Published in final edited form as:

Methods Mol Biol. 2011 ; 696: 247–257. doi:10.1007/978-1-60761-987-1_15.

Data Management and Data Integration in the HUPO Plasma Proteome Project

Gilbert S. Omenn

Abstract

The Human Plasma Proteome Project (HPPP) is an international collaboration coordinated by the Human Proteome Organisation (HUPO). Its Pilot Phase generated the 2005 *Proteomics* special issue “Exploring the Human Plasma Proteome” (Omenn et al. *Proteomics* 5:3226–3245, 2005) and a book with the same title (Omenn GS (ed) (2006) *Exploring the human plasma proteome*. Wiley-Liss, Weinheim, pp 372). Data management for that Pilot Phase included collection, integration, analysis, and dissemination of findings from participating laboratories and data repositories. Many investigators face the same challenges of integration of data from complex, dynamic serum, and plasma specimens. The PPP workflow assembled a representative Core Dataset of 3,020 protein identifications, overcoming ambiguity and redundancy in the heterogeneous contributed identifications and redundancy and updates in the protein sequence databases. The results were made available with alternative thresholds from the University of Michigan, yielding a range of numbers of protein identifications. Data were submitted to EBI/PRIDE and to ISB/PeptideAtlas. The current phase of the PPP employs Proteome Xchange to link submission of well-annotated primary datasets to EBI/PRIDE, distributed file sharing by Tranche/ Proteome Commons.org, and reanalysis from the primary raw spectra at ISB/PeptideAtlas. Such human plasma proteome datasets are available for data mining comparisons with the proteomes of other organs and biofluids in health and disease.

1. Introduction

The database of 3,020 protein identifications from the large collaborative Human Plasma Proteome Project (HPPP)(1, 2), organized as the first initiative of the Human Proteome Organisation (HUPO) in 2002, has been widely utilized and has been cited 252 times as of 21 January 2010. Thus, it is desirable for users to understand its organization and especially the data management and data integration features that are critical to cross-comparison of findings from different studies. The challenges of data management and data integration across dozens of participating laboratories remain highly relevant in the field, especially the objective of obtaining full annotation of samples. The HUPO Protein Standards Initiative (PSI) has addressed many aspects of standardization of data formats and data submission (psidev.sf.net).

Compromises typically must be accepted on the level of detail of experimental methodology, starting with the protocol and variation in collection and processing of blood specimens; the choices of reference specimens; the capture of information that is embedded in free text; the

uncertainty of identifications when laboratories are mandated to push the limitation of detectability; the parameters used by various mass spectrometry instruments; the design of data storage systems; and the choice of sequence database (and version) used for analysis (see Note 1).

This chapter describes the guidelines for data submission, the creation of the data repository, the array of specially prepared reference specimens, the handling of MS/MS data, the data integration workflow algorithm, and the consolidation and annotation of datasets from 18 laboratories that submitted MS/MS findings on the HUPO PPP reference specimens. Results from these and other platforms were published in (1).

2. Methods

2.1. Creating a Data Repository

The HPPP adopted a data model (3) focused on identifications of whole proteins with a high-level, concise description of experimental results and a minimum of data input, transmission, and reformatting for the collaborating submitters. Guidance specified protein accession numbers and names, binary description of the confidence of the protein identification (with common parameters), lists of identified peptides, and free text descriptions of experimental protocols, estimates of relative abundance, and any information about posttranslational modifications (see Note 2). Identification datasets were stored as peptide lists, reflecting the fact that many laboratories applied intact protein fractionation before tryptic digestion and mass spectrometry. During the PPP Pilot Phase, we later requested peak lists and raw spectra in the instrument native format. Participating laboratories used different search databases and different algorithms to assemble protein identifications from the search output. The guidance anticipated the guidelines subsequently mandated by *Molecular and Cellular Proteomics* (4) and other journals, and the publication ((3), Table 1) explicitly compared the PPP data model with the Carr et al. guidelines (4). Laboratories received two distinct identifiers: a numeric public identifier used for interactions with the submission centers and other laboratories, and a three-character private code known only to the laboratory and the central data analysis group, used to create data surveys without disclosing the identity of submitters ((1), Tables 1 and 2).

The data repository was built with a Structured Query Language (SQL) relational database server, an intermediate structure presenting an exact copy of the data submitted, and the main data structure designed to hold the integrated project data. The database captured three sets of protein identifiers from the same experiment: (1) protein IDs made by data producers, in the entity identification; (2) results of peptide list searches performed by the data integration center, in the entity ProteinByPeptides; and (3) analyses by others, through the MsRun branch of the database. The entire repository structure is available in Fig. 1 of

¹Highly collaborative studies utilizing a range of technology platforms and a variety of specimens are hard to fit into a tight uniform protocol. Dealing with heterogeneous datasets requires special procedures and cross-checking, which can be enhanced by targeted data mining and data integration. Some of these features are well demonstrated in the HUPO HPPP.

