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


†Analytical Biochemistry, Department of Pharmacy, University of Groningen, A. Deusinglaan 1, 9713 AV Groningen, The Netherlands
‡Science for Life Laboratory, KTH - Royal Institute of Technology, SE-171 21 Stockholm, Sweden
§Institute of Chemistry, Academia Sinica, 128 Academia Road Sec. 2, Taipei 115, Taiwan
∥Institute of Information Science, Academia Sinica, 128 Academia Road Sec. 2, Taipei 115, Taiwan
⊥The State Key Laboratory of Proteomeics, Beijing Proteome Research Center, Beijing Institute of Radiation Medicine, No. 27 Taiping Road, Haidian District, Beijing 100850, China
#Department of Biochemistry and Molecular Biology, Monash University, Clayton, Victoria 3800, Australia
△Department of Chemistry and Biомolecular Sciences and ARC Centre of Excellence in Bioinformatics, Macquarie University, Sydney, New South Wales 2109, Australia
▼Australian School of Advanced Medicine, Macquarie University, Sydney, NSW 2109, Australia
○Proteomics Unit, Institute of Chemistry, Federal University of Rio de Janeiro, Cidade Universitária, Av Athos da Silveira Ramos 149, CT-AS42, 21941-909 Rio de Janeiro, RJ, Brazil
■Beijing Institute of Genomics and BGI Shenzhen, No. 1 Beichen West Road, Chaoyang District, Beijing 100101, China
◆Key Laboratory of Functional Protein Research of Guangdong Higher Education Institutes, College of Life Science and Technology, Jinan University, Guangzhou 510632, China
●Department of Computational Medicine & Bioinformatics, University of Michigan, 100 Washtenaw Avenue, Ann Arbor, Michigan 48109-2218, United States
▽Departments of Computational Medicine & Bioinformatics and Computer Sciences, University of Michigan, 100 Washtenaw Avenue, Ann Arbor, Michigan 48109-2218, United States
▲ProteoRed-ISCIII, Biomolecular and Bioinformatics Resources Platform (PRB2), Spanish Consortium of C-HPP (Chr-16), CIMA, University of Navarra, 31008 Pamplona, Spain
◆Chr16 SpHPP Consortium, CIMA, University of Navarra, 31008 Pamplona, Spain
◇Department of Cellular and Molecular Medicine, Centro de Investigaciones Biológicas (CIB-CSIC), 28040 Madrid, Spain
★Centro Nacional de Biotecnología (CNB-CSIC), Cantoblanco, 28049 Madrid, Spain
☆Cancer Research Center. Proteomics Unit and General Service of Cytometry, Department of Medicine, University of Salamanca-CSIC, IBISAL, Campus Miguel de Unamuno s/n, 37007 Salamanca, Spain
$qUniversity of Victoria-Genome British Columbia Proteomics Centre, Vancouver Island Technology Park, #3101–4464 Markham Street, Victoria, British Columbia V8Z 7X8, Canada
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 the key approaches and their roles in the project. We present new analytical technologies that improve the chemical space and lower detection limits coupled to 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 content has 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.

KEYWORDS: missing proteins, Chromosome-Centric Human Proteome Project, LC–MS, antibody enrichment, proteomics, bioinformatics

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 biomedical 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).1,2 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”,3,4 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)\(^3\)\(^-\)\(^5\) 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 proteomic data sets, containing at least the raw MS data accompanied by experimental and technical metadata. By November 2014, ProteomeXchange resources stored ∼1500 data sets (∼50% of which are publicly available) from a wide variety of sources, with humans being the most-represented species. Once the data sets are made publicly available, they are usually reprocessed by PeptideAtlas\(^9\)\(^-\)\(^11\) using the X!Tandem\(^19\)\(^,\)\(^20\) database search tool. Basically, every one 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\(^21\)\(^,\)\(^22\) is adding information at the genomic, transcriptomic, and proteomic levels to the corpus of information available in UniProtKB.\(^23\)\(^-\)\(^25\) In particular, neXtProt integrates

