (DIVET), University of Milano, Milano, Italy
[22]Institute of Nephrology, Graduate School of Medical and Dental Sciences, Niigata University, Niigata
[23]Orechovich Institute of Biomedical Chemistry, Moscow, Russia
[24]Department of Pharmacology and Toxicology, The University of Texas Medical Branch, Galveston, TX 77555-1074, U.S.A.
[25]Northwestern University, Evanston, IL, USA
[26]Clinical Protein Science & Imaging, Department of Biomedical Engineering, Lund University, BMC D13, 221 84 Lund, Sweden
[27]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, CB10 1SD, Hinxton, Cambridge, UK.
[28]Institute for Systems Biology, 401 Terry Avenue North, Seattle, WA 98109, USA
[29]SIB Swiss Institute of Bioinformatics, Geneva, Switzerland
[30]Department of Human Protein Science, Faculty of medicine, University of Geneva, Switzerland
[31]Gilbert S. Omenn, Departments of Computational Medicine & Bioinformatics, Internal Medicine, Human Genetics and School of Public Health, University of Michigan, Ann Arbor, MI, USA, 48109-2218
[32]Departments of Integrated Omics for Biomedical Science & Biochemistry, College of Life Science and Technology, Yonsei Proteome Research Center, Yonsei University, Seoul, 120-749, Korea
[33]The Barnett Institute of Chemical and Biological Analysis, Northeastern University, 140 The Fenway, Boston 02115, Massachusetts, United States

*Corresponding Author: p.l.horvatovich@rug.nl; Tel: +31-50-363-3341; fax: +31-50-363-7582.

** The authors want to pay tribute to Juan Pablo Albar, a friend and recognized scientist, who passed away in July 2014 in Madrid during preparations of the HUPO Congress 2014.

TITLE RUNNING HEAD: A quest for missing proteins

Keyword: missing proteins, chromosome centric human proteome project, LC-MS, antibody enrichment, proteomics, bioinformatics

ABBREVIATIONS

ASV	Alternative Splicing Variant
ATAQS	Automated and Targeted Analysis with Quantitative SRM
B/D HPP	Biology/Disease driven Human Proteome Project
CAPER	Chromosome Assembled human Proteome browser
CHCPC	Chinese Human Chromosome Proteome Consortium
C-HPP	Chromosome Centric Human Proteome Project
CPTAC	Clinical Proteomic Tumor Analysis Consortium
DHS	DNase I hypersensitivity
EBI	European Bioinformatics Institute
ENCODE	Encyclopedia of the DNA Elements
ETD	Electron Transfer Dissociation
EThcD	combination of electron-transfer dissociation (ETD) and higher-energy collision dissociation (HCD)
FDR	False Discovery Rate
GEO	Gene Expression Omnibus
GPCR	G-protein coupled receptors
gpmDB	Global Proteome Machine database
GSC	*Glioma* stem cells
HPA	Human Protein Atlas
HPPP	Human Plasma Proteome Project
HTML	HyperText Markup Language
HUPO	HUman Proteome Organisation
ISBER	International Society for Biological and Environmental Repositories

JSON	JavaScript Object Notation
MATF	Monash Antibody Technologies Facility (Monash University)
MS	Mass Spectrometry
NIST	National Institute of Standards and Technology
ORF	Open Reading Frame
PASSEL	PeptideAtlas SRM Experiment Library
PE	Protein Existence
PrESTs	Protein Epitope Signature Tags
PSM	Peptide Spectrum Match
REST	Representational State Transfer
SAV	Single Amino acid Variant
SILAC	Stable Isotope Labeling by Amino Acids
SISCAPA	Stable Isotope Standard Capture with Anti-Peptide Antibodies
SOP	Standard Operating Procedure
SPARQL	Protocol and RDF Query Language
SPE	Solid Phase Extraction
SRM	Single Reaction Monitoring
SWATH-MS	Sequential Window Acquisition of all Theoretical Fragment ion Mass Spectrometry
TPP	Trans Proteomic Pipeline
XML	Extensible Markup Language

WEB LINKS

BioPortal	http://bioportal.bioontology.org/
C-HPP Web	http://www.c-hpp.org
C-HPP Wiki	http://c-hpp.webhosting.rug.nl/
dasHPPboard	http://sphppdashboard.cnb.csic.es/
Disease Ontology	http://disease-ontology.org/
neXtProt REST service	www.nextprot.org/rest/
Protannotator	http://biolinfo.org/protannotator
Ontology Lookup Service	http://www.ebi.ac.uk/ontology-lookup/

Abstract

This paper summarizes the recent activities of the Chromosome-Centric Human Proteome Project (C-HPP) consortium, which develops new technologies to identify yet-to-be annotated proteins (termed "missing proteins") in biological samples that lack sufficient experimental evidence at the protein level for confident protein identification. The C-HPP also aims to identify new protein forms that may be caused by genetic variability, post-translational modifications, and alternative splicing. Proteogenomic data integration forms the basis of the C-HPP's activities; therefore, we have summarized some of key approaches and their roles in the project. We present new analytical technologies that improve the chemical space and lower detection limits coupled with bioinformatics tools and some publicly available resources that can be used to improve data analysis or support the development of analytical assays. Most of this paper's contents have been compiled from posters, slides, and discussions presented in the series of C-HPP workshops held during 2014. All data (posters, presentations) used are available at the C-HPP Wiki (http://c-hpp.webhosting.rug.nl/) and in the supporting information.

Introduction

Proteins such as those acting as enzymes, regulatory proteins, transporters, and receptors are the active macromolecules of human biology and are thus central to understanding biological molecular processes. Understanding the diversity and complexity of these biological molecular interactions is a central focus of bio-medical research today. It is important that all protein forms of human genes are eventually studied so that their biological functions and roles in healthy and disease states can be determined [**posters 17** and **18** in **Table 1**][1-3]. Proteins cannot be amplified and are chemically much more heterogeneous than DNA and RNA. Their analysis therefore represents a much more significant analytical challenge.

To meet this challenge, the Human Proteome Organization (HUPO) announced in 2010 at the HUPO Congress in Sydney, Australia, the formation of the Human Proteome Project (HPP) to sequentially catalogue the protein products of human genes, both to identify proteins that have little or no evidence at the protein level, termed "missing proteins"[4, 5], and to discover and characterize protein sequence variability with genetic origin and post-translational modifications (PTM) of known proteins. The Chromosome-Centric Human Proteome Project (C-HPP)[4-6] is a large multidisciplinary international effort to identify all human protein forms and catalogue them on the basis of the chromosome location of their coding genes. In the C-HPP, one national or multinational team is responsible for the identification and annotation of protein products of the genes in each chromosome. Evidence at the protein level means that a protein has been detected, preferably by mass spectrometry (MS) and preferably from multiple peptides unique to the protein observed in multiple experiments and possibly in diverse biological sample types. Multiple data resources help to provide information on protein evidence.

The sharing of data, protocols, and other electronic resources such as proteome annotations is crucial in the C-HPP and proteomics community to enable the reuse of collected information (e.g., in form of high-quality spectral libraries for spectral library identification or the development of selected reaction monitoring [SRM] assays) and to develop and improve new data analysis protocols. The current status and developments in the shared proteomics data and proteome knowledge by the main stakeholders are presented below. An overview of current bioinformatics resources and the data flow to support the creation of complete part lists of the human proteome is shown in **Figure 1.**

The ProteomeXchange[7] consortium, led by PRIDE[8, 9] at the European Bioinformatics Institute (Hinxton/Cambridge, UK) and by PeptideAtlas[10-12] at the Institute for Systems Biology (Seattle, Washington, USA), is devoted to the standardization of data submission and dissemination of MS-based proteomics data and to the promotion of public sharing of proteomics data in the public domain.

In addition, it promotes the use of community data standards developed by the Proteomics Standards Initiative (PSI). As such, ProteomeXchange resources store original MS datasets, containing at least the raw MS data accompanied by the processed results (peptide and protein identifications, but possibly quantitative information as well) and by suitable experimental and technical metadata. By November 2014, ProteomeXchange resources stored approximately 1500 datasets (~50% of which are publicly available) from a wide variety of sources, with humans being the most-represented species. Once the datasets are made publicly available, they are usually reprocessed by PeptideAtlas[10-12] using the Trans Proteomic Pipeline (TPP)[13-17] and by the Global Proteome Machine Database (gpmDB)[18, 19] using the X!Tandem[20, 21] database search tool. Basically, everyone in the community can reanalyze the raw data available in ProteomeXchange for different purposes.

To provide a comprehensive view of the human proteome and its diversity, neXtProt[22, 23] is adding information at the genomic, transcriptomic and proteomic levels to the corpus of information available in UniProtKB[24-26]. In particular, neXtProt integrates genomic variation data from dbSNP[27] and COSMIC[28], transcriptomic data from BGee[29], antibody-based protein evidence from the Human Protein Atlas (HPA)[30, 31], MS-based information from PeptideAtlas[10-12], three-dimensional structural information from the Protein Data Bank[32] and various PTM information from manually curated literature. Based on this combined information, neXtProt attributes a protein existence (PE) level to each entry originally defined in UniProtKB. The PE1 level (experimental evidence at the protein level) denotes entries with credible evidence by protein expression and identification by MS, immunohistochemical analysis, three-dimensional structure, or amino acid sequencing. The PE2 level (experimental evidence at transcript level) refers to proteins with transcript expression evidence but without evidence of protein detection. The PE3 level (protein inferred from homology) is attributed to proteins without human protein or transcript evidence but with strong evidence on homologous protein in another species. The PE4 level (protein predicted) is for proteins that are hypothesized from gene

models, and the PE5 level (protein and gene uncertain) refers to "dubious" or "uncertain" genes that at one time seemed to have some protein-level evidence but have since been deemed doubtful. The PE5 category generally corresponds to pseudogenes or non-coding RNAs according to the protein annotation from different resources (HGNC[33], RefSeq[34, 35], HAVANA, CCDS[36-38], UniProtKB/Swiss-Prot[39]). Among the 643 entries in the PE5 category in neXtProt in August 2011, 119 have already become obsolete in UniProtKB and have been deleted from neXtProt, 13 have been upgraded to the PE1 category due to manual curation of publications and/or convincing proteomics data, and 11 have been upgraded to the PE2 or PE3 categories. Based on these numbers, one can estimate that less than 5% (< 30) of the remaining PE5 proteins in neXtProt are true proteins. Given this low probability, any MS identification of PE5 proteins must be carefully checked. Proteins in the PE2-4 categories are awaiting experimental confirmation at the protein level and are called "missing proteins" in the context of C-HPP.

One of the primary tasks of the C-HPP is to determine why no protein products have been identified for certain genes showing open reading frame for translation, i.e., genes coding for the so-called "missing proteins with no or poor protein evidence." There are five main reasons for the existence of "missing proteins" (**Figure 2**)[40]. (1) The current mainstream proteomics technology cannot identify them, possibly because of the low abundance of the proteins, because the sequences do not contain tryptic cleavage sites or generate peptides which can uniquely identify the proteins, or because the protein digestion results of peptides that are lost during the sample preparation and analysis. (2) They are expressed only in rarely studied tissues or cell types, or are expressed only as a result of a stimulus or perturbation. (3) They are not expressed at all and are part of the silent information of the human genome. (4) They reflect erroneous annotation of the genome, which results in incorrectly predicted protein sequences. (5) Many highly homologous proteins or proteins with large sequence variability are missed or not counted due to the parsimonious protein assembly of shotgun MS/MS protein

identification or due to large sequence variability such as immunoglobulins. PeptideAtlas and gpmDB select only one "representative protein" among highly homologous members of protein families[40] when the available sequence coverage cannot distinguish these related proteins (see the Cedar scheme in Farrah et al.[41] and **Figure S1** in the supporting information).

When the C-HPP began (2012), it was announced that no satisfactory evidence existed at the protein level for 6568 (33%) of the 20,059 protein-coding genes[42]. NeXtProt released a new version as of September 19, 2014, that contains 20,055 entries, among which 16,491 are PE1 proteins, 2948 lack protein evidence (PE2, PE3, PE4), and 616 are dubious (PE5). Hence, there is still insufficient evidence at the protein level for approximately 15% (if we exclude PE5) or 18% of the human proteome. It is encouraging to see that the number of missing proteins has been reduced considerably since the initial assessment by the C-HPP in 2012.

This paper presents an overview of the new technologies and new resources that have been used during the last 4 years by members of the C-HPP consortium and others to identify missing proteins. Most of the contents described here have been obtained from data presented during several HUPO workshops in 2014, including the 9th, 10th and 11th C-HPP Workshops in Busan, South-Korea (26 March), Bangkok, Thailand (9 August) and Segovia, Spain (9 October), respectively, and during the C-HPP and HPP sessions of the HUPO 2014 Congress in Madrid, Spain (5-8 October).

Proteogenomics

Analytical technologies and bioinformatics are the key components for the identification and quantification of proteins in a complex biological sample. The current workhorse of proteomics analysis is shotgun LC-MS/MS, typically using a C18 stationary phase and acetonitrile/water eluent pairs, resulting in sequence coverage typically lower than 30% for identified proteins. Additionally, most of the collected MS/MS spectra contain gaps in the fragment ion series, thus preventing de novo

peptide sequence spectra interpretation and identification[43]. The most widely used approach for protein identification is database searching, which requires a list of protein sequences that are expected to be present in the analyzed samples. UniProtKB is the most frequently used protein sequence database. It has two main components: UniProtKB/SwissProt (which contains manually curated sequences) and UniProtKB/TrEMBL[44] (which contains computationally generated records from DNA sequences that have not been manually curated). Canonical sequences are used to represent the most prevalent sequences that are most similar to those of other species and in which the length or amino acid composition allows the clearest description of protein domains, splice isoforms, polymorphisms, and PTMs. UniProtKB contains some degree of protein isoform information and sequence variation due to genetic variability, but it is far from complete. This results in a low level of identification of peptides that arise from alternative splicing, coding non-synonymous single-nucleotide polymorphisms (SNPs), and single amino acid variants (SAVs) due to RNA editing[45, 46]. Therefore a proteogenomic approach using DNA and mRNA data to build a protein database that contains all genetic variability or sample-specific protein sequence information is becoming more and more popular and has contributed to the identification of new protein forms[47]. Conversely, peptide-level data can serve to fill gaps or correct errors in the DNA and RNA databases[47, 48]. There are two main proteogenomic approaches: (1) a protein sequence database is constructed from publicly available databases that contain sequence information with genetic variability such as dbSNP[27, 49] or H-INVDB[50], or (2) a customized protein sequence database is constructed from annotated DNA and mRNA transcript data obtained from the same sample[51-53]. However, proteogenomic analysis generally results in a larger database than does using the UniProtKB canonical sequences, and should be followed through false discovery rate (FDR) analysis at both the peptide and protein levels[13, 47], especially when the database search is performed in multiple steps[47, 54]. PeptideProphet[55] from TPP and Percolator[56-58] can be used for FDR calculations for peptide spectrum matches (PSMs). ProteinProphet[59] and MAYU[17] (both from TPP) serve to estimate

FDR for protein inference. PeptideShaker[60, 61] provides a solution for both PSM and protein FDR calculation. The statistical power of PSM in separating correct and incorrect PSM distributions can be enhanced by including the measurable and predictable physico-chemical properties of peptides in addition to *m/z*, such as the retention time in liquid chromatography[62] or the high-resolution isoelectric point[63]. Identification of missing proteins can be enhanced by identifying cell lines and tissue samples with transcriptomics evidence [**poster 12**], analyzing samples of different ages, and including samples acquired under special stress conditions and biological perturbations.