²Gaining sufficient annotation of preanalytical variables (20), fractionation of specimens, MSMS analytical and search engine variables, and database matching procedures is another major challenge, with the content often less than desired.

Adamski et al. (3). Data were transmitted primarily as Excel or Word documents, even though assistance was available and promoted to prepare XML schema-based file formats.

2.2. PPP Reference Specimens

The investigators collectively decided to have a range of reference specimens to be able to address alternatives in anti-coagulation, compare plasma versus serum, and obtain preliminary results on ethnic group differences. BD Diagnostics prepared the requested specimens from pairs of donors of Caucasian–American (BD1), African–American (BD2), and Asian–American (BD3) backgrounds, after informed consent (1). Sets of four specimens were prepared: serum, EDTA-plasma, heparin-plasma, and citrate-plasma (making 12, see Note 3). A similar set of four specimens was prepared by the Chinese Academy of Medical Sciences (CAMS). Finally, the UK National Institute of Biological Standards and Control made available a lyophilized citrated plasma prepared from a pool of 25 human donors (NIBSC). Of 55 laboratories that originally committed to participate, 41 requested and received the BD1 specimens, 27 the BD2 and BD3 sets, 15 the CAMS set, and 45 the NIBSC sample. Laboratories varied markedly on how many of the specimens they actually analyzed, and how extensively they fractionated and analyzed each specimen.

2.3. Inference from Peptides to Proteins

MS/MS spectra yield sequence information for peptides, primarily but not only tryptic peptides, to be matched against protein databases. Often the search returns a cluster of proteins, all of which contain the same set of matching peptides. For a uniformly collected dataset, probabilities are readily applied with PeptideProphet/ProteinProphet (5). However, the extremely heterogeneous, collaborative nature of this dataset, with various instruments and various search engines (6), required an alternative, which is shown in Subheading 2.4. The concept of this workflow algorithm is that proteins most likely truly present are more likely to be detected across independent experiments and to have been annotated more extensively. The outcome is the choice of one protein as the representative entry from several overlapping clusters of equivalent protein identifications. As discussed later, we retained the full list to permit comparisons with proteins identified by others using different integration strategies (or none at all).

2.4. HPPP Data Integration Workflow Algorithm from Adamski et al. (3)

1. Assemble peptide sequence lists retaining all source information
2. Search the peptide lists against the IPI v2.21 database (periodically updated). Require 100% identity between the sequences; disregard flanking residues.
3. Select one representative protein from each cluster of equivalent protein matches, or intersection of several clusters.
4. Each protein entry in the reference database receives three integer scores:

³Based on the results of the HPPP Pilot Phase, we prefer and recommend the use of plasma over serum for proteomics analyses and the use of EDTA-plasma among the plasma options (1, 21).

- a. Number of labs reporting a peptide sequence list containing a sequence which maps to a cluster, including this protein
- b. Number of distinct experiments (labs $\times$ specimens $\times$ protocols) reporting a peptide list with this protein
- c. Number of identifications (labs $\times$ specimens $\times$ protocols $\times$ clusters) for clusters, including this protein. Choose cluster member with largest value of (a). In case of tie scores, proceed to (b), (c), and (d) to (h).
- d. Prefer proteins that are products of a well-described gene (not “hypothetical,” “similar to,” etc.) from Ensembl.
- e. Well-described protein-product of any gene
- f. Well-described protein not assigned to any gene
- g. Protein not assigned to any gene, described only as a fragment or similar to, etc.
- h. Select the protein having the lower IPI number (in IPI v2.21).

Score (a) counts each laboratory only once, no matter from how many specimens or with how many different peptide sequence lists the laboratory identified this protein. Next in importance, score (b) counts the number of independent experiments in which the protein was identified. Score (c) counts all reported peptide sequence lists, even if several results are from the same experiment. Criteria (d-g) indicate the level of annotation for each database entry, facilitating the selection of the best-described proteins.