Table 1. List of Posters Presented at C-HPP Poster Session on October 7, 2014 at HUPO 2014 (Madrid) and Used as Second Reference (Poster Number) in This Paper

<table>
<thead>
<tr>
<th>poster number</th>
<th>title</th>
<th>Chr team</th>
<th>main topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mining Missing Membrane Proteins from Lung Cancer Tissues and Cell lines</td>
<td>4</td>
<td>proteomics technology: opening chemical space</td>
</tr>
<tr>
<td>2</td>
<td>CAPER 3.0: a scalable cloud-based pipeline for the data-intensive analysis of proteomic data sets</td>
<td>1, 8, 20</td>
<td>bioinformatics: proteome browser</td>
</tr>
<tr>
<td>3</td>
<td>Chromosome 18: Master Proteome</td>
<td>18</td>
<td>proteomics technology: lowering detection limit</td>
</tr>
<tr>
<td>4</td>
<td>GenomewidePDB v 2.0: Update on the transcriptomic and proteomic expression data with alternatively spliced products layered in a genome-wide manner</td>
<td>13</td>
<td>bioinformatics: proteome browser</td>
</tr>
<tr>
<td>5</td>
<td>Development of improved subcellular fractionation procedures of the placental membrane proteins for discovering disease biomarkers and missing proteins</td>
<td>13</td>
<td>proteomics technology: lower detection limit</td>
</tr>
<tr>
<td>6</td>
<td>Chromosome X</td>
<td>X</td>
<td>missing protein: identification strategy</td>
</tr>
<tr>
<td>7</td>
<td>The Mitochondrial Human Proteome Project - MT-HPP</td>
<td>mitochondria</td>
<td>missing protein: identification strategy</td>
</tr>
<tr>
<td>8</td>
<td>Revisiting the Identification of Canonical Splice Isoforms through Integration of Functional Genomics and Proteomics Evidence from the Chromosome 17 Human Proteome Project</td>
<td>17</td>
<td>exploring effect of genetic variability</td>
</tr>
<tr>
<td>9</td>
<td>Subcellular fractionation enhances Chromosome 16 Proteome Coverage</td>
<td>16</td>
<td>proteomics technology: lowering detection limit</td>
</tr>
<tr>
<td>10</td>
<td>Proteogenomics Dashboard for the Human Proteome Project</td>
<td>16</td>
<td>bioinformatics: proteogenomic data mining</td>
</tr>
<tr>
<td>11</td>
<td>Missing proteins in Chromosome 16 Spanish HPP</td>
<td>16</td>
<td>missing protein: identification strategy</td>
</tr>
<tr>
<td>12</td>
<td>Transcriptomic profiling toward the localization of the missing proteins</td>
<td>16</td>
<td>missing proteins: identification with proteogenomic</td>
</tr>
<tr>
<td>13</td>
<td>The Chromosome 19 Strategy to Characterize Novel Proteoforms and Missing Proteins Using ENCODE Resources</td>
<td>19</td>
<td>missing protein: identification with proteogenomic</td>
</tr>
<tr>
<td>14</td>
<td>Targeting proteins of chromosome 16</td>
<td>16</td>
<td>proteomics technology: lower detection limit</td>
</tr>
<tr>
<td>15</td>
<td>Dissecting Chromosome 16 proteome</td>
<td>16</td>
<td>missing protein: identification strategy</td>
</tr>
<tr>
<td>16</td>
<td>Unified Human Proteome Browser Initiative</td>
<td>all</td>
<td>bioinformatics: proteome browser</td>
</tr>
<tr>
<td>17</td>
<td>Expression of αvβ6 integrin enhances both plasminogen and latent-transforming growth factor-β1 dependent proliferation, invasion and ERK1/2 signaling in colorectal cancer cells</td>
<td>7</td>
<td>proteomics technology: application</td>
</tr>
<tr>
<td>18</td>
<td>Overexpression of αvβ6 integrin alters the colorectal cancer cell proteome in favor of elevated proliferation and a switching in cellular adhesion which increases invasion</td>
<td>7</td>
<td>proteomics technology: application</td>
</tr>
<tr>
<td>19</td>
<td>Approaching the organelar brain proteome to understand the molecular basis of Schizophrenia</td>
<td>15</td>
<td>protein quantification: application. missing proteins</td>
</tr>
<tr>
<td>20</td>
<td>Splice Variants in Aggressive Human Triple Negative Breast Cancer</td>
<td>17</td>
<td>exploring effect of genetic variability</td>
</tr>
<tr>
<td>21</td>
<td>Revisiting the Identification of Canonical Splice Isoforms through Integration of Functional Genomics and Proteomics Evidence from the Chromosome 17 Human Proteome Project</td>
<td>17</td>
<td>exploring effect of genetic variability</td>
</tr>
</tbody>
</table>