A general drawback of bottom-up shotgun LC-MS/MS approaches is that complete protein forms cannot be reconstituted from peptide fragments. A top-down approach that allows the peptide-protein interference problem to be avoided may provide a solution for determination of the accurate distribution of whole protein forms, also called proteoforms[64]. Proteoforms are the most recent nomenclature of protein forms introduced by the Top Down Proteomics Consortium, which "designates all of the different molecular forms in which the protein product of a single gene can be found, including changes due to genetic variations, alternatively spliced RNA transcripts and post-translational modifications". The relationship of the proteoform terminology to the UniProt canonical sequences and other protein sequence variability or modifications is shown in **Figure S2** in the supporting information. Importantly, unlike bottom-up protocols in which detailed information on PTMs and sequence variants is compromised because of enzymatic digestion, intact proteins are analyzed in top-down approaches, which allows the unequivocal identification and location of specific modifications. However, they require relatively pure protein samples, they are restricted to proteins of less than 30 kDa, the available fragmentation spectra are often far from complete, and the obtained complex spectra are often difficult to interpret[64-67].

Analysis of mRNA has an advantage in that sequences can be amplified to provide nearly complete sequence coverage using current RNA sequencing technologies. The challenge is to accurately annotate

the resulting raw DNA and RNA data, which is generally performed using the Ensembl genome browser[68]. Ensembl contains a reference genome and includes annotation from the Encyclopedia of the DNA Elements (ENCODE)[69, 70] which is a "comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active." However, protein-coding gene annotations such as GENCODE[71] are based on the protein sequences stored in public databases such as UniProtKB or NCBI RefSeq[35] and gene models that predict the long open reading frames (ORFs) that are most likely to code a protein, which can lead to errors in the annotation of these databases. Therefore, besides revealing protein forms due to genetic variability, a proteogenomic approach can help to confirm the existence of the 616 dubious human proteins currently annotated as PE5 in neXtProt. It can also support identification of new ORFs and translated non-coding mRNA, or redefine the starting and ending parts of protein coding regions, as reported by Kim et al[1, 72]. However when protein identification is performed exclusively with a translated mRNA sequence the much shorter half-life of mRNA compared to proteins[73, 74] should be taken into account in the integration of proteogenomics data. The half-life difference between these two molecular species could result in proteins without mRNA when proteins and mRNA are measured in the same sample and at a single time point. Time series sampling could be used to overcome this issue, when it is possible. This is the case for cell cultures, blood or animal experiments, or tissues for which multiple samples are available from the same specimen at different time. For other cases, the use of a combined databases from translated mRNA sequences and the UniProt database is an option for the detection of proteins with a half-life much longer than that of mRNA.

Translating mRNA, which is directly upstream of protein expression, thus serves as a useful resource for protein identification[53]. Wang et al.[75] performed the first translated mRNA sequencing (RNC-seq) in human lung cancer cell lines and observed an improved correlation of RNC-mRNA abundance with

translated protein when the RNC-mRNA length was taken into consideration. The same group showed that the genes with translation evidence represent an improved reference for the identification of proteins, the detection of sequence variations (SAV, RNA editing and alternative splicing), and integration of the MS data[53].

Furthermore, missing proteins with mRNA evidence and more stringent conditions with ribosome-bound mRNA (RNC-mRNA) evidence are most probably translated, but the current proteomics technology does not allow their detection because of a restricted chemical space or because the detection sensitivity is not sufficient. According to the presentation from Zhang et al. (submitted manuscript) at the C-HPP workshop during the HUPO 2014 Congress in Madrid, ~5% of transcribed mRNAs are typically not translated in a single cell line, and these non-translated mRNAs are highly cell-type specific and/or tissue specific. This allows the focus to be placed on missing proteins in samples with translation evidence and the development of targeted SRM assays and specific sample preparation methods, e.g., the use of antibody enrichment of missing proteins for low abundant peptides or the use of different proteases when missing proteins do not contain identifiable unique tryptic peptides with the detected tryptic peptide set.

An example of a proteogenomic study in which translated mRNA analysis, proteomics data integration, and the use of antibodies were performed to enrich low abundant proteins was presented by Chang et al.[52] from the Chinese Human Chromosome Proteome Consortium covering chromosomes 1, 8, and 20. In their study, three hepatocellular carcinoma cell lines (Hep3B, HCCLM3, and MHCC97H) were submitted for mRNA and RNC-mRNA analysis and to comprehensive analysis with deep proteomics and antibody-enriched transcription factor proteomics. Based on the integrated data, they concluded that only 50.2% of the protein-coding genes with translation evidence were found in the proteomic data. This result is comparable to that of a previous study on the RNC-mRNA and MS data of Caco-2 cells: 52.6% of the proteins with translation evidence were missing from the LC-MS/MS data acquired

from institutions[53]. The inability to detect certain proteins by LC-MS/MS was most probably a results of the translation control mechanisms and analytical limitations of MS-based shotgun identification of peptides and proteins. This warrants a survey of missing proteins in other resources and strategies, such as detergent-insoluble fractions of cell/tissue lysates and forced gene expression through epigenetic manipulations.

Integrating alternatively spliced transcripts with proteomics information allows the study of the transcriptional regulation of proteins in both healthy and diseased tissue [**poster 8** and **13**]. This effect was shown by Menon et al.[76], who integrated RNA-seq and proteomics data as part of the chromosome 17 team and identified more than one splice variant for each of 1167 genes expressed in at least one of three breast cancer cell line models ERBB2+SKBR3, ERBB2+ SUM190, and EGFR(ERBB1)+SUM149, of hormone receptor–negative breast cancers. Their data analysis showed high differences between alternative splicing distributions in the three different cell lines, which were distinctively enriched for different key cell functions such as amino acid and sugar metabolism, caspase activity, and endocytosis in SKBR3; aspects of metabolism, especially of lipids in SUM190; and cell adhesion, integrin and ERK1/ERK2 signaling, and translational control in SUM149. In **poster 20**, Menon and Omenn presented findings of recurrent non–canonical splice variants of interesting proteins in 126 triple-negative breast cancer specimens using data available from EBI/PRIDE.

Another dimension of the proteogenomic splice isoform studies was presented in **poster 21** by Li et al, who undertook genome-wide isoform-level protein connectivity analysis. The isoform with the highest connectivity seems to be more highly associated with function than the choice of a canonical protein isoform based on the sequence length or the abundance of the isoform, which are the methods commonly used in established databases. The genome-wide isoform analysis in mice has been reported[77] and is under development for humans by the Chromosome 17 team (Li et al., unpublished).

Glioma stem cells (GSCs) isolated from patient tumors possess both stem-like and oncogenic patterns of protein expression and are thus a potential source of missing proteins. The Chromosome 19 team has characterized their expression profiles at both the transcript and protein levels. They analyzed 1382 chromosome 19 genes in GSCs using a transcription microarray and showed that 70-75% of them were expressed in each of the studied cell lines[78]. The customized analyses identified differential gene expression patterns specific to chromosome 19 between subtypes of GSCs. It was found that roughly 20% of the transcripts were differentially expressed in the proneural and classical subtypes in comparison to transcription patterns in human neuronal stem cells[79]. The chromosome 19 transcripts that potentially encoded candidate unidentified ORFs proteins were also investigated; 43 ORFs were represented on the arrays, of which 31 (72%) were expressed in the GSC lines. GSCs are also a source of protein variants. Recently, proteomic searches of high-resolution LC-MS/MS data of GSC protein digests identified 19 SAVs in 17 chromosome 19 proteins[1, 2]. Several of the protein variants may have oncogenic potential and are the subjects of further investigation. Furthermore, the integration of RNA-seq and proteomic data made possible the study of the somatic-proteomic landscape of GSCs, thereby allowing the contribution of new knowledge regarding novel fusion proteins in GSC pathobiology. To summarize the current status of chromosome 19, **Figure 3** shows the numbers of genes, mRNA, and proteins, including the number of "missing" proteins and the number of predicted molecular forms (such as mutant proteoforms) and known PTMs. This study and the preceding studies from the Chromosome 17 team illustrate the potential of involving new protein forms that arise from genetic variability and alternative splicing into the investigation of new biology.

Enlarging the analyzed chemical space

Proteins are composed of 20 amino acids and are known to be modified by more than 300 types of PTMs[80-82], embracing a wide chemical space that should be covered by the proteomics analytical approach. In addition, artificial modifications introduced by the sampling protocol need to be

considered. This large chemical space is well covered, but not completely ~~covered~~, by the widely used acetonitrile/water C18 LC-MS/MS protocols. For example, studies in multiple tissues and cells lines performed by the Chinese Human Chromosome Proteome Consortium[52] showed that hydrophobicity (28%) and a low molecular mass (<30 kDa; 75%) are important physicochemical properties that predict unsuccessful detection of a protein. In contrast, the isoelectric point and half-life do not seem to play important roles in detectability. Unidentified proteins in hepatocellular carcinoma cell lines were enriched in specific cellular processes such as olfaction with non-liver function or mainly localize in the cell membrane, supporting the hydrophobicity-negative bias of the currently dominant method of proteomics analysis. Tissue transcript analysis showed that transcripts for the missing proteins are abundant in the testis. Interestingly, a recent analysis of data in the HPA has shown that more tissue-specific proteins are made in the testis than in any other tissue in the body[83]. Analysis of the DNase I hypersensitivity of mRNA and RNC-mRNA data suggests that the missing proteins without a detectable signal are relatively enriched in the chromatin regions with low DNase I hypersensitivity (~40% of the missing proteins), which suggests that the specific structure of chromatin can repress the transcriptional process. Chromosome 11 (and to a lesser extent chromosome 19) showed a greater number of missing proteins without transcript evidence, and those missing proteins were densely clustered in several well-defined chromosome regions. One major group of these missing proteins is presumed to have olfaction function.

Missing protein identification can be enhanced by developing specific enrichment methods such as the use of Proteominer beads[84, 85] and enrichment of protein aggregates (Yang Chen, Yaxing Li, Jiayong Zhong, et al.; manuscript under review in JPR); specific analytical methods for hydrophobic proteins; a specific fractionation method such as the analysis of subcellular fractions [**poster 5, 9** and **19**]; and methods to increase protein sequence coverage (e.g., by using multiple proteases for protein cleavage[49] or by using a more efficient method of peptide fragmentation such as EThcD[86-88]). The membrane

subproteome was suggested to be a rich source of missing proteins. A deep sequencing strategy using complementary two-dimensional chromatography with a combination of high-pH reversed phase (RP), strong anion exchange and low-pH RP stationary phases was used to increase the measured dynamic concentration range. The preliminary results of the enriched membrane proteome from the group led by Yu-Ju Chen [**poster 1**] showed that high-pH RP columns enhanced the retention of hydrophobic peptides and increased the identification coverage of the missing membrane proteins (unpublished results).

Lowering the detection limit with targeted SRM, SWATH analysis, ProteomeAnalyzer, and antibody enrichment

SRM assays have been used for decades to quantify small compounds by MS. The laboratory of Ruedi Aebersold has further developed this approach into a standard method for proteomics to enable simultaneous multiplexed quantification of several hundred proteins in complex biological samples with a wide concentration dynamic range. Picotti et al.[89, 90] showed the power of this method by detecting almost the complete proteome of yeast, covering 4.5 orders of magnitude of the dynamic concentration range. Large-scale application of SRM assays for targeted quantification of long human protein lists required not only the increased speed of the triple quadrupole instruments, but also the creation of such important informatics resources as high-quality spectral libraries (e.g., NIST spectral libraries, SRMAtlas[91]), repository of SRM assay results (PASSEL[92]), a database of ranked peptides and SRM transitions for all proteins in selected proteomes (SRMAtlas[91]), and a database of peptides and transitions with quantification calibration curves (SRMQuantAtlas). SRM assay development requires the identification of proteotypic peptides that not only map uniquely to a single protein or isoform but also are readily ionized and can be detected by MS with a high probability. The proteotypic sequence and SRM transitions must be unique to unequivocally identify the protein form among all other protein forms in the human proteome. This task, coupled with the processing and analysis of the acquired data,

is supported by step-specific algorithms and comprehensive bioinformatics pipelines[93] to plan SRM assays for missing proteins, such as ATAQS[94], mQuest[95], MaRiMba[96], SMRBuilder[97], and Skyline[98]. The PeptidePicker tool developed by Mohammed Y et al.[99] can help to select the most appropriate surrogate peptides for a given protein list in human and mouse proteomes to be used in targeted SRM assays based on the current knowledge of the community as presented in UniProtKB, PeptideAtlas, gpmDB, PRIDE and dbSNP. The tool identified has already reported peptides in online databases for missing proteins, although the quality of the data in these databases varies considerably.

The data-independent sequential window acquisition workflow (SWATH-MS) allows collection of non-targeted fragment spectra by fragmenting large windows of precursor ions (typically 20 to 25 ~~Da~~m/z). The resulting MS/MS data can be reconstituted from the co-eluted fragment ions with liquid chromatography retention time using deconvolution methods. The SWATH approach also can be seen as a generalization of the SRM approach, in which each detectable fragment ion is measured and can be reconstituted from the acquired data without being restricted to a targeted list of transitions as in SRM. Recently, SWATHAtlas was introduced, which stores a human library of MS/MS spectra acquired on a TripleTOF instrument for 10,000 human proteins[100]. This library was obtained from 331 measurements on cell lines, blood, and other human tissues and is intended to be used by PeakView, the OpenSWATH tool[101], and other analogous processing software, providing 51% of coverage of canonical UniProtKB/Swiss-Prot[39] entries.

Another important resource for the identification of missing proteins is SRMAtlas[91], which contains a high-confidence “gold standard” quality SRM assay for at least one unique peptide for 99.9% of the canonical UniProtKB/Swiss-Prot[39] entries. This high coverage was achieved by including MS/MS spectra obtained from a large campaign of production and analysis of synthetic peptides for the complete human proteome. Another source of MS/MS spectra and spectral libraries for phosphorylated and unmodified synthetic peptides is available for assay development[102].

The development and application of SRM assays to complex biological samples is a well-established technology for protein quantification that requires expensive instrumentation and experienced personnel, which limits its utility in replacement of the commonly used western blot analysis for quantification of proteins. Following a planning period led by the HUPO Industrial Advisory Board and a survey of 266 participants, mostly from biology and clinically oriented laboratories, HUPO launched the ProteomeAnalyzer initiative in collaboration with instrument vendors with the goal of developing affordable SRM instrumentation capable of quantifying of 100 to 2000 proteins.

Antibodies are effective reagents for the specific detection and enrichment of missing proteins[103]. The availability of highly specific and validated antibodies is crucial for the detection of low abundant missing proteins and the spatial characterization of their expression pattern in cells and tissues. The implementation of high-throughput production of validated high-affinity monoclonal antibodies using automated production systems will provide renewable resources[104, 105]. SISCAPA can enhance the sensitivity of SRM analyses by enriching specific peptides[106-108].