2.5. Summary of Collaborative Data

The 18 laboratories that contributed MS/MS data (MALDI, LC-ESI, and FT-ICR-MS) submitted a total of 12,667 distinct protein accession numbers, using the IPI, SwissProt, and NCBI Inr databases, with IPI version 2.21 (5) the standard we chose for this project. Over time, new versions of IPI appeared, a problem for any longitudinal study or even a snapshot study with several to many months from data collection to publication. We locked in and referred back to v2.21. After integration, we had 9,504 unique proteins of 6 amino acids in length based on spectra for one or more peptides, and 3,020 proteins based on two or more peptides (see Note 4). The article (3) described in great detail the thresholds individual scientists might apply to the publicly available primary datasets. In the course of the project, we held a Jamboree Workshop (June 2004) at which participating scientists and teams from the various laboratories and informatics specialists worked together on the primary data. Several labs agreed to standardize their LCQ-MSMS SEQUEST searches to use Xcorr 1.9, 2.2, 3.75 for 1+, 2+, and 3+ ions, respectively, plus deltaCn 0.1 and Rsp 4 as the threshold for “high confidence” sequences of tryptic peptides. Note that delta Cn and Rsp are not

⁴Protein lists from the HPPP are available at: www.ebi.ac.uk/PRIDE for HUPO HPPP and individual lab submissions; <http://www.bioinformatics.med.umich.edu/hupo/ppp> includes protein lists for the 3,020 protein core dataset and its peptides, plus its corresponding 5,102 protein matches before integration. The 9,504 and 889 protein lists are also posted in this site; and embedded in www.peptideatlas.org, human plasma proteome datasets from multiple sources (not all of the HPPP datasets were included, see (14)).

always employed; they increase stringency and confidence of protein identifications. The number of lab-reported high-confidence identifications ranged from 21 to 789.

We gave emphasis to cross-laboratory confirmation of identifications (see refs. (1–3) for many details). Figure 1 shows the numbers of protein identifications according to the number of peptides per protein detected across experiments and laboratories; the dark subset in each bar represents the proportion confirmed in a second lab.

We presented a schema in Fig. 2 in the form of a diamond-shaped parallelogram with sets of proteins. The entire post-integration list of 9,504 (“all identifications”) was divided into two more stringent categories, identifications called “high confidence” by the participating investigators (2,857 proteins) and 3,020 proteins for which two or more distinct peptides were reported across all 18 laboratories reporting MSMS results, following integration. The final point in the diamond represents “high-confidence multi-peptide identifications” with 1,555 proteins. This latter set was used for comparison with the number of identifications in the HUPO Human Brain Proteome report (7).

We also published an even more restricted set of 889 proteins (8) in which we applied the Bonferroni adjustment for multiple statistical comparisons of the protein match with $p > 0.95$ among 43,730 IPI entries, as well as an adjustment for protein length to account for more opportunities for matching peptides the longer the protein sequence. The Bonferroni is a very common adjustment in large-scale transcriptomics analyses, but it is seldom utilized in proteomics. Given the many families of proteins, it is likely that this analysis, treating each protein as an independent observation, is overly stringent.

2.6. False-Positive Identifications

False-positive (FP) identifications are a widely acknowledged problem. A standard solution is to match the peptide sequences against a reversed sequence version of the protein database, such that each “reverse hit” would be a representation for a FP. In our HPPP analysis, we were dealing, as noted, with highly heterogeneous datasets and a variety of search engines (9) and database matches, so we applied a different concept. We posited that FP and true-positive identifications would show opposite behavior as one accumulates large numbers of peptide IDs. FPs would be expected to accumulate roughly proportional to the total peptide IDs, without two or more FP peptide IDs coinciding on the same database entry at any rate greater than random. In contrast, for a protein which is truly present at a detectable concentration in the specimen, increased sampling should identify the same peptides mapping to the same correct database entry, as described in ref. (1). The actual criteria varied across the 18 laboratories.

2.7. Correlating Immunoassay Quantitation of Proteins with Estimates of Abundance Based on Number of Peptides

The HPPP had a specific subproject on quantitative estimates of protein concentrations, utilizing immunoassays from several laboratories (see also Note 5). This topic received a lot of attention at the Jamboree Workshop and in the subsequent publication by Haab et al. (10). The peptide counting method, the average number of different peptides found for that IPI number across the labs reporting that IPI identification, may be regarded as a precursor to

the now-popular label-free spectral counting approach. A major challenge is figuring out what epitopes account for the immunoassay results and which of multiple proteins in a family or cluster may cross-react with the antibody, or not do so. The conclusion, with appropriate caveats, was that we obtained a log-linear relationship between immunoassay-based concentrations and number of peptides detected for a wide concentration range of proteins (see Fig. 6b in (1)). The correlation coefficient (of the log-linear relationship) for the total number of peptides matching that protein, based on quantitative immunoassays of 49 proteins among the 3,020 protein dataset, was $r = 0.86$ (1). These proteins cover quite a range of eight orders of magnitude in concentration.

2.8. Comparisons of Protein Identifications Across Different Studies

In the overview paper for the Plasma Proteome Project, we compared the protein identifications of the HPPP with those of several other authors ((1), Table 4). The amount of overlap between and among these reports was not high, reflecting especially incomplete detection of low abundance proteins, as well as uncertain numbers of false positives. An important methodological point in these comparisons is the fact that different investigators use different methods for integration of multiple matches or clusters, if they do integration at all. We found it necessary to go back to our larger datasets, both the unintegrated list of 5,102 proteins for the 3,020 core dataset and the 9,504 integrated IDs, including single peptide hits, to pick up additional matches with these datasets from different sources. Many biologically significant annotations can be generated with data mining of the HPPP (see refs (1, 2)) for numerous examples).