The ProteomeXchange\(^6\) consortium, led by PRIDE\(^7\)\(^,\)\(^8\) at the European Bioinformatics Institute (Hinxton/Cambridge, U.K.) and by PeptideAtlas\(^9\)\(^-\)\(^11\) at the Institute for Systems Biology (Seattle, WA), 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 data sets, 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 ∼1500 data sets (∼50% of which are publicly available) from a wide variety of sources, with humans being the most-represented species. Once the data sets are made publicly available, they are usually reprocessed by PeptideAtlas\(^9\)\(^-\)\(^11\) using the Trans Proteomic Pipeline (TPP)\(^12\)\(^-\)\(^16\) and by the Global Proteome Machine Database (GPMDB)\(^17\)\(^,\)\(^18\) using the X!Tandem\(^19\)\(^,\)\(^20\) database search tool. Basically, every one 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\(^21\)\(^,\)\(^22\) is adding information at the genomic, transcriptomic, and proteomic levels to the corpus of information available in UniProtKB.\(^23\)\(^-\)\(^25\) In particular, neXtProt integrates
genomic variation data from dbSNP and COSMIC, transcriptomic data from BGee, antibody-based protein evidence from the Human Protein Atlas (HPA), MS-based information from PeptideAtlas, 3D structural information from the Protein Data Bank, and various PTM information from manually curated literature. On the basis of 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, 3D 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 of 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 noncoding RNAs according to the protein annotation from different resources (HGNC, RefSeq, HAVANA, CCDS, UniProtKB/Swiss-Prot). 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 or convincing proteomics data, and 11 have been upgraded to the PE2 or PE3 categories. On the basis of these numbers, one can estimate that <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, that is, 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). (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 in
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 on the human genome. (4) They reflect erroneous annotation of the genome, which results in incorrectly predicted protein sequences; the genome is undergoing reassessment on a frequent basis. (5) Many highly homologous proteins or proteins with large sequence variability are missed or not counted due to the parsimonious protein assembly 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 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 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 2. 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 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 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.

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 <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. The most widely used approach for protein identification is
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. 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 <30 kDa, the available fragmentation spectra are often far from complete, and the obtained complex spectra are often difficult to interpret. 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. Ensembl contains a reference genome and includes annotation from the Encyclopedia of DNA Elements (ENCODE), 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 are based on the protein sequences stored in public databases such as UniProtKB or NCBI RefSeq 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 propose the existence of one or another of the 616 dubious human proteins currently annotated as PES in neXtProt. It can also support identification of new ORFs and translated noncoding mRNA or redefine the starting and ending parts of protein-coding regions, as reported by Kim et al. However, when protein identification is performed exclusively with a translated mRNA sequence, the much shorter half life of mRNA compared with proteins 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 times. For other cases, the use of 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. Wang et al. 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 (SAAV, RNA editing, and alternative splicing), and integration of the MS data. 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 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, for example, 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 was performed to enrich low abundant proteins was presented by Chang et al. 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. On the basis of 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 protein entries with translation evidence were missing from the LC-MS/MS data acquired from institutions. The inability to detect certain proteins by LC-MS/MS was most probably a result 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 [posters 8 and 13]. This effect was shown by Menon et al., 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 noncanonical 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 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. 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 with transcription patterns in human neuronal stem cells. 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, proteome searches of high-resolution LC–MS/MS data of GSC protein digests identiﬁed 19 SAAVs in 17 chromosome 19 proteins. 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. In summary, the current status of chromosome 19 in 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.