The HPA[31, 109] project has systematically generated affinity purified polyclonal antibodies using proteospecific recombinant protein fragment and Protein Epitope Signature Tags (PrESTs)[110]. After a rigorous validation scheme, the approved antibodies are used to assess the spatial distributions of the proteins in a multitude of human cells and tissues by immunohistochemical analysis. The November 2014 HPA release (version 13.0) contains more than 13 million images of protein expression patterns generated by the use of 23,968 validated antibodies targeting 16,943 genes. In addition to protein evidence, expression levels, and subcellular localization, the HPA contains mRNA expression levels for the majority of tissues and cell lines involved in the HPA[111]. The resources from HPA are highly valuable for the identification of cell lines and tissues that express missing proteins or for cross-validation of MS or HPA antibody protein evidence. Methods for the use of PrEST antigens as spike-in reagents for quantitative MS were recently demonstrated[112]. Immuno-SILAC has proved capable of

absolute quantification of proteins in complex samples based on HPA antibodies and stable isotope-labeled PrESTs to allow affinity enrichment before MS analysis and accurate quantification[113].

In a recent collaboration between the HPA group in Stockholm and the high throughput monoclonal antibody facility at Monash University in Melbourne a number of monoclonal antibodies against missing proteins, important signaling molecules and proteins of interest to the Chromosome 7 and 17 groups were generated using the same PrESTs as immunogens, which will allow a direct comparison between monoclonal and polyclonal antibodies raised against the same prEST and generate new reagents for the proteomics community. Interestingly, in some cases it was possible to raise monoclonal antibodies to targets that had failed to generate polyclonals. This finding provides an additional route for completion of the task of generating renewable antibodies to all human proteins using the existing antigen resources. Lambert et al.[114] recently showed that coupling affinity enrichment with quantitative MS techniques such as SWATH analysis provides the most sensitive detection method for low abundant missing proteins.

Human sample resources

Human samples are collected and stored in various locations worldwide and are crucial to the C-HPP project and to proteomics and disease research in general. Even if sensitive analytical methods are available to uniquely identify and detect missing proteins, high-quality human samples collected under strict standard operating procedures for collection, processing, and storage must be available to characterize protein expression. Although many countries have recognized this need and have established biobanks for the collection and storage of human samples available from local or regional resources, they have not always been collected under the optimal conditions required for the maintenance of the initial integrity of the protein constituent of samples for proteomics studies. Here, many factors leading to protein degradation need to be identified and addressed by the community. Therefore, as demonstrated by several groups[115-121], sample collection and storage protocols should be

assessed and optimized in this respect for each sample type. For the C-HPP initiative, in addition to ensuring the sample quality, it is also important to exchange samples between laboratories in different countries, for which legal and ethical regulations should be in place. To facilitate the exchange of samples, HUPO will join forces with ISBER[122, 123], an international organization that has worked out regulatory and ethical protocols and Best Practice guidelines[124] for such purposes.

Controlled vocabularies and ontologies pioneered by SNOMED[125] providing standardised anatomical descriptors related to tissue types (BRENDA)[126, 127], cell types (Cell Ontology)[128], and human diseases (DOID, http://disease-ontology.org/)[129] and common descriptions of clinical details, sampling, sample handling, and sample storage data are crucial to effectively compare and search metadata of the samples stored in biobanks and to enable studies that make use of samples from multiple biobanks. Biology and clinically related ontologies are accessible through the Ontology Lookup Service[130] hosted at EBI (http://www.ebi.ac.uk/ontology-lookup/) or at BioPortal (http://bioportal.bioontology.org/).

Integration of the HUPO Biology/Disease Human Proteome (B/D HPP) and C-HPP initiatives will be beneficial for both consortia because C-HPP can provide new assays for missing proteins or protein isoforms whose role and function can be immediately studied by B/D HPP teams in the context of health and disease. G-protein–coupled receptors, and especially olfactory receptors, are overrepresented among the missing proteins. This protein family is low abundant and shows highly specific tissue expression, and expression of the approximately 900 human olfactory receptors that are responsible for the detection of odorant compounds is only expected in nasal tissue. An assessment of the number of identified olfactory receptors in Kim et al.[72] and Wilhelm et al.[131] by Ezkurdia et al.[132] showed that these two large-scale studies with poor MS/MS spectra identified more than 100 olfactory receptors, despite the fact that they did not include data from nasal tissue. This quality assessment shows the importance of critical error analysis of peptide and protein identification in large-scale data analysis projects. The use of a 1% threshold for FDR limited only to PSM or peptide levels is not

sufficient to provide a high-quality list of identified proteins in large aggregated datasets. Therefore the statistical criteria must be a 1% FDR or better calculated at the protein level for the combined dataset as adopted by PeptideAtlas[11, 17]. Using a 1% FDR threshold at the PSM or peptide level would result in a large number of misidentified or indistinguishable proteins when analyzing a large amounts of data. These incorrect PSMs map to proteins randomly, which results in a greater FDR at the protein level. Setting an FDR should take into account the number of identified peptides and proteins in large datasets. For example, if a million PSM pass a threshold of 1% FDR, this implies that there are 10,000 false PSMs, and these tend to map to proteins with one peptide per protein, which results in large FDR at the protein level. For datasets from which 3000 proteins are identified, a 1% protein-level FDR implies only 30 incorrect protein identifications. However, for very large datasets from which 15,000 proteins are identified, a 1% protein-level FDR would result in 150 misidentified proteins, which is a considerable number. In this case, lowering the FDR to 0.1% for example, would keep the number of misidentified proteins at more acceptable number of approximately 15. C-HPP will stringently identify olfactory receptors in nasal tissue accompanied with thorough FDR analysis at the PSM, peptide and protein levels.

Bioinformatics resources

High-level bioinformatics support is crucial for the success of the C-HPP initiative and goes beyond the already-listed sequence knowledge bases, MS databases and SRM assay development support, and evaluation pipelines. Many groups have developed Human Proteome Browsers to support the chromosome-centric integration, processing, and visualization of proteogenomic data or MS/MS repositories such as the Gene-Centric Knowledgebase[133, 134], GenomeWideDB[135] [**poster 4**], Human Proteome Map[72], proteomicsDB[131], gpmDB[18], PeptideAtlas[11, 12, 91], HPA[6, 30, 31, 109], The Proteome Browser[136], CAPER[137, 138] [**poster 2**], and Human Proteinpedia[139-141]. These resources are currently

being developed in isolation, which makes it difficult to further interrogate the diverse types of information stored in these resources. With the participation of the major database developers listed previously (**Figure 1**), an initiative to create at Unified Human Proteome Browser [**poster 16**] as an advanced knowledge-mining system was established at HUPO 2014 in Madrid. This builds on the strengths of existing browsers and their development teams to provide a unified platform for further detailed analysis of the acquired proteogenomic data from the perspectives of chromosomes, biology, and disease. This will lead to a better overview of the existing proteogenomic information that can be developed to suit the needs of the global proteomics community and to improve the current standards of data processing, visualization, and interpretation. It will be essential to subject the component resources and their overall performance to comparisons of assumptions, methods, or findings.

The importance of the quality of bioinformatics workflows and use of false-discovery thresholds was demonstrated by Eric Deutsch, who showed that the addition of four large datasets (the CPTAC repository[142] and those of Kim et al.[72], Wilhelm et al.[131], and Guo et al.[143]) to PeptideAtlas[11, 12, 91] only increased the amount of level 1 protein evidence for approximately 1365 neXtProt entries using stringent error thresholds of 0.000091 FDR for PSMs, 0.00028 FDR at peptide level, and 0.011 FDR at protein level identification. The successive increments in HumanAll build database of PeptideAtlas from these new large studies were 541, 591, 231, and 2 proteins. gpmDB, PeptideAtlas, and neXtProt each estimated the high-quality protein identifications from Kim et al.[39] and Wilhelm et al.[60] to be about 13,000, not 17,294 or 18,059, as reported. Further scientific scrutiny of the many reasons for these large discrepancies will be desirable, involving all parties, as launched in Madrid.

The proteomics community has a great deal of experience with over-calling protein identifications when stringent FDR thresholds are not maintained. The sensitivity to protein matching protocols can be illustrated with the results from the HUPO Human Plasma Proteome Project (HPPP). The original HPPP team paper[144] highlighted a “core dataset” of 3020 proteins with two or more peptide matches,

but clearly delineated a broad range of values with other criteria. In contrast, States et al.[145] published a uniquely stringent version of the same heterogeneous data utilizing Bonferroni-type adjustment for multiple comparisons with 889 protein identifications. In 2011, Farrah et al.[41] published a Cedar scheme (**Figure S1** in the supporting information) for HPPP that demonstrated stepwise recognition of 1929 canonical proteins (1% protein-level FDR) + 236 possibly distinguished, totaling 2165 not subsumed; + 2507 subsumed = 4672 peptide-set unique; + 5686 indistinguishable = 9358 sequence-unique; + 10,102 identical = 19,460 exhaustive list (suitable for cross-checking a different canonical set to see whether the match was lost in the choice of a "representative protein"; see **Figure S1** in the supporting information). By 2014 the Human Plasma Proteome had grown to 4005 canonical proteins, as documented in the comparison of kidney, urine, and plasma proteomes[10].

The Spanish chromosome 16 team developed a method using transcription data from public repositories (GEO[146, 147]) obtained with cancer samples, cell lines, and healthy tissues to identify samples that showed enrichment for missing proteins [**poster 12**]. The data analysis showed that 2861 missing protein-coding genes were expressed at the mRNA level in at least one sample, and that the majority of the genes showed sample specificity. Their study confirmed that the missing proteins are typically shorter and of lower abundance than those that have been identified. Transmembrane, cytoskeleton, signal transduction, spermatogenesis, zinc finger domains, synapses, neurotransmitter activity, and olfactory transduction are enriched cellular functions among the missing proteins[148]. All of these data will be available in the dasHPPboard webtool (http://sphppdashboard.cnb.csic.es/) [**poster 10**], which has the goal of creating a similar initiative for storing and accessing the processed data generated by the C-HPP projects in a manner similar to that of ENCODE[46, 47].

To support the C-HPP initiative, Islam et al.[149] developed the Protannotator tool to provide extensive annotation of missing proteins. Protannotator consists of a generic pipeline incorporating bioinformatics and annotation tools to identify homologues and to map putative functional signatures,

gene ontology, and biochemical pathways. Sequential BLAST searches originally developed for chromosome 7[150] can be used to identify homologues from nonhuman mammalian proteins with strong protein evidence or homologues with validated human proteins. The Protannotator tool identified nonhuman mammalian homologues with protein evidence for 1271 missing proteins in other mammalian species, and 564 missing protein sequences were homologues to the reviewed human proteins. Functional annotations for the remaining missing proteins support the identification of possible biological sources and conditions under which the remaining missing proteins may be expressed. The tool also generates in silico proteotypic peptides, which facilitate the development of SRM assays. A search of these proteotypic peptides in ENCODE[46, 47] revealed proteomic evidence for 107 missing proteins, with evidence for an additional 15 missing proteins using the data of a recent membrane proteomic study[149].

NeXtProt provides primarily web-based protein evidence information, but also enables retrieval of data in various output formats (HTML, JavaScript Object Notation [JSON] and XML) using the REST Application Programming Interface (www.nextprot.org/rest/). In addition, neXtProt provides "chromosome reports" on its ftp server to support C-HPP projects. At the workshop in Segovia, the neXtProt team announced the development of an advanced search engine based on SPARQL that will enable complex and powerful queries, including federated queries with external resources[151].

Conclusions

The reduction of the proportion of missing proteins in the human proteome from 33% to 18% (or 15%) over the last four years shows the clear progress of the C-HPP, which is mainly due to the application of improved proteomics technology such as specific sample preparation (e.g., antibody-based enrichment and enrichment of hydrophobic peptides), the use of advanced spectrometers, the application of SRM and SWATH assays for missing proteins [**poster 3** and **14**], and the analysis of unusual human sample types [**posters 6, 7, 11, 13,** and **15**]. As the results approach saturation of the

parts list for protein-coding genes, it will be ever more important to apply stringent FDR criteria to the claims of protein matches and to confirm the findings with orthogonal methods. "One-hit wonders," especially of short peptides, and claims of matches in tissue or cell types without transcript expression or that have not previously shown evidence of such proteins with modern instruments should be viewed with skepticism. The quality of the spectra must be examined, keeping in mind that when Ezkurdia et al.[132] examined the spectra for hundreds of olfactory receptor proteins claimed by Kim et al.[72] and by Wilhelm et al.[131], none survived scrutiny. Likewise, peptides with multiple matches may be more likely to represent known, highly expressed proteins with a single mutation or an RNA-edited site than a "missing protein." The C-HPP has also encouraged analyses of amplicons (*cis*-regulated genes in specific chromosomal segments) and of protein families, as well as the recognition of proteins that are unlikely to be detected for the reasons outlined in **Figure 2**.

Proteogenomic analysis integrating data from genomics, transcriptomics, and proteomics is gaining momentum and results in an addition to the human proteome protein forms that arise from genetic variability, such as SAVs, RNA-editing, and alternative splicing. Proteogenomic technology now allows the routine study of these new protein forms in biological processes to unravel their roles in various diseases. Spectral libraries of synthetic peptides for almost all human proteins, together with the large number of antibodies generated by the HPA, permits the functional analysis of proteins and protein forms in biological experiments with complex designs. Bioinformatics support for the C-HPP has been largely developed during the last four years and has contributed to its success not only by reducing the number of missing proteins, but also in aiding the discovery of multiple new protein forms.

It is clear that work must still be undertaken to confirm the presence of the remaining missing proteins, which will become more and more challenging as the completion of the MS-based evidence of the human proteome on the gene basis is reached. C-HPP members are increasing their activities to find

evidence for the remaining missing human proteins and to discover more and more complete sets of protein forms that reflect genetic variability and post-translational modifications.

The C-HPP posters presented at HUPO 2014 in Madrid (**Table 1**) are available online at the *Journal of Proteome Research* as supporting information, including the poster's abstract, and most of the oral presentations can be found at C-HPP Wiki (http://c-hpp.webhosting.rug.nl/).

Acknowledgements

J.A.V. acknowledges the EU FP7 grants 'ProteomeXchange' [grant number 260558] and PRIME-XS [grant number 262067]. G.S.O. acknowledges grant U54ES017885 from the NIH. Carol L. Nilsson acknowledges the Cancer Prevention and Research Institute of Texas (CPRIT, RML 1122) and the University of Texas Medical Branch. Y.K.P acknowledges the C-HPP grant from the Korean Ministry of Health and Welfare (to Y.K.P., HI13C2098).

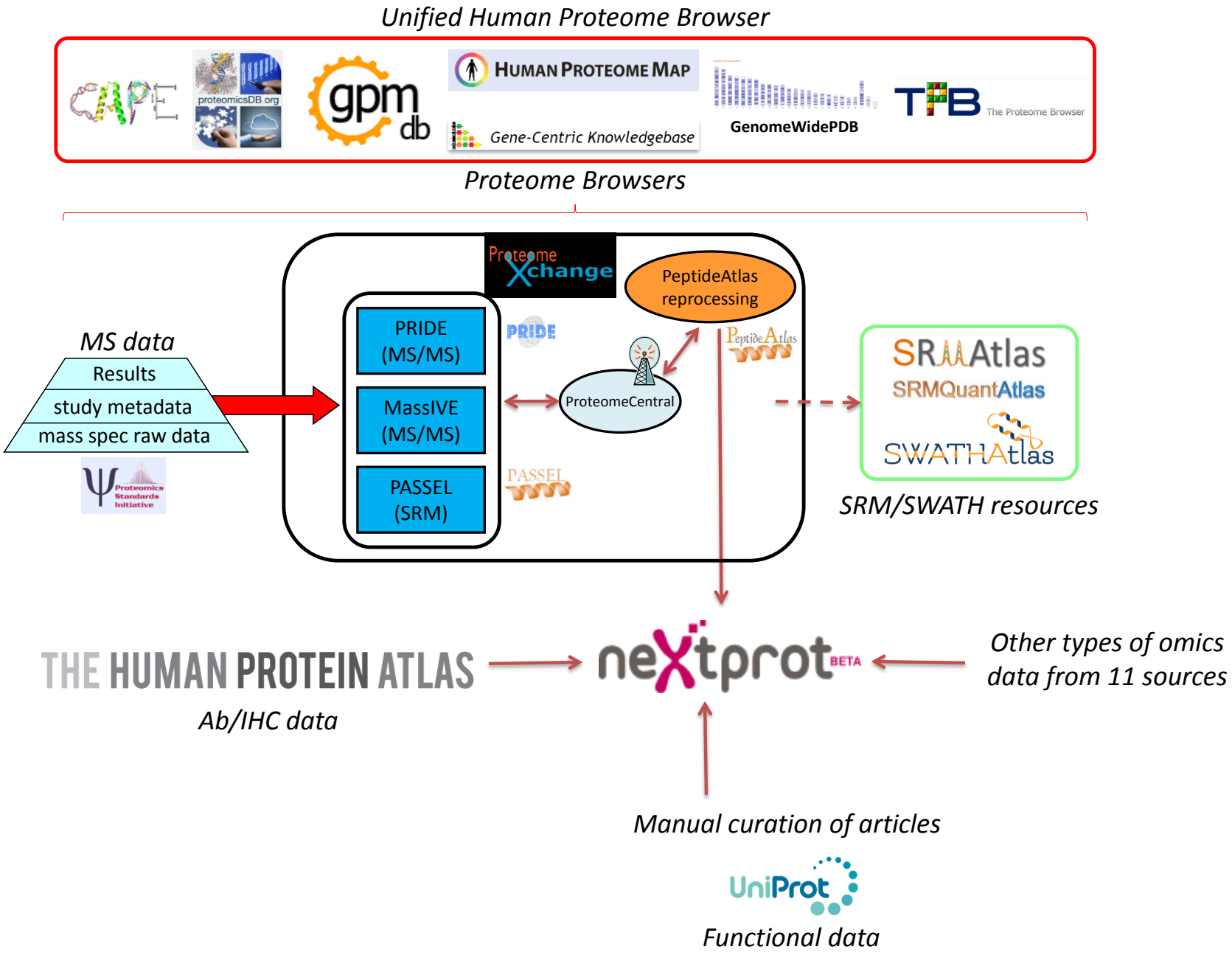
Unified Human Proteome Browser
CAPE
proteomicsDB.org
gpm db
HUMAN PROTEOME MAP
Gene-Centric Knowledgebase
GenomeWidePDB
TPB The Proteome Browser
Proteome Browsers
MS data
Results
study metadata
mass spec raw data
Proteomics Standards Initiative
ProteomeXchange
PRIDE (MS/MS)
MassIVE (MS/MS)
PASSEL (SRM)
PRIDE
PASSEL
ProteomeCentral
PeptideAtlas reprocessing
PeptideAtlas
SRMAtlas
SRMQuantAtlas
SWATHAtlas
SRM/SWATH resources
THE HUMAN PROTEIN ATLAS
Ab/IHC data
neXtprot BETA
Other types of omics data from 11 sources
Manual curation of articles
UniProt
Functional data

Figure 1. Bioinformatics resources to support the discovery, cataloging and browsing of protein part lists. Raw MS data acquired in different laboratories are deposited in the storage resources from the ProteomeXchange consortium. At present MS/MS datasets are fully supported by PRIDE and MassIVE, whereas SRM datasets are supported by PASSEL. Once is made publicly available, data in ProteomeXchange can be further used by many resources, for example reprocessing by PeptideAtlas using the Trans Proteomic Pipeline, or e.g. used by other Protein Browser resources. SRMAtlas, SRMQuanAtlas, SWATHAtlas with PeptideAtlas and other spectral libraries form the rich resources to develop or implement SRM assays. NeXtProt integrates data on proteins using 14 different resources and classifies proteins in 5 existence categories. C-HPP at HUPO 2014 in Madrid launched the Unified Human Proteome Browser Initiative to provide a unified view of the acquired proteogenomic data from a chromosome, biology and disease perspective.

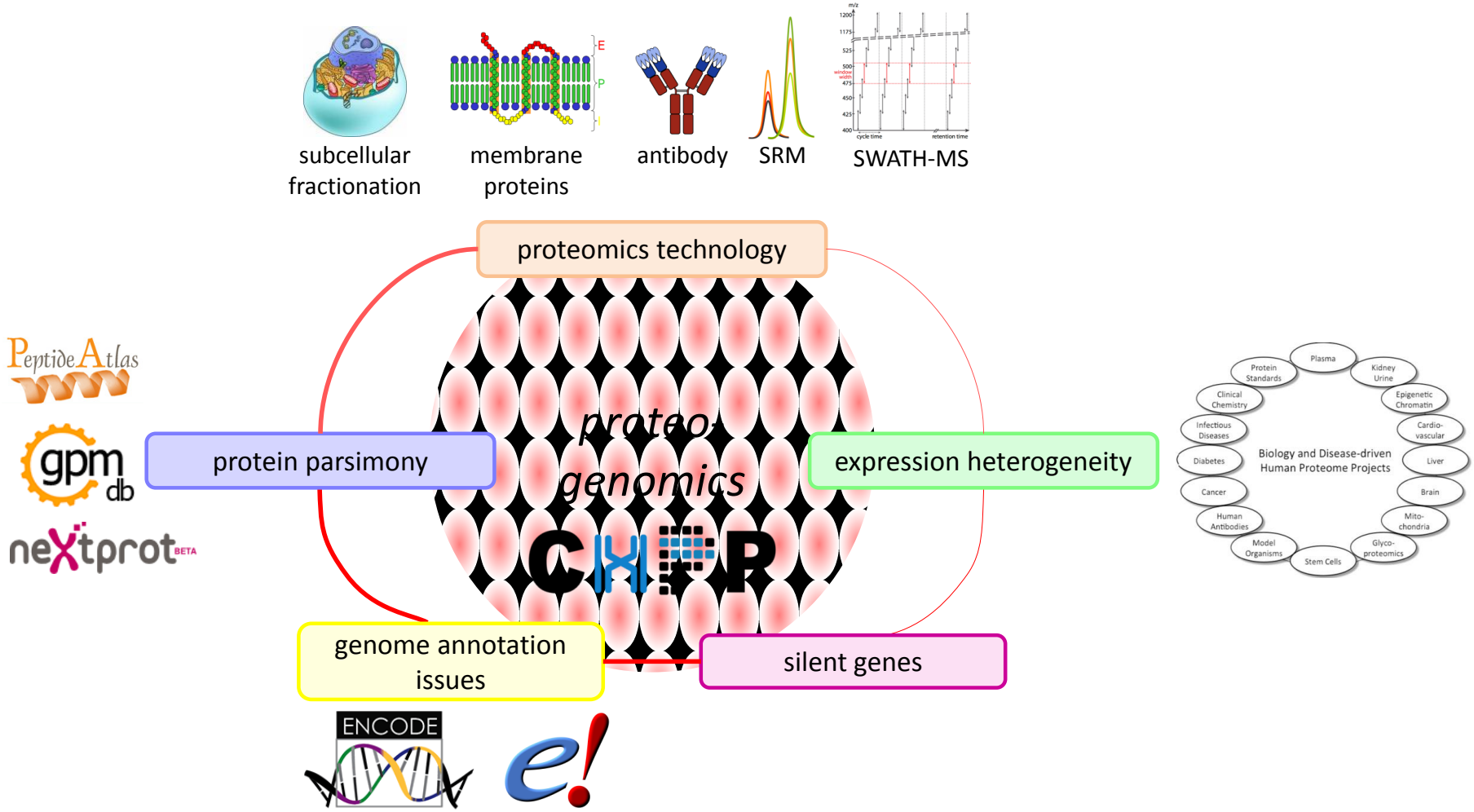

Figure 2. The five main reasons for proteins without evidence at the protein level (missing proteins) are: (1) current proteomics technology is not able to detect them due to uncovered chemical space of the applied mainstream analytical method, (2) expression heterogeneity of protein present only in rare and not yet analyzed samples, (3) silent genes present only in the genome, but never expressed, and (4) error in genome annotation, or (5) proteins missed or not counted due to parsimonious protein identification of shotgun proteomics database search

approach leading to simplification of protein representation of highly homologous proteins of the same protein family in databases such as PeptideAtlas, neXtProt, and gpmDB[40] or to large sequence variability such as immunoglobulins. Proteomics technology (1) can be improved with techniques such as subcellular fractionation or specific enrichment of membrane proteins, while low abundant proteins can be detected either with enrichment using monoclonal and polyclonal antibodies or/and using sensitive SRM and SWATH analysis. Expression heterogeneity (4) can be improved by joining forces with biology/disease driven research groups for example by enhancing collaboration between C-HPP and B/D HPP teams. Proteogenomic approach integrating genome, transcriptome with proteome data helps in general to (5) identify protein forms originating from genetic variability and (3) may correct for genome annotation errors. Use of multiple protease enhance protein coverage and can lead to distinction of highly homologous proteins (5) in main protein evidence databases. The most challenging group of missing proteins are silent genes (2) which are not normally expressed during the life cycle of an individual but can be activated by mutation, recombination, insertion elements, or other genetic mechanisms.

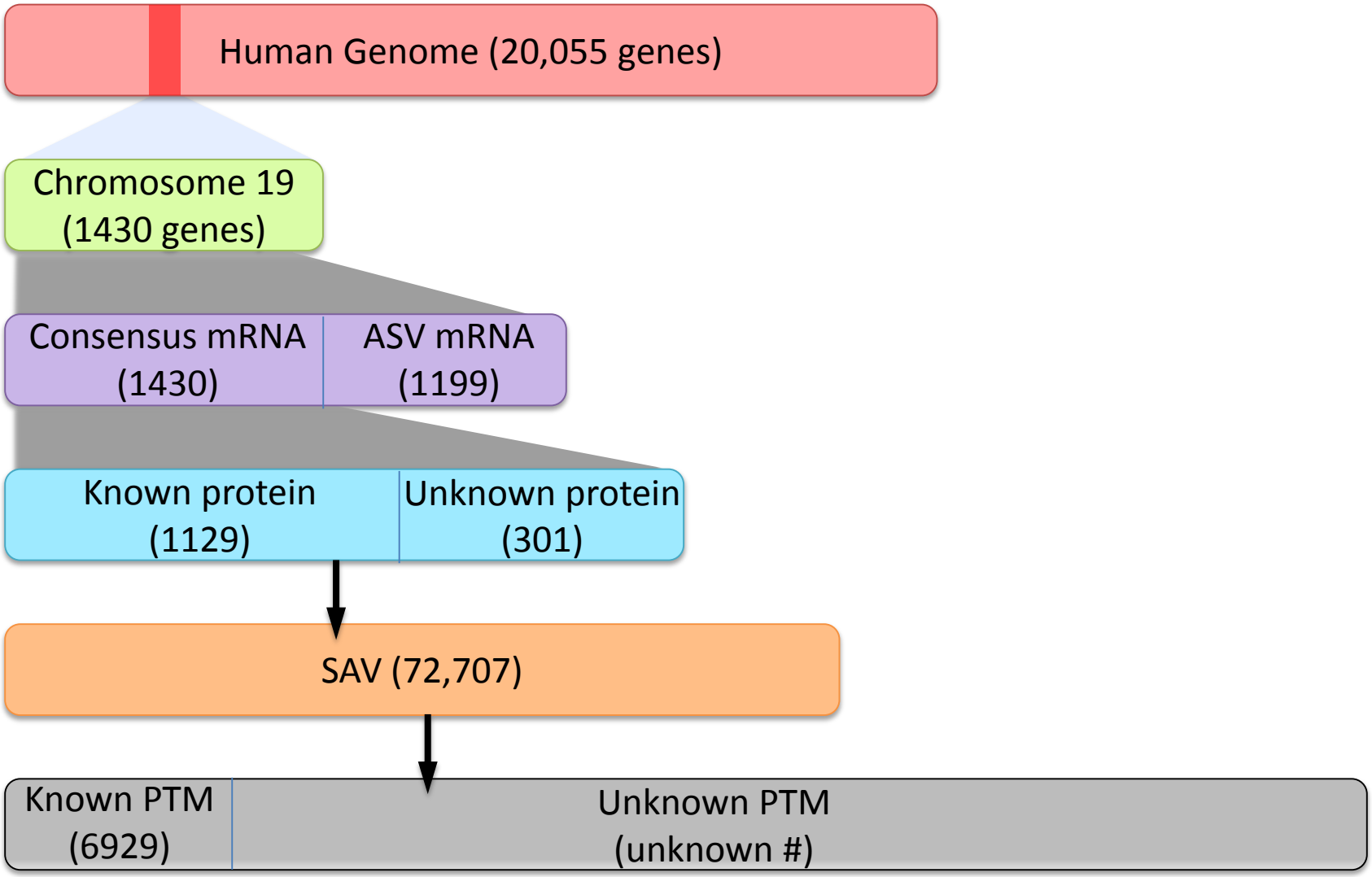

Figure 3. The number of chromosome 19 genes[152] and the identified molecular entities at transcript and expression levels (mRNA and proteins) are illustrated as a proteogenomic analysis of *Glioma* stem cells addressed the challenges integrating genomics, transcriptomics and

proteomics data. Although, the figure presents the current status of chromosome 19, the number of “missing” consensus proteins and their alternative forms, including ASV, new ORFs and new SAVs, is proportionally similar of other human chromosomes.