---

**FIGURE 3.** Number of chromosome 19 genes and the identiﬁed molecular entities at transcript and expression levels (mRNA and proteins) are illustrated as a proteogenomic analysis of glioma stem cells addressing 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 SAAVs, is proportionally similar to other human chromosomes.

---

**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 embracing a wide chemical space that should be covered by the proteomics analytical approach. In addition, artiﬁcial modiﬁcations intro¬duced by the sampling protocol need to be considered. This large chemical space is well-covered, but not completely, 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 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 nonliver 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. 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 olfactory function.

Missing protein identification can be enhanced by developing specific enrichment methods such as the use of Proteominer beads and enrichment of protein aggregates (Chen Y, Li Y, Zhong J, Zhang J, Chen Z, Yang L, Cao X, He QY, Zhang G, Wang T. J Proteome Res. 2015 Jun 25. PMID:26108252); specific analytical methods for hydrophobic proteins; a specific fractionation method such as the analysis of subcellular fractions (posters S, 9, and 19); and methods to increase protein sequence coverage (e.g., by using multiple proteases for protein cleavage or by using a more efficient method of peptide fragmentation such as ETHcD). The membrane subproteome was suggested to be a rich source of missing proteins. A deep sequencing strategy using complementary 2D 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. 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), repository of SRM assay results (PASSEL), a database of ranked peptides and SRM transitions for all proteins in selected proteomes (SRMAtlas) 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 to the processing and analysis of the acquired data, is supported by step-specific algorithms and comprehensive bioinformatics pipelines to plan SRM assays for missing proteins, such as ATAQs, mQuest, MaRiMba, SMRBuilder, and Skyline. The PeptidePicker tool developed by Mohammed et al. 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, GPMDM, 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 nontargeted fragment spectra by fragmenting large windows of precursor ions (typically 20 to 25 m/z). The resulting MS/MS data can be reconstituted from the coeluted 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. 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 and other analogous processing software, providing 51% of coverage of canonical UniProtKB/Swiss-Prot entries.

Another important resource for the identification of missing proteins is SRMAtlas, 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 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.

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. 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. SISCAPA can enhance the sensitivity of SRM analyses by enriching specific peptides.

The HPA project has systematically generated affinity-purified polyclonal antibodies using proteospecific recombinant protein fragment and Protein Epitope Signature Tags (PrESTs). 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. 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. Immuno-SILAC has proved to be 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.

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. 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, 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, an international organization that has worked out regulatory and ethical protocols and Best Practice guidelines for such purposes.

Controlled vocabularies and ontologies pioneered by SNOMED providing standardized anatomical descriptors related to tissue types (BRENDA), cell types (Cell Ontology), and human diseases (DOID, http://disease-ontology.org/) 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 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 and olfactory cortex tissue. An assessment of the number of reported olfactory receptors in Kim et al. and Wilhelm et al. by Ezkurdi et al. showed that these two large-scale studies had used poor MS/MS spectra to identify more than 100 olfactory receptors each, 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 data sets. Therefore, the statistical criteria must be a 1% FDR or better calculated at the protein level for the combined data set, as adopted by PeptideAtlas. Using a 1% FDR threshold at the PSM or peptide level would result in a large number of misidentified or indistinguishable proteins when analyzing 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 data sets. For example, if one million PSMs 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 data sets from which 3000 proteins are identified, a 1% protein-level FDR implies only 30 incorrect protein identifications. However, for very large data sets from which 15,000 proteins are identified, a 1% protein-level FDR would result in 150 misidentified proteins, which is a considerable number. As described by Savitski et al. and Omenn et al., new methods are being introduced to
optimize finding true positives while controlling false positives in large heterogeneous data sets. C-HPP will stringently identify olfactory receptors in nasal tissue accompanied by 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, GenomeWideDB (poster 4), Human Proteome Map, proteomicsDB, PeptideAtlas, HPA, The Proteome Browser, CAPER (poster 2), and Human Proteinpedia. 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 a 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 data sets (the CPTAC repository and those of Kim et al., Wilhelm et al., Guo et al.) to PeptideAtlas only increased the amount of level 1 protein evidence for 1365 neXtProt entries using stringent error thresholds of 0.00009 FDR for PSMs, 0.0003 FDR at peptide level, and 0.011 FDR at protein level identification. The successive increments in HumanAllBuild 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. and Wilhelm et al. to be around 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, including the primary research groups, as launched in Madrid.

The proteomics community has a great deal of experience with overcalling 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 highlighted a “core dataset” of 3020 proteins with two or more peptide matches but clearly delineated a broad range of numbers with other criteria. In contrast, States et al. 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. 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.

The Spanish chromosome 16 team developed a method using transcription data from public repositories (GEO) 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. All of these data will be available in the dashHPPBoard 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.

To support the C-HPP initiative, Islam et al. 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 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 of 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 revealed proteomic evidence for 107 missing proteins, with evidence of an additional 15 missing proteins using the data of a recent membrane proteomic study.

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.

**Contributors**

Eric Deutsch, Unit of Protein Mass Spectrometry, CNB-CSIC, Madrid, Spain; Farrah et al.

**Acknowledgements**

The authors would like to acknowledge the support of the European Molecular Biology Laboratory (EMBL) and the Jackson Laboratory for Genomic Medicine (JLGM) for their contributions to the Human Plasma Proteome Project (HPPP). We also thank the Human Protein Reference Database (HPRD) and the Human Plasma Proteome Project (HPPP) for their contributions to the Human Proteome Browsers to support the chromosome-centric integration, processing, and visualization of proteogenomic data.
CONCLUSIONS

The reduction of the proportion of missing proteins in the human proteome from 33 to 18% (or 15%) over the last 3 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 (posters 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.\textsuperscript{133} examined the spectra for hundreds of olfactory receptor proteins claimed by Kim et al.\textsuperscript{73} and by Wilhelm et al.,\textsuperscript{132} 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 arises from genetic variability, such as SAAVs, 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 4 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 of 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 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/).

ASSOCIATED CONTENT

Supporting Information

Sets of protein identifications at various levels of redundancy under the Cedar scheme. Peptide-centric illustration of six protein sequences in a hypothetical ProteinProphet protein group, in order of descending ProteinProphet probability. Origins of the proteform terminology to clearly describe biological variability at the level of protein primary structure. All data (posters, presentations) used in the series of C-HPP workshops held during 2014. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/pr5013009.

AUTHOR INFORMATION

Corresponding Author

E-mail: p.l.horvatovich@rug.nl. Tel: +31-50-363-3341. Fax: +31-50-363-7582.

Notes

The authors declare no competing financial interest.

We 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.

We pay tribute to Thomas Edward Fehninger, a friend and recognized scientist, who passed away in July 2015 in Lund.

ACKNOWLEDGMENTS

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. C.L.N. 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).

ABBREVIATIONS

AV, 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; CHDPC, 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; ETThD, 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; SAAV, 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.
REFERENCES


(49) 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.

(50) Sheynkman, G. M.; Shortreed, M. R.; Frey, B. L.; Scaf1, M.; Smith, L. M. Large-scale mass spectrometric detection of variant peptides resulting from nonsynonymous nucleotide differences. J. Proteome Res. 2014, 13 (1), 228–240.