Poster number	Title	Chr team	Main topic
1	Mining Missing Membrane Proteins from Lung Cancer Tissues and Cell lines	4	Proteomics technology: opening chemical space
2	CAPER 3.0: a scalable cloud-based pipeline for the data-intensive analysis of proteomic datasets	1, 8, 20	Bioinformatics: proteome browser
3	Chromosome 18: Master Proteome	18	Proteomics technology: lowering detection limit
4	GenomewidePDB v 2.0: Update on the transcriptomic and proteomic expression data with alternatively spliced products layered in a genome-wide manner	13	Bioinformatics: proteome browser
5	Development of improved subcellular fractionation procedures of the placental membrane proteins for discovering disease biomarkers and missing proteins	13	Proteomics technology: lower detection limit
6	Chromosome X	X	Missing protein: identification strategy
7	The Mitochondrial Human Proteome Project - MT-HPP	Mitochondria	Missing protein: identification strategy
8	Revisiting the Identification of Canonical Splice lsoforms through Integration of Functional Genomics and Proteomics Evidence from the Chromosome 17 Human Proteome Project	17	Exploring effect of genetic variability
9	Subcellular fractionation enhances Chromosome 16 Proteome Coverage	16	Proteomics technology: lowering detection limit
10	Proteogenomics Dashboard for the Human Proteome Project	16	Bioinformatics: proteogenomic data mining
11	Missing proteins in Chromosome 16 Spanish HPP	16	Missing protein: identification strategy
12	Transcriptomic profiling towards the localization of the missing proteins	16	Missing proteins: identification with proteogenomic
13	The Chromosome 19 Strategy to Characterize Novel Proteoforms and Missing Proteins Using ENCODE Resources	19	Missing protein: identification with proteogenomic
14	Targeting proteins of chromosome 16	16	Proteomics technology: lower detection limit
15	Dissecting Chromosome 16 proteome	16	Missing protein: identification strategy
16	Unified Human Proteome Browser Initiative	all	Bioinformatics: proteome

			browser
17	Expression of αvβ6 integrin enhances both plasminogen and latent-transforming growth factor-β1 dependent proliferation, invasion and ERK1/2 signaling in colorectal cancer cells	7	Proteomics technology: application
18	Overexpression of αvβ6 integrin alters the colorectal cancer cell proteome in favor of elevated proliferation and a switching in cellular adhesion which increases invasion	7	Proteomics technology: application
19	Approaching the organellar brain proteome to understand the molecular basis of Schizophrenia	15	Protein quantification: application. Missing proteins.
20	Splice Variants in Aggressive Human Triple Negative Breast Cancer	17	Exploring effect of genetic variability
21	Revisiting the Identification of Canonical Splice Isoforms through Integration of Functional Genomics and Proteomics Evidence from the Chromosome 17 Human Proteome Project	17	Exploring effect of genetic variability

Table 1. List of posters presented at C-HPP poster session on 7 October 2014 at HUPO 2014 (Madrid), and used as second reference (poster number) in this paper.

References

1. Lichti, C. F.; Mostovenko, E.; Wadsworth, P.; Pettitt, B. M.; Sulman, E. P.; Wang, Q.; Lang, F. F.; Rezeli, M.; Marko-Varga, G.; Végvári, Á.; Nilsson, C. L., Systematic Identification of Single Amino Acid Polymorphisms in Glioma Stem Cell-Derived Chromosome 19 Proteins. *Journal of Proteome Reseach* **2015,** 14, (1).
2. Lichti, C. F.; Mostovenko, E.; Wadsworth, P. A.; Lynch, G. C.; Pettitt, B. M.; Sulman, E. P.; Wang, Q.; Lang, F. F.; Rezeli, M.; Marko-Varga, G.; Vegvari, A.; Nilsson, C. L., Systematic Identification of Single Amino Acid Variants in Glioma Stem-Cell-Derived Chromosome 19 Proteins. *J Proteome Res* **2014**.
3. Nilsson, C. L.; Mostovenko, E.; Lichti, C. F.; Ruggles, K.; Fenyo, D.; Rosenbloom, K. R.; Hancock, W. S.; Paik, Y. K.; Omenn, G. S.; LaBaer, J.; Kroes, R. A.; Uhlen, M.; Hober, S.; Vegvari, A.; Andren, P. E.; Sulman, E. P.; Lang, F. F.; Fuentes, M.; Carlsohn, E.; Emmett, M. R.; Moskal, J. R.; Berven, F. S.; Fehniger, T. E.; Marko-Varga, G., Use of ENCODE Resources to Characterize Novel Proteoforms and Missing Proteins in the Human Proteome. *J Proteome Res* **2014**.
4. Paik, Y. K.; Jeong, S. K.; Omenn, G. S.; Uhlen, M.; Hanash, S.; Cho, S. Y.; Lee, H. J.; Na, K.; Choi, E. Y.; Yan, F.; Zhang, F.; Zhang, Y.; Snyder, M.; Cheng, Y.; Chen, R.; Marko-Varga, G.; Deutsch, E. W.; Kim, H.; Kwon, J. Y.; Aebersold, R.; Bairoch, A.; Taylor, A. D.; Kim, K. Y.; Lee, E. Y.; Hochstrasser, D.; Legrain, P.; Hancock, W. S., 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.; Aebersold, R.; Bairoch, A.; Yamamoto, T.; Legrain, P.; Lee, H. J.; Na, K.; Jeong, S. K.; He, F.; Binz, P. A.; Nishimura, T.; Keown, P.; Baker, M. S.; Yoo, J. S.; Garin, J.; Archakov, A.; Bergeron, J.; Salekdeh, G. H.; Hancock, W. S., Standard guidelines for the chromosome-centric human proteome project. *J Proteome Res* **2012,** 11, (4), 2005-13.
6. Legrain, P.; Aebersold, R.; Archakov, A.; Bairoch, A.; Bala, K.; Beretta, L.; Bergeron, J.; Borchers, C. H.; Corthals, G. L.; Costello, C. E.; Deutsch, E. W.; Domon, B.; Hancock, W.; He, F.; Hochstrasser, D.; Marko-Varga, G.; Salekdeh, G. H.; Sechi, S.; Snyder, M.; Srivastava, S.; Uhlen, M.; Wu, C. H.; Yamamoto, T.; Paik, Y. K.; Omenn, G. S., The human proteome project: current state and future direction. *Mol Cell Proteomics* **2011,** 10, (7), M111 009993.
7. Vizcaino, J. A.; Deutsch, E. W.; Wang, R.; Csordas, A.; Reisinger, F.; Rios, D.; Dianes, J. A.; Sun, Z.; Farrah, T.; Bandeira, N.; Binz, P. A.; Xenarios, I.; Eisenacher, M.; Mayer, G.; Gatto, L.; Campos, A.; Chalkley, R. J.; Kraus, H. J.; Albar, J. P.; Martinez-Bartolome, S.; Apweiler, R.; Omenn, G. S.; Martens, L.; Jones, A. R.; Hermjakob, H., ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat Biotechnol* **2014,** 32, (3), 223-6.
8. Vizcaino, J. A.; Cote, R. G.; Csordas, A.; Dianes, J. A.; Fabregat, A.; Foster, J. M.; Griss, J.; Alpi, E.; Birim, M.; Contell, J.; O'Kelly, G.; Schoenegger, A.; Ovelleiro, D.; Perez-Riverol, Y.; Reisinger, F.; Rios, D.; Wang, R.; Hermjakob, H., The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res* **2013,** 41, (Database issue), D1063-9.
9. Martens, L.; Hermjakob, H.; Jones, P.; Adamski, M.; Taylor, C.; States, D.; Gevaert, K.; Vandekerckhove, J.; Apweiler, R., PRIDE: the proteomics identifications database. *Proteomics* **2005,** 5, (13), 3537-45.
10. Farrah, T.; Deutsch, E. W.; Omenn, G. S.; Sun, Z.; Watts, J. D.; Yamamoto, T.; Shteynberg, D.; Harris, M. M.; Moritz, R. L., State of the human proteome in 2013 as viewed through PeptideAtlas: comparing the kidney, urine, and plasma proteomes for the biology- and disease-driven Human Proteome Project. *J Proteome Res* **2014,** 13, (1), 60-75.

11. Farrah, T.; Deutsch, E. W.; Hoopmann, M. R.; Hallows, J. L.; Sun, Z.; Huang, C. Y.; Moritz, R. L., The state of the human proteome in 2012 as viewed through PeptideAtlas. *J Proteome Res* **2013,** 12, (1), 162-71.
12. Deutsch, E. W., The PeptideAtlas Project. *Methods Mol Biol* **2010,** 604, 285-96.
13. Deutsch, E. W.; Mendoza, L.; Shteynberg, D.; Farrah, T.; Lam, H.; Tasman, N.; Sun, Z.; Nilsson, E.; Pratt, B.; Prazen, B.; Eng, J. K.; Martin, D. B.; Nesvizhskii, A. I.; Aebersold, R., A guided tour of the Trans-Proteomic Pipeline. *Proteomics* **2010,** 10, (6), 1150-9.
14. Keller, A.; Shteynberg, D., Software pipeline and data analysis for MS/MS proteomics: the trans-proteomic pipeline. *Methods Mol Biol* **2011,** 694, 169-89.
15. Pedrioli, P. G., Trans-proteomic pipeline: a pipeline for proteomic analysis. *Methods Mol Biol* **2010,** 604, 213-38.
16. Nesvizhskii, A. I.; Vitek, O.; Aebersold, R., Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat Methods* **2007,** 4, (10), 787-97.
17. Reiter, L.; Claassen, M.; Schrimpf, S. P.; Jovanovic, M.; Schmidt, A.; Buhmann, J. M.; Hengartner, M. O.; Aebersold, R., Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Mol Cell Proteomics* **2009,** 8, (11), 2405-17.
18. Zhang, C. C.; Rogalski, J. C.; Evans, D. M.; Klockenbusch, C.; Beavis, R. C.; Kast, J., In silico protein interaction analysis using the global proteome machine database. *J Proteome Res* **2011,** 10, (2), 656-68.
19. Beavis, R. C., Using the global proteome machine for protein identification. *Methods Mol Biol* **2006,** 328, 217-28.
20. Bjornson, R. D.; Carriero, N. J.; Colangelo, C.; Shifman, M.; Cheung, K. H.; Miller, P. L.; Williams, K., X!!Tandem, an improved method for running X!tandem in parallel on collections of commodity computers. *J Proteome Res* **2008,** 7, (1), 293-9.
21. Duncan, D. T.; Craig, R.; Link, A. J., Parallel tandem: a program for parallel processing of tandem mass spectra using PVM or MPI and X!Tandem. *J Proteome Res* **2005,** 4, (5), 1842-7.
22. Lane, L.; Argoud-Puy, G.; Britan, A.; Cusin, I.; Duek, P. D.; Evalet, O.; Gateau, A.; Gaudet, P.; Gleizes, A.; Masselot, A.; Zwahlen, C.; Bairoch, A., neXtProt: a knowledge platform for human proteins. *Nucleic Acids Res* **2012,** 40, (Database issue), D76-83.
23. Gaudet, P.; Argoud-Puy, G.; Cusin, I.; Duek, P.; Evalet, O.; Gateau, A.; Gleizes, A.; Pereira, M.; Zahn-Zabal, M.; Zwahlen, C.; Bairoch, A.; Lane, L., neXtProt: organizing protein knowledge in the context of human proteome projects. *J Proteome Res* **2013,** 12, (1), 293-8.
24. Magrane, M.; Consortium, U., UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)* **2011,** 2011, bar009.
25. Jain, E.; Bairoch, A.; Duvaud, S.; Phan, I.; Redaschi, N.; Suzek, B. E.; Martin, M. J.; McGarvey, P.; Gasteiger, E., Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinformatics* **2009,** 10, 136.
26. Apweiler, R.; Bairoch, A.; Wu, C. H., Protein sequence databases. *Curr Opin Chem Biol* **2004,** 8, (1), 76-80.
27. Sayers, E. W.; Barrett, T.; Benson, D. A.; Bolton, E.; Bryant, S. H.; Canese, K.; Chetvernin, V.; Church, D. M.; Dicuccio, M.; Federhen, S.; Feolo, M.; Fingerman, I. M.; Geer, L. Y.; Helmberg, W.; Kapustin, Y.; Krasnov, S.; Landsman, D.; Lipman, D. J.; Lu, Z.; Madden, T. L.; Madej, T.; Maglott, D. R.; Marchler-Bauer, A.; Miller, V.; Karsch-Mizrachi, I.; Ostell, J.; Panchenko, A.; Phan, L.; Pruitt, K. D.; Schuler, G. D.; Sequeira, E.; Sherry, S. T.; Shumway, M.; Sirotkin, K.; Slotta, D.; Souvorov, A.; Starchenko, G.; Tatusova, T. A.; Wagner, L.; Wang, Y.; Wilbur, W. J.; Yaschenko, E.; Ye, J., Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **2012,** 40, (Database issue), D13-25.
28. Forbes, S. A.; Bindal, N.; Bamford, S.; Cole, C.; Kok, C. Y.; Beare, D.; Jia, M.; Shepherd, R.; Leung, K.; Menzies, A.; Teague, J. W.; Campbell, P. J.; Stratton, M. R.; Futreal, P. A., COSMIC: mining complete

cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res* **2011,** 39, (Database issue), D945-50.
29. Bastian, F.; Parmentier, G.; Roux, J.; Moretti, S.; Laudet, V.; Robinson-Rechavi, M., Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species. In *Data Integration in the Life Sciences*, Bairoch, A.; Cohen-Boulakia, S.; Froidevaux, C., Eds. Springer Berlin Heidelberg: 2008; Vol. 5109, pp 124-131.
30. Uhlen, M.; Oksvold, P.; Fagerberg, L.; Lundberg, E.; Jonasson, K.; Forsberg, M.; Zwahlen, M.; Kampf, C.; Wester, K.; Hober, S.; Wernerus, H.; Bjorling, L.; Ponten, F., Towards a knowledge-based Human Protein Atlas. *Nat Biotechnol* **2010,** 28, (12), 1248-50.
31. Uhlen, M.; Bjorling, E.; Agaton, C.; Szigyarto, C. A.; Amini, B.; Andersen, E.; Andersson, A. C.; Angelidou, P.; Asplund, A.; Asplund, C.; Berglund, L.; Bergstrom, K.; Brumer, H.; Cerjan, D.; Ekstrom, M.; Elobeid, A.; Eriksson, C.; Fagerberg, L.; Falk, R.; Fall, J.; Forsberg, M.; Bjorklund, M. G.; Gumbel, K.; Halimi, A.; Hallin, I.; Hamsten, C.; Hansson, M.; Hedhammar, M.; Hercules, G.; Kampf, C.; Larsson, K.; Lindskog, M.; Lodewyckx, W.; Lund, J.; Lundeberg, J.; Magnusson, K.; Malm, E.; Nilsson, P.; Odling, J.; Oksvold, P.; Olsson, I.; Oster, E.; Ottosson, J.; Paavilainen, L.; Persson, A.; Rimini, R.; Rockberg, J.; Runeson, M.; Sivertsson, A.; Skollermo, A.; Steen, J.; Stenvall, M.; Sterky, F.; Stromberg, S.; Sundberg, M.; Tegel, H.; Tourle, S.; Wahlund, E.; Walden, A.; Wan, J.; Wernerus, H.; Westberg, J.; Wester, K.; Wrethagen, U.; Xu, L. L.; Hober, S.; Ponten, F., A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol Cell Proteomics* **2005,** 4, (12), 1920-32.
32. Bernstein, F. C.; Koetzle, T. F.; Williams, G. J.; Meyer, E. F., Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M., The Protein Data Bank: a computer-based archival file for macromolecular structures. *J Mol Biol* **1977,** 112, (3), 535-42.
33. Gray, K. A.; Daugherty, L. C.; Gordon, S. M.; Seal, R. L.; Wright, M. W.; Bruford, E. A., Genenames.org: the HGNC resources in 2013. *Nucleic Acids Res* **2013,** 41, (Database issue), D545-52.
34. Tatusova, T.; Ciufo, S.; Fedorov, B.; O'Neill, K.; Tolstoy, I., RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res* **2014,** 42, (Database issue), D553-9.
35. Pruitt, K. D.; Brown, G. R.; Hiatt, S. M.; Thibaud-Nissen, F.; Astashyn, A.; Ermolaeva, O.; Farrell, C. M.; Hart, J.; Landrum, M. J.; McGarvey, K. M.; Murphy, M. R.; O'Leary, N. A.; Pujar, S.; Rajput, B.; Rangwala, S. H.; Riddick, L. D.; Shkeda, A.; Sun, H.; Tamez, P.; Tully, R. E.; Wallin, C.; Webb, D.; Weber, J.; Wu, W.; DiCuccio, M.; Kitts, P.; Maglott, D. R.; Murphy, T. D.; Ostell, J. M., RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res* **2014,** 42, (Database issue), D756-63.
36. Farrell, C. M.; O'Leary, N. A.; Harte, R. A.; Loveland, J. E.; Wilming, L. G.; Wallin, C.; Diekhans, M.; Barrell, D.; Searle, S. M.; Aken, B.; Hiatt, S. M.; Frankish, A.; Suner, M. M.; Rajput, B.; Steward, C. A.; Brown, G. R.; Bennett, R.; Murphy, M.; Wu, W.; Kay, M. P.; Hart, J.; Rajan, J.; Weber, J.; Snow, C.; Riddick, L. D.; Hunt, T.; Webb, D.; Thomas, M.; Tamez, P.; Rangwala, S. H.; McGarvey, K. M.; Pujar, S.; Shkeda, A.; Mudge, J. M.; Gonzalez, J. M.; Gilbert, J. G.; Trevanion, S. J.; Baertsch, R.; Harrow, J. L.; Hubbard, T.; Ostell, J. M.; Haussler, D.; Pruitt, K. D., Current status and new features of the Consensus Coding Sequence database. *Nucleic Acids Res* **2014,** 42, (Database issue), D865-72.
37. Harte, R. A.; Farrell, C. M.; Loveland, J. E.; Suner, M. M.; Wilming, L.; Aken, B.; Barrell, D.; Frankish, A.; Wallin, C.; Searle, S.; Diekhans, M.; Harrow, J.; Pruitt, K. D., Tracking and coordinating an international curation effort for the CCDS Project. *Database (Oxford)* **2012,** 2012, bas008.
38. Pruitt, K. D.; Harrow, J.; Harte, R. A.; Wallin, C.; Diekhans, M.; Maglott, D. R.; Searle, S.; Farrell, C. M.; Loveland, J. E.; Ruef, B. J.; Hart, E.; Suner, M. M.; Landrum, M. J.; Aken, B.; Ayling, S.; Baertsch, R.; Fernandez-Banet, J.; Cherry, J. L.; Curwen, V.; Dicuccio, M.; Kellis, M.; Lee, J.; Lin, M. F.; Schuster, M.; Shkeda, A.; Amid, C.; Brown, G.; Dukhanina, O.; Frankish, A.; Hart, J.; Maidak, B. L.; Mudge, J.; Murphy, M. R.; Murphy, T.; Rajan, J.; Rajput, B.; Riddick, L. D.; Snow, C.; Steward, C.; Webb, D.; Weber, J. A.; Wilming, L.; Wu, W.; Birney, E.; Haussler, D.; Hubbard, T.; Ostell, J.; Durbin, R.; Lipman, D., The

consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res* **2009,** 19, (7), 1316-23.
39. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res* **2014,** 42, (Database issue), D191-8.
40. Omenn, G. S., The strategy, organization, and progress of the HUPO Human Proteome Project. *J Proteomics* **2014,** 100, 3-7.
41. Farrah, T.; Deutsch, E. W.; Omenn, G. S.; Campbell, D. S.; Sun, Z.; Bletz, J. A.; Mallick, P.; Katz, J. E.; Malmstrom, J.; Ossola, R.; Watts, J. D.; Lin, B.; Zhang, H.; Moritz, R. L.; Aebersold, R., A high-confidence human plasma proteome reference set with estimated concentrations in PeptideAtlas. *Mol Cell Proteomics* **2011,** 10, (9), M110 006353.
42. 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.
43. Horvatovich, P.; Franke, L.; Bischoff, R., Proteomic studies related to genetic determinants of variability in protein concentrations. *J Proteome Res* **2014,** 13, (1), 5-14.
44. Bairoch, A.; Apweiler, R., The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* **2000,** 28, (1), 45-8.
45. Tang, W.; Fei, Y.; Page, M., Biological significance of RNA editing in cells. *Mol Biotechnol* **2012,** 52, (1), 91-100.
46. Chepelev, I., Detection of RNA editing events in human cells using high-throughput sequencing. *Methods Mol Biol* **2012,** 815, 91-102.
47. Nesvizhskii, A. I., Proteogenomics: concepts, applications and computational strategies. *Nat Methods* **2014,** 11, (11), 1114-25.
48. Jaffe, J. D.; Berg, H. C.; Church, G. M., Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics* **2004,** 4, (1), 59-77.
49. Sheynkman, G. M.; Shortreed, M. R.; Frey, B. L.; Scalf, M.; Smith, L. M., Large-scale mass spectrometric detection of variant peptides resulting from nonsynonymous nucleotide differences. *J Proteome Res* **2014,** 13, (1), 228-40.
50. Imanishi, T.; Nagai, Y.; Habara, T.; Yamasaki, C.; Takeda, J.; Mikami, S.; Bando, Y.; Tojo, H.; Nishimura, T., Full-length transcriptome-based H-InvDB throws a new light on chromosome-centric proteomics. *J Proteome Res* **2013,** 12, (1), 62-6.
51. Wang, Q.; Wen, B.; Wang, T.; Xu, Z.; Yin, X.; Xu, S.; Ren, Z.; Hou, G.; Zhou, R.; Zhao, H.; Zi, J.; Zhang, S.; Gao, H.; Lou, X.; Sun, H.; Feng, Q.; Chang, C.; Qin, P.; Zhang, C.; Li, N.; Zhu, Y.; Gu, W.; Zhong, J.; Zhang, G.; Yang, P.; Yan, G.; Shen, H.; Liu, X.; Lu, H.; Zhong, F.; He, Q. Y.; Xu, P.; Lin, L.; Liu, S., Omics evidence: single nucleotide variants transmissions on chromosome 20 in liver cancer cell lines. *J Proteome Res* **2014,** 13, (1), 200-11.
52. Chang, C.; Li, L.; Zhang, C.; Wu, S.; Guo, K.; Zi, J.; Chen, Z.; Jiang, J.; Ma, J.; Yu, Q.; Fan, F.; Qin, P.; Han, M.; Su, N.; Chen, T.; Wang, K.; Zhai, L.; Zhang, T.; Ying, W.; Xu, Z.; Zhang, Y.; Liu, Y.; Liu, X.; Zhong, F.; Shen, H.; Wang, Q.; Hou, G.; Zhao, H.; Li, G.; Liu, S.; Gu, W.; Wang, G.; Wang, T.; Zhang, G.; Qian, X.; Li, N.; He, Q. Y.; Lin, L.; Yang, P.; Zhu, Y.; He, F.; Xu, P., Systematic analyses of the transcriptome, translatome, and proteome provide a global view and potential strategy for the C-HPP. *J Proteome Res* **2014,** 13, (1), 38-49.
53. Zhong, J.; Cui, Y.; Guo, J.; Chen, Z.; Yang, L.; He, Q. Y.; Zhang, G.; Wang, T., Resolving chromosome-centric human proteome with translating mRNA analysis: a strategic demonstration. *J Proteome Res* **2014,** 13, (1), 50-9.
54. Jeong, K.; Kim, S.; Bandeira, N., False discovery rates in spectral identification. *BMC Bioinformatics* **2012,** 13 Suppl 16, S2.

55. Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R., Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* **2002,** 74, (20), 5383-92.
56. Kall, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J., Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods* **2007,** 4, (11), 923-5.
57. Kall, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S., Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J Proteome Res* **2008,** 7, (1), 29-34.
58. Kall, L.; Storey, J. D.; Noble, W. S., Non-parametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry. *Bioinformatics* **2008,** 24, (16), i42-8.
59. Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R., A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* **2003,** 75, (17), 4646-58.
60. Vaudel, M.; Burkhart, J. M.; Zahedi, R. P.; Oveland, E.; Berven, F. S.; Sickmann, A.; Martens, L.; Barsnes, H., PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat Biotechnol* **2015,** 33, (1), 22-4.
61. Kroksveen, A. C.; Guldbrandsen, A.; Vedeler, C.; Myhr, K. M.; Opsahl, J. A.; Berven, F. S., Cerebrospinal fluid proteome comparison between multiple sclerosis patients and controls. *Acta Neurol Scand Suppl* **2012**, (195), 90-6.
62. Baczek, T.; Kaliszan, R., Predictions of peptides' retention times in reversed-phase liquid chromatography as a new supportive tool to improve protein identification in proteomics. *Proteomics* **2009,** 9, (4), 835-47.
63. Branca, R. M.; Orre, L. M.; Johansson, H. J.; Granholm, V.; Huss, M.; Perez-Bercoff, A.; Forshed, J.; Kall, L.; Lehtio, J., HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. *Nat Methods* **2014,** 11, (1), 59-62.
64. Smith, L. M.; Kelleher, N. L., Proteoform: a single term describing protein complexity. *Nat Methods* **2013,** 10, (3), 186-7.
65. Kelleher, N. L.; Thomas, P. M.; Ntai, I.; Compton, P. D.; LeDuc, R. D., Deep and quantitative top-down proteomics in clinical and translational research. *Expert Rev Proteomics* **2014**, 1-3.
66. Catherman, A. D.; Skinner, O. S.; Kelleher, N. L., Top Down proteomics: facts and perspectives. *Biochem Biophys Res Commun* **2014,** 445, (4), 683-93.
67. Lisitsa, A.; Moshkovskii, S.; Chernobrovkin, A.; Ponomarenko, E.; Archakov, A., Profiling proteoforms: promising follow-up of proteomics for biomarker discovery. *Expert Rev Proteomics* **2014,** 11, (1), 121-9.
68. Flicek, P.; Amode, M. R.; Barrell, D.; Beal, K.; Billis, K.; Brent, S.; Carvalho-Silva, D.; Clapham, P.; Coates, G.; Fitzgerald, S.; Gil, L.; Giron, C. G.; Gordon, L.; Hourlier, T.; Hunt, S.; Johnson, N.; Juettemann, T.; Kahari, A. K.; Keenan, S.; Kulesha, E.; Martin, F. J.; Maurel, T.; McLaren, W. M.; Murphy, D. N.; Nag, R.; Overduin, B.; Pignatelli, M.; Pritchard, B.; Pritchard, E.; Riat, H. S.; Ruffier, M.; Sheppard, D.; Taylor, K.; Thormann, A.; Trevanion, S. J.; Vullo, A.; Wilder, S. P.; Wilson, M.; Zadissa, A.; Aken, B. L.; Birney, E.; Cunningham, F.; Harrow, J.; Herrero, J.; Hubbard, T. J.; Kinsella, R.; Muffato, M.; Parker, A.; Spudich, G.; Yates, A.; Zerbino, D. R.; Searle, S. M., Ensembl 2014. *Nucleic Acids Res* **2014,** 42, (Database issue), D749-55.
69. An integrated encyclopedia of DNA elements in the human genome. *Nature* **2012,** 489, (7414), 57-74.
70. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **2004,** 306, (5696), 636-40.
71. Harrow, J.; Frankish, A.; Gonzalez, J. M.; Tapanari, E.; Diekhans, M.; Kokocinski, F.; Aken, B. L.; Barrell, D.; Zadissa, A.; Searle, S.; Barnes, I.; Bignell, A.; Boychenko, V.; Hunt, T.; Kay, M.; Mukherjee, G.; Rajan, J.; Despacio-Reyes, G.; Saunders, G.; Steward, C.; Harte, R.; Lin, M.; Howald, C.; Tanzer, A.; Derrien, T.; Chrast, J.; Walters, N.; Balasubramanian, S.; Pei, B.; Tress, M.; Rodriguez, J. M.; Ezkurdia, I.; van Baren, J.; Brent, M.; Haussler, D.; Kellis, M.; Valencia, A.; Reymond, A.; Gerstein, M.; Guigo, R.;

Hubbard, T. J., GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **2012,** 22, (9), 1760-74.
72. Kim, M. S.; Pinto, S. M.; Getnet, D.; Nirujogi, R. S.; Manda, S. S.; Chaerkady, R.; Madugundu, A. K.; Kelkar, D. S.; Isserlin, R.; Jain, S.; Thomas, J. K.; Muthusamy, B.; Leal-Rojas, P.; Kumar, P.; Sahasrabuddhe, N. A.; Balakrishnan, L.; Advani, J.; George, B.; Renuse, S.; Selvan, L. D.; Patil, A. H.; Nanjappa, V.; Radhakrishnan, A.; Prasad, S.; Subbannayya, T.; Raju, R.; Kumar, M.; Sreenivasamurthy, S. K.; Marimuthu, A.; Sathe, G. J.; Chavan, S.; Datta, K. K.; Subbannayya, Y.; Sahu, A.; Yelamanchi, S. D.; Jayaram, S.; Rajagopalan, P.; Sharma, J.; Murthy, K. R.; Syed, N.; Goel, R.; Khan, A. A.; Ahmad, S.; Dey, G.; Mudgal, K.; Chatterjee, A.; Huang, T. C.; Zhong, J.; Wu, X.; Shaw, P. G.; Freed, D.; Zahari, M. S.; Mukherjee, K. K.; Shankar, S.; Mahadevan, A.; Lam, H.; Mitchell, C. J.; Shankar, S. K.; Satishchandra, P.; Schroeder, J. T.; Sirdeshmukh, R.; Maitra, A.; Leach, S. D.; Drake, C. G.; Halushka, M. K.; Prasad, T. S.; Hruban, R. H.; Kerr, C. L.; Bader, G. D.; Iacobuzio-Donahue, C. A.; Gowda, H.; Pandey, A., A draft map of the human proteome. *Nature* **2014,** 509, (7502), 575-81.
73. Schwanhausser, B.; Busse, D.; Li, N.; Dittmar, G.; Schuchhardt, J.; Wolf, J.; Chen, W.; Selbach, M., Corrigendum: Global quantification of mammalian gene expression control. *Nature* **2013,** 495, (7439), 126-7.
74. Schwanhausser, B.; Busse, D.; Li, N.; Dittmar, G.; Schuchhardt, J.; Wolf, J.; Chen, W.; Selbach, M., Global quantification of mammalian gene expression control. *Nature* **2011,** 473, (7347), 337-42.
75. Wang, T.; Cui, Y.; Jin, J.; Guo, J.; Wang, G.; Yin, X.; He, Q. Y.; Zhang, G., Translating mRNAs strongly correlate to proteins in a multivariate manner and their translation ratios are phenotype specific. *Nucleic Acids Res* **2013,** 41, (9), 4743-54.
76. Menon, R.; Im, H.; Zhang, E. Y.; Wu, S. L.; Chen, R.; Snyder, M.; Hancock, W. S.; Omenn, G. S., Distinct splice variants and pathway enrichment in the cell-line models of aggressive human breast cancer subtypes. *J Proteome Res* **2014,** 13, (1), 212-27.
77. Eksi, R.; Li, H. D.; Menon, R.; Wen, Y.; Omenn, G. S.; Kretzler, M.; Guan, Y., Systematically differentiating functions for alternatively spliced isoforms through integrating RNA-seq data. *PLoS Comput Biol* **2013,** 9, (11), e1003314.
78. Nilsson, C. L.; Berven, F.; Selheim, F.; Liu, H.; Moskal, J. R.; Kroes, R. A.; Sulman, E. P.; Conrad, C. A.; Lang, F. F.; Andren, P. E.; Nilsson, A.; Carlsohn, E.; Lilja, H.; Malm, J.; Fenyo, D.; Subramaniyam, D.; Wang, X.; Gonzales-Gonzales, M.; Dasilva, N.; Diez, P.; Fuentes, M.; Vegvari, A.; Sjodin, K.; Welinder, C.; Laurell, T.; Fehniger, T. E.; Lindberg, H.; Rezeli, M.; Edula, G.; Hober, S.; Marko-Varga, G., Chromosome 19 annotations with disease speciation: a first report from the Global Research Consortium. *J Proteome Res* **2013,** 12, (1), 135-50.
79. Verhaak, R. G.; Hoadley, K. A.; Purdom, E.; Wang, V.; Qi, Y.; Wilkerson, M. D.; Miller, C. R.; Ding, L.; Golub, T.; Mesirov, J. P.; Alexe, G.; Lawrence, M.; O'Kelly, M.; Tamayo, P.; Weir, B. A.; Gabriel, S.; Winckler, W.; Gupta, S.; Jakkula, L.; Feiler, H. S.; Hodgson, J. G.; James, C. D.; Sarkaria, J. N.; Brennan, C.; Kahn, A.; Spellman, P. T.; Wilson, R. K.; Speed, T. P.; Gray, J. W.; Meyerson, M.; Getz, G.; Perou, C. M.; Hayes, D. N., Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **2010,** 17, (1), 98-110.
80. Bischoff, R.; Schluter, H., Amino acids: chemistry, functionality and selected non-enzymatic post-translational modifications. *J Proteomics* **2012,** 75, (8), 2275-96.
81. Zhao, Y.; Jensen, O. N., Modification-specific proteomics: strategies for characterization of post-translational modifications using enrichment techniques. *Proteomics* **2009,** 9, (20), 4632-41.
82. Walsh, C. T.; Garneau-Tsodikova, S.; Gatto, G. J., Jr., Protein posttranslational modifications: the chemistry of proteome diversifications. *Angew Chem Int Ed Engl* **2005,** 44, (45), 7342-72.
83. Uhlen, M.; Fagerberg, L.; Hallstrom, B. M.; Lindskog, C.; Oksvold, P.; Mardinoglu, A.; Sivertsson, A.; Kampf, C.; Sjostedt, E.; Asplund, A.; Olsson, I.; Edlund, K.; Lundberg, E.; Navani, S.; Szigyarto, C. A.; Odeberg, J.; Djureinovic, D.; Takanen, J. O.; Hober, S.; Alm, T.; Edqvist, P. H.; Berling, H.; Tegel, H.;

Mulder, J.; Rockberg, J.; Nilsson, P.; Schwenk, J. M.; Hamsten, M.; von Feilitzen, K.; Forsberg, M.; Persson, L.; Johansson, F.; Zwahlen, M.; von Heijne, G.; Nielsen, J.; Ponten, F., Proteomics. Tissue-based map of the human proteome. *Science* **2015,** 347, (6220), 1260419.
84. Boschetti, E.; Righetti, P. G., The ProteoMiner in the proteomic arena: a non-depleting tool for discovering low-abundance species. *J Proteomics* **2008,** 71, (3), 255-64.
85. Righetti, P. G.; Boschetti, E., The ProteoMiner and the FortyNiners: searching for gold nuggets in the proteomic arena. *Mass Spectrom Rev* **2008,** 27, (6), 596-608.
86. Frese, C. K.; Altelaar, A. F.; van den Toorn, H.; Nolting, D.; Griep-Raming, J.; Heck, A. J.; Mohammed, S., Toward full peptide sequence coverage by dual fragmentation combining electron-transfer and higher-energy collision dissociation tandem mass spectrometry. *Anal Chem* **2012,** 84, (22), 9668-73.
87. Liu, F.; van Breukelen, B.; Heck, A. J., Facilitating Protein Disulfide Mapping by a Combination of Pepsin Digestion, Electron Transfer Higher Energy Dissociation (EThcD), and a Dedicated Search Algorithm SlinkS. *Mol Cell Proteomics* **2014,** 13, (10), 2776-86.
88. Mommen, G. P.; Frese, C. K.; Meiring, H. D.; van Gaans-van den Brink, J.; de Jong, A. P.; van Els, C. A.; Heck, A. J., Expanding the detectable HLA peptide repertoire using electron-transfer/higher-energy collision dissociation (EThcD). *Proc Natl Acad Sci U S A* **2014,** 111, (12), 4507-12.
89. Picotti, P.; Bodenmiller, B.; Mueller, L. N.; Domon, B.; Aebersold, R., Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics. *Cell* **2009,** 138, (4), 795-806.
90. Picotti, P.; Lam, H.; Campbell, D.; Deutsch, E. W.; Mirzaei, H.; Ranish, J.; Domon, B.; Aebersold, R., A database of mass spectrometric assays for the yeast proteome. *Nat Methods* **2008,** 5, (11), 913-4.
91. Kusebauch, U.; Deutsch, E. W.; Campbell, D. S.; Sun, Z.; Farrah, T.; Moritz, R. L., Using PeptideAtlas, SRMAtlas, and PASSEL: Comprehensive Resources for Discovery and Targeted Proteomics. *Curr Protoc Bioinformatics* **2014,** 46, 13 25 1-13 25 28.
92. Farrah, T.; Deutsch, E. W.; Kreisberg, R.; Sun, Z.; Campbell, D. S.; Mendoza, L.; Kusebauch, U.; Brusniak, M. Y.; Huttenhain, R.; Schiess, R.; Selevsek, N.; Aebersold, R.; Moritz, R. L., PASSEL: the PeptideAtlas SRMexperiment library. *Proteomics* **2012,** 12, (8), 1170-5.
93. Perez-Riverol, Y.; Wang, R.; Hermjakob, H.; Muller, M.; Vesada, V.; Vizcaino, J. A., Open source libraries and frameworks for mass spectrometry based proteomics: a developer's perspective. *Biochim Biophys Acta* **2014,** 1844, (1 Pt A), 63-76.
94. Brusniak, M. Y.; Kwok, S. T.; Christiansen, M.; Campbell, D.; Reiter, L.; Picotti, P.; Kusebauch, U.; Ramos, H.; Deutsch, E. W.; Chen, J.; Moritz, R. L.; Aebersold, R., ATAQS: A computational software tool for high throughput transition optimization and validation for selected reaction monitoring mass spectrometry. *BMC Bioinformatics* **2011,** 12, 78.
95. Reiter, L.; Rinner, O.; Picotti, P.; Huttenhain, R.; Beck, M.; Brusniak, M. Y.; Hengartner, M. O.; Aebersold, R., mProphet: automated data processing and statistical validation for large-scale SRM experiments. *Nat Methods* **2011,** 8, (5), 430-5.
96. Sherwood, C. A.; Eastham, A.; Lee, L. W.; Peterson, A.; Eng, J. K.; Shteynberg, D.; Mendoza, L.; Deutsch, E. W.; Risler, J.; Tasman, N.; Aebersold, R.; Lam, H.; Martin, D. B., MaRiMba: a software application for spectral library-based MRM transition list assembly. *J Proteome Res* **2009,** 8, (10), 4396-405.
97. Sheng, Q.; Wu, C.; Su, Z.; Zeng, R., SRMBuilder: a user-friendly tool for selected reaction monitoring data analysis. *J Bioinform Comput Biol* **2011,** 9 Suppl 1, 51-62.
98. MacLean, B.; Tomazela, D. M.; Shulman, N.; Chambers, M.; Finney, G. L.; Frewen, B.; Kern, R.; Tabb, D. L.; Liebler, D. C.; MacCoss, M. J., Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **2010,** 26, (7), 966-8.

99. Mohammed, Y.; Domanski, D.; Jackson, A. M.; Smith, D. S.; Deelder, A. M.; Palmblad, M.; Borchers, C. H., PeptidePicker: a scientific workflow with web interface for selecting appropriate peptides for targeted proteomics experiments. *J Proteomics* **2014,** 106, 151-61.
100. Rosenberger, G.; Koh, C. C.; Guo, T.; Röst, H. L.; Kouvonen, P.; Collins, B. C.; Heusel, M.; Liu, Y.; Caron, E.; Vichalkovski, A.; Faini, M.; Schubert, O. T.; Faridi, P.; Ebhardt, H. A.; Matondo, M.; Lam, H.; Bader, S. L.; Campbell, D. S.; Deutsch, E. W.; Moritz, R. L.; Tate, S.; Aebersold, R., A repository of assays to quantify 10,000 human proteins by SWATH-MS. *Scientific Data* **2014,** 1.
101. Rost, H. L.; Rosenberger, G.; Navarro, P.; Gillet, L.; Miladinovic, S. M.; Schubert, O. T.; Wolski, W.; Collins, B. C.; Malmstrom, J.; Malmstrom, L.; Aebersold, R., OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat Biotechnol* **2014,** 32, (3), 219-23.
102. Marx, H.; Lemeer, S.; Schliep, J. E.; Matheron, L.; Mohammed, S.; Cox, J.; Mann, M.; Heck, A. J.; Kuster, B., A large synthetic peptide and phosphopeptide reference library for mass spectrometry-based proteomics. *Nat Biotechnol* **2013,** 31, (6), 557-64.
103. Wang, K.; Huang, C.; Nice, E., Recent advances in proteomics: towards the human proteome. *Biomed Chromatogr* **2014,** 28, (6), 848-57.
104. Colwill, K.; Graslund, S., A roadmap to generate renewable protein binders to the human proteome. *Nat Methods* **2011,** 8, (7), 551-8.
105. Layton, D.; Laverty, C.; Nice, E., Design and operation of an automated high-throughput monoclonal antibody facility. *Biophysical Reviews* **2013,** 5, (1), 47-55.
106. Razavi, M.; Frick, L. E.; LaMarr, W. A.; Pope, M. E.; Miller, C. A.; Anderson, N. L.; Pearson, T. W., High-throughput SISCAPA quantitation of peptides from human plasma digests by ultrafast, liquid chromatography-free mass spectrometry. *J Proteome Res* **2012,** 11, (12), 5642-9.
107. Zhao, L.; Whiteaker, J. R.; Pope, M. E.; Kuhn, E.; Jackson, A.; Anderson, N. L.; Pearson, T. W.; Carr, S. A.; Paulovich, A. G., Quantification of proteins using peptide immunoaffinity enrichment coupled with mass spectrometry. *J Vis Exp* **2011**, (53).
108. Anderson, N. L.; Anderson, N. G.; Haines, L. R.; Hardie, D. B.; Olafson, R. W.; Pearson, T. W., Mass spectrometric quantitation of peptides and proteins using Stable Isotope Standards and Capture by Anti-Peptide Antibodies (SISCAPA). *J Proteome Res* **2004,** 3, (2), 235-44.
109. Persson, A.; Hober, S.; Uhlen, M., A human protein atlas based on antibody proteomics. *Curr Opin Mol Ther* **2006,** 8, (3), 185-90.
110. Larsson, K.; Wester, K.; Nilsson, P.; Uhlen, M.; Hober, S.; Wernerus, H., Multiplexed PrEST immunization for high-throughput affinity proteomics. *J Immunol Methods* **2006,** 315, (1-2), 110-20.
111. Fagerberg, L.; Hallstrom, B. M.; Oksvold, P.; Kampf, C.; Djureinovic, D.; Odeberg, J.; Habuka, M.; Tahmasebpoor, S.; Danielsson, A.; Edlund, K.; Asplund, A.; Sjostedt, E.; Lundberg, E.; Szigyarto, C. A.; Skogs, M.; Takanen, J. O.; Berling, H.; Tegel, H.; Mulder, J.; Nilsson, P.; Schwenk, J. M.; Lindskog, C.; Danielsson, F.; Mardinoglu, A.; Sivertsson, A.; von Feilitzen, K.; Forsberg, M.; Zwahlen, M.; Olsson, I.; Navani, S.; Huss, M.; Nielsen, J.; Ponten, F.; Uhlen, M., Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol Cell Proteomics* **2014,** 13, (2), 397-406.
112. Zeiler, M.; Straube, W. L.; Lundberg, E.; Uhlen, M.; Mann, M., A Protein Epitope Signature Tag (PrEST) library allows SILAC-based absolute quantification and multiplexed determination of protein copy numbers in cell lines. *Mol Cell Proteomics* **2012,** 11, (3), O111 009613.
113. Edfors, F.; Bostrom, T.; Forsstrom, B.; Zeiler, M.; Johansson, H.; Lundberg, E.; Hober, S.; Lehtio, J.; Mann, M.; Uhlen, M., Immunoproteomics using polyclonal antibodies and stable isotope-labeled affinity-purified recombinant proteins. *Mol Cell Proteomics* **2014,** 13, (6), 1611-24.
114. Lambert, J. P.; Ivosev, G.; Couzens, A. L.; Larsen, B.; Taipale, M.; Lin, Z. Y.; Zhong, Q.; Lindquist, S.; Vidal, M.; Aebersold, R.; Pawson, T.; Bonner, R.; Tate, S.; Gingras, A. C., Mapping differential

interactomes by affinity purification coupled with data-independent mass spectrometry acquisition. *Nat Methods* **2013,** 10, (12), 1239-45.
115. Welinder, C.; Jonsson, G.; Ingvar, C.; Lundgren, L.; Olsson, H.; Breslin, T.; Vegvari, A.; Laurell, T.; Rezeli, M.; Jansson, B.; Baldetorp, B.; Marko-Varga, G., Establishing a Southern Swedish Malignant Melanoma OMICS and biobank clinical capability. *Clin Transl Med* **2013,** 2, (1), 7.
116. Rosenling, T.; Stoop, M. P.; Smolinska, A.; Muilwijk, B.; Coulier, L.; Shi, S.; Dane, A.; Christin, C.; Suits, F.; Horvatovich, P. L.; Wijmenga, S. S.; Buydens, L. M.; Vreeken, R.; Hankemeier, T.; van Gool, A. J.; Luider, T. M.; Bischoff, R., The impact of delayed storage on the measured proteome and metabolome of human cerebrospinal fluid. *Clin Chem* **2011,** 57, (12), 1703-11.
117. Rosenling, T.; Slim, C. L.; Christin, C.; Coulier, L.; Shi, S.; Stoop, M. P.; Bosman, J.; Suits, F.; Horvatovich, P. L.; Stockhofe-Zurwieden, N.; Vreeken, R.; Hankemeier, T.; van Gool, A. J.; Luider, T. M.; Bischoff, R., The effect of preanalytical factors on stability of the proteome and selected metabolites in cerebrospinal fluid (CSF). *J Proteome Res* **2009,** 8, (12), 5511-22.
118. Govorukhina, N. I.; de Vries, M.; Reijmers, T. H.; Horvatovich, P.; van der Zee, A. G.; Bischoff, R., Influence of clotting time on the protein composition of serum samples based on LC-MS data. *J Chromatogr B Analyt Technol Biomed Life Sci* **2009,** 877, (13), 1281-91.
119. Marko-Varga, G., BioBanking as the central tool for translational medicine CTM issue 2013. *Clin Transl Med* **2013,** 2, (1), 4.
120. Marko-Varga, G.; Vegvari, A.; Welinder, C.; Lindberg, H.; Rezeli, M.; Edula, G.; Svensson, K. J.; Belting, M.; Laurell, T.; Fehniger, T. E., Standardization and utilization of biobank resources in clinical protein science with examples of emerging applications. *J Proteome Res* **2012,** 11, (11), 5124-34.
121. Vegvari, A.; Welinder, C.; Lindberg, H.; Fehniger, T. E.; Marko-Varga, G., Biobank resources for future patient care: developments, principles and concepts. *J Clin Bioinforma* **2011,** 1, (1), 24.
122. Pugh, R. S., Overview of the International Society for Biological and Environmental Repositories (ISBER) Working Groups. *Biopreserv Biobank* **2014,** 12, (5), 358-60.
123. Betsou, F.; Gunter, E.; Clements, J.; DeSouza, Y.; Goddard, K. A.; Guadagni, F.; Yan, W.; Skubitz, A.; Somiari, S.; Yeadon, T.; Chuaqui, R., Identification of evidence-based biospecimen quality-control tools: a report of the International Society for Biological and Environmental Repositories (ISBER) Biospecimen Science Working Group. *J Mol Diagn* **2013,** 15, (1), 3-16.
124. Pitt, K.; Betsou, F., The ISBER Best Practices Self Assessment Tool (SAT): Lessons learned after three years of collecting responses. *Biopreserv Biobank* **2012,** 10, (6), 548-9.
125. Mayer, G.; Jones, A. R.; Binz, P. A.; Deutsch, E. W.; Orchard, S.; Montecchi-Palazzi, L.; Vizcaino, J. A.; Hermjakob, H.; Oveillero, D.; Julian, R.; Stephan, C.; Meyer, H. E.; Eisenacher, M., Controlled vocabularies and ontologies in proteomics: overview, principles and practice. *Biochim Biophys Acta* **2014,** 1844, (1 Pt A), 98-107.
126. Chang, A.; Schomburg, I.; Placzek, S.; Jeske, L.; Ulbrich, M.; Xiao, M.; Sensen, C. W.; Schomburg, D., BRENDA in 2015: exciting developments in its 25th year of existence. *Nucleic Acids Res* **2014**.
127. Gremse, M.; Chang, A.; Schomburg, I.; Grote, A.; Scheer, M.; Ebeling, C.; Schomburg, D., The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Res* **2011,** 39, (Database issue), D507-13.
128. Sarntivijai, S.; Lin, Y.; Xiang, Z.; Meehan, T.; Diehl, A.; Vempati, U.; Schurer, S.; Pang, C.; Malone, J.; Parkinson, H.; Liu, Y.; Takatsuki, T.; Saijo, K.; Masuya, H.; Nakamura, Y.; Brush, M.; Haendel, M.; Zheng, J.; Stoeckert, C.; Peters, B.; Mungall, C.; Carey, T.; States, D.; Athey, B.; He, Y., CLO: The cell line ontology. *Journal of Biomedical Semantics* **2014,** 5, (1), 37.
129. Schriml, L. M.; Arze, C.; Nadendla, S.; Chang, Y. W.; Mazaitis, M.; Felix, V.; Feng, G.; Kibbe, W. A., Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Res* **2012,** 40, (Database issue), D940-6.

130. Cote, R.; Reisinger, F.; Martens, L.; Barsnes, H.; Vizcaino, J. A.; Hermjakob, H., The Ontology Lookup Service: bigger and better. *Nucleic Acids Res* **2010,** 38, (Web Server issue), W155-60.
131. Wilhelm, M.; Schlegl, J.; Hahne, H.; Moghaddas Gholami, A.; Lieberenz, M.; Savitski, M. M.; Ziegler, E.; Butzmann, L.; Gessulat, S.; Marx, H.; Mathieson, T.; Lemeer, S.; Schnatbaum, K.; Reimer, U.; Wenschuh, H.; Mollenhauer, M.; Slotta-Huspenina, J.; Boese, J. H.; Bantscheff, M.; Gerstmair, A.; Faerber, F.; Kuster, B., Mass-spectrometry-based draft of the human proteome. *Nature* **2014,** 509, (7502), 582-7.
132. Ezkurdia, I.; Vazquez, J.; Valencia, A.; Tress, M., Analyzing the First Drafts of the Human Proteome. *J Proteome Res* **2014**.
133. Shargunov, A. V.; Krasnov, G. S.; Ponomarenko, E. A.; Lisitsa, A. V.; Shurdov, M. A.; Zverev, V. V.; Archakov, A. I.; Blinov, V. M., Tissue-specific alternative splicing analysis reveals the diversity of chromosome 18 transcriptome. *J Proteome Res* **2014,** 13, (1), 173-82.
134. Zgoda, V. G.; Kopylov, A. T.; Tikhonova, O. V.; Moisa, A. A.; Pyndyk, N. V.; Farafonova, T. E.; Novikova, S. E.; Lisitsa, A. V.; Ponomarenko, E. A.; Poverennaya, E. V.; Radko, S. P.; Khmeleva, S. A.; Kurbatov, L. K.; Filimonov, A. D.; Bogolyubova, N. A.; Ilgisonis, E. V.; Chernobrovkin, A. L.; Ivanov, A. S.; Medvedev, A. E.; Mezentsev, Y. V.; Moshkovskii, S. A.; Naryzhny, S. N.; Ilina, E. N.; Kostrjukova, E. S.; Alexeev, D. G.; Tyakht, A. V.; Govorun, V. M.; Archakov, A. I., Chromosome 18 transcriptome profiling and targeted proteome mapping in depleted plasma, liver tissue and HepG2 cells. *J Proteome Res* **2013,** 12, (1), 123-34.
135. Jeong, S. K.; Lee, H. J.; Na, K.; Cho, J. Y.; Lee, M. J.; Kwon, J. Y.; Kim, H.; Park, Y. M.; Yoo, J. S.; Hancock, W. S.; Paik, Y. K., GenomewidePDB, a proteomic database exploring the comprehensive protein parts list and transcriptome landscape in human chromosomes. *J Proteome Res* **2013,** 12, (1), 106-11.
136. Goode, R. J.; Yu, S.; Kannan, A.; Christiansen, J. H.; Beitz, A.; Hancock, W. S.; Nice, E.; Smith, A. I., The proteome browser web portal. *J Proteome Res* **2013,** 12, (1), 172-8.
137. Wang, D.; Liu, Z.; Guo, F.; Diao, L.; Li, Y.; Zhang, X.; Huang, Z.; Li, D.; He, F., CAPER 2.0: an interactive, configurable, and extensible workflow-based platform to analyze data sets from the Chromosome-centric Human Proteome Project. *J Proteome Res* **2014,** 13, (1), 99-106.
138. Guo, F.; Wang, D.; Liu, Z.; Lu, L.; Zhang, W.; Sun, H.; Zhang, H.; Ma, J.; Wu, S.; Li, N.; Jiang, Y.; Zhu, W.; Qin, J.; Xu, P.; Li, D.; He, F., CAPER: a chromosome-assembled human proteome browsER. *J Proteome Res* **2013,** 12, (1), 179-86.
139. Muthusamy, B.; Thomas, J. K.; Prasad, T. S.; Pandey, A., Access guide to human proteinpedia. *Curr Protoc Bioinformatics* **2013,** Chapter 1, Unit 1 21.
140. Kandasamy, K.; Keerthikumar, S.; Goel, R.; Mathivanan, S.; Patankar, N.; Shafreen, B.; Renuse, S.; Pawar, H.; Ramachandra, Y. L.; Acharya, P. K.; Ranganathan, P.; Chaerkady, R.; Keshava Prasad, T. S.; Pandey, A., Human Proteinpedia: a unified discovery resource for proteomics research. *Nucleic Acids Res* **2009,** 37, (Database issue), D773-81.
141. Mathivanan, S.; Ahmed, M.; Ahn, N. G.; Alexandre, H.; Amanchy, R.; Andrews, P. C.; Bader, J. S.; Balgley, B. M.; Bantscheff, M.; Bennett, K. L.; Bjorling, E.; Blagoev, B.; Bose, R.; Brahmachari, S. K.; Burlingame, A. S.; Bustelo, X. R.; Cagney, G.; Cantin, G. T.; Cardasis, H. L.; Celis, J. E.; Chaerkady, R.; Chu, F.; Cole, P. A.; Costello, C. E.; Cotter, R. J.; Crockett, D.; DeLany, J. P.; De Marzo, A. M.; DeSouza, L. V.; Deutsch, E. W.; Dransfield, E.; Drewes, G.; Droit, A.; Dunn, M. J.; Elenitoba-Johnson, K.; Ewing, R. M.; Van Eyk, J.; Faca, V.; Falkner, J.; Fang, X.; Fenselau, C.; Figeys, D.; Gagne, P.; Gelfi, C.; Gevaert, K.; Gimble, J. M.; Gnad, F.; Goel, R.; Gromov, P.; Hanash, S. M.; Hancock, W. S.; Harsha, H. C.; Hart, G.; Hays, F.; He, F.; Hebbar, P.; Helsens, K.; Hermeking, H.; Hide, W.; Hjerno, K.; Hochstrasser, D. F.; Hofmann, O.; Horn, D. M.; Hruban, R. H.; Ibarrola, N.; James, P.; Jensen, O. N.; Jensen, P. H.; Jung, P.; Kandasamy, K.; Kheterpal, I.; Kikuno, R. F.; Korf, U.; Korner, R.; Kuster, B.; Kwon, M. S.; Lee, H. J.; Lee, Y. J.; Lefevre, M.; Lehvaslaiho, M.; Lescuyer, P.; Levander, F.; Lim, M. S.; Lobke, C.; Loo, J. A.; Mann, M.; Martens, L.; Martinez-Heredia,

J.; McComb, M.; McRedmond, J.; Mehrle, A.; Menon, R.; Miller, C. A.; Mischak, H.; Mohan, S. S.; Mohmood, R.; Molina, H.; Moran, M. F.; Morgan, J. D.; Moritz, R.; Morzel, M.; Muddiman, D. C.; Nalli, A.; Navarro, J. D.; Neubert, T. A.; Ohara, O.; Oliva, R.; Omenn, G. S.; Oyama, M.; Paik, Y. K.; Pennington, K.; Pepperkok, R.; Periaswamy, B.; Petricoin, E. F.; Poirier, G. G.; Prasad, T. S.; Purvine, S. O.; Rahiman, B. A.; Ramachandran, P.; Ramachandra, Y. L.; Rice, R. H.; Rick, J.; Ronnholm, R. H.; Salonen, J.; Sanchez, J. C.; Sayd, T.; Seshi, B.; Shankari, K.; Sheng, S. J.; Shetty, V.; Shivakumar, K.; Simpson, R. J.; Sirdeshmukh, R.; Siu, K. W.; Smith, J. C.; Smith, R. D.; States, D. J.; Sugano, S.; Sullivan, M.; Superti-Furga, G.; Takatalo, M.; Thongboonkerd, V.; Trinidad, J. C.; Uhlen, M.; Vandekerckhove, J.; Vasilescu, J.; Veenstra, T. D.; Vidal-Taboada, J. M.; Vihinen, M.; Wait, R.; Wang, X.; Wiemann, S.; Wu, B.; Xu, T.; Yates, J. R.; Zhong, J.; Zhou, M.; Zhu, Y.; Zurbig, P.; Pandey, A., Human Proteinpedia enables sharing of human protein data. *Nat Biotechnol* **2008,** 26, (2), 164-7.

142. Whiteaker, J. R.; Halusa, G. N.; Hoofnagle, A. N.; Sharma, V.; MacLean, B.; Yan, P.; Wrobel, J. A.; Kennedy, J.; Mani, D. R.; Zimmerman, L. J.; Meyer, M. R.; Mesri, M.; Rodriguez, H.; Paulovich, A. G., CPTAC Assay Portal: a repository of targeted proteomic assays. *Nat Methods* **2014,** 11, (7), 703-4.

143. Guo, X.; Trudgian, D. C.; Lemoff, A.; Yadavalli, S.; Mirzaei, H., Confetti: a multiprotease map of the HeLa proteome for comprehensive proteomics. *Mol Cell Proteomics* **2014,** 13, (6), 1573-84.

144. Omenn, G. S.; States, D. J.; Adamski, M.; Blackwell, T. W.; Menon, R.; Hermjakob, H.; Apweiler, R.; Haab, B. B.; Simpson, R. J.; Eddes, J. S.; Kapp, E. A.; Moritz, R. L.; Chan, D. W.; Rai, A. J.; Admon, A.; Aebersold, R.; Eng, J.; Hancock, W. S.; Hefta, S. A.; Meyer, H.; Paik, Y. K.; Yoo, J. S.; Ping, P.; Pounds, J.; Adkins, J.; Qian, X.; Wang, R.; Wasinger, V.; Wu, C. Y.; Zhao, X.; Zeng, R.; Archakov, A.; Tsugita, A.; Beer, I.; Pandey, A.; Pisano, M.; Andrews, P.; Tammen, H.; Speicher, D. W.; Hanash, S. M., 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, (13), 3226-45.

145. States, D. J.; Omenn, G. S.; Blackwell, T. W.; Fermin, D.; Eng, J.; Speicher, D. W.; Hanash, S. M., Challenges in deriving high-confidence protein identifications from data gathered by a HUPO plasma proteome collaborative study. *Nat Biotechnol* **2006,** 24, (3), 333-8.

146. Barrett, T.; Wilhite, S. E.; Ledoux, P.; Evangelista, C.; Kim, I. F.; Tomashevsky, M.; Marshall, K. A.; Phillippy, K. H.; Sherman, P. M.; Holko, M.; Yefanov, A.; Lee, H.; Zhang, N.; Robertson, C. L.; Serova, N.; Davis, S.; Soboleva, A., NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res* **2013,** 41, (Database issue), D991-5.

147. Edgar, R.; Domrachev, M.; Lash, A. E., Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **2002,** 30, (1), 207-10.

148. Guruceaga, E.; Sanchez Del Pino, M. M.; Corrales, F. J.; Segura, V., Prediction of a missing protein expression map in the context of the human proteome project. *J Proteome Res* **2015,** 14, (3), 1350-60.

149. Islam, M. T.; Garg, G.; Hancock, W. S.; Risk, B. A.; Baker, M. S.; Ranganathan, S., Protannotator: a semiautomated pipeline for chromosome-wise functional annotation of the "missing" human proteome. *J Proteome Res* **2014,** 13, (1), 76-83.

150. Ranganathan, S.; Khan, J. M.; Garg, G.; Baker, M. S., Functional annotation of the human chromosome 7 "missing" proteins: a bioinformatics approach. *J Proteome Res* **2013,** 12, (6), 2504-10.

151. P, G.; PA, M.; M, Z.-Z.; I, C.; P.D, D.; O, E.; A, G.; A, G.; M, P.; D, T.; Y, Z.; L, L.; A, B., The neXtProt knowledgebase on human proteins: current status. *Nucleic Acids Research* **in press**.

152. Gaudet, P.; Argoud-Puy, G.; Cusin, I.; Duek, P.; Evalet, O.; Gateau, A.; Gleizes, A.; Pereira, M.; Zahn-Zabal, M.; Zwahlen, C.; Bairoch, A.; Lane, L., neXtProt: Organizing Protein Knowledge in the Context of Human Proteome Projects. *J Proteome Res* **2012,** 12, (1), 293-298.