Comparisons of other organ proteomes with the plasma (or serum) proteome remain to be pursued. Such comparisons have been frequent statements of intent across the HUPO Initiatives, including liver, brain, kidney/urine, and cardiovascular. As noted, there has been a comparison of plasma and brain (7) and a comparison of plasma and the salivary fluid proteome (11), using the HUPO PPP for the plasma comparisons.

The PPP is a major component of the Human Plasma PeptideAtlas (12, 13). It is now desirable to utilize the entire complement of studies in the PeptideAtlas, which has the very special advantage that all of these datasets have been re-analyzed from the raw spectra at the Institute for Systems Biology with the TransProteomicPipeline, eliminating numerous sources of variation due to instrument settings, search engine parameters, and database matching and integration. Deutsch et al. published their first Human Plasma PeptideAtlas as part of the HPPP Pilot Phase publication; they identified 960 proteins from datasets that partially overlapped the datasets contributed to the HPPP (14).

In an update of the Human Plasma PeptideAtlas, Farrah et al. have reanalyzed the data from 14 of the reporting laboratories in the Pilot Phase of the HPPP using the latest TPP pipeline and the SpectraST spectral library searching tool (15) searched against the latest NIST human library (version 3.0; available at <http://www.peptideatlas.org/specplib/>). Applying

⁵One of the many interesting side analyses was the matching of our peak lists for six small datasets against microbial genomes in the NCBI Microbial (nonhuman) GenBank (June 2004 release), using X!Tandem for RefSeq protein sequence identification. We found notable bacterial and mycobacterial matches (1), a clue to the usefulness of this approach for the now very popular work on metagenomics of the huge microbial populations who share our bodies and influence many physiological functions.

extremely stringent PeptideProphet FDR thresholds, they identified 10,893 unique peptides and inferred 1,186 proteins at a 5% decoy-estimated protein false-discovery rate, and 9,807 peptides and 930 proteins at a 1% protein FDR; the entire Human Plasma PeptideAtlas, including additional HPPP current phase submissions, has 2,249 proteins at 1% FDR as of January, 2010 [data provided by Drs. Terry Farrah and Eric Deutsch].

2.9. The Next Phase, now Current Phase, of the HUPO HPPP

Under current cochairs Ruedi Aebersold, Mark Baker (succeeding Young-Ki Paik), and Gil Omenn, the HUPO HPPP continues with the intent to collect large, well-annotated datasets on human plasma in normal individuals and as part of disease-oriented studies with both organ and plasma specimen analyses (16, 17) (also see Note 6). The aims of the PPP-2 are (1) to stimulate submission of high-quality, large datasets of human plasma proteome findings with advanced technology platforms; (2) to establish a robust, value-added informatics scheme involving EBI/PRIDE, University of Michigan/ProteomeCommons/Tranche, and Institute for Systems Biology/PeptideAtlas; and (3) to collaborate with other HUPO organ-based and disease-related initiatives to make plasma the common pathway for biomarker development and application.

The initial datasets and the PRIDE Web site were demonstrated at the HPPP Workshop at the Amsterdam HUPO Congress (18). The data processing and data mining scheme calls for use of the Proteome Xchange (Fig. 3): submission of the fully annotated experimental datasets with the investigator's interpretations to EBI/PRIDE; automatic transfer to Tranche/ProteomeCommons.org for distributed file sharing globally; and automatic transfer to PeptideAtlas for full reanalysis from the raw spectra across all submissions to the new HPPP. A major element is the development of heavy-labeled proteotypic peptides based on N-glycosite peptide isolation, which will be a major resource for many kinds of proteomics studies, including high-throughput-targeted proteomics. We plan on consolidated analyses with other plasma proteome datasets already in the PeptideAtlas or received subsequently. Tranche will also contribute these datasets to the Peptidome at NIH/NCBI and to the GPMdb in Canada, and to any scientists requesting such resources. It is expected that all HUPO initiatives will contribute to the large-scale Gene-Centric Human Proteome Project now under discussion (19).

Acknowledgments

I thank all the investigators and core staff for the Pilot Phase of the HUPO HPPP (see (1)) and especially the bioinformatics team headquartered at the University of Michigan who were coauthors on the original description of the Data Management and Data Integration plan for this project: Marcin Adamski, Thomas Blackwell, Rajasree Menon, and David States of the University of Michigan and Lennart Martens, Chris Taylor, and Henning Hermjakob of the European Bioinformatics Institute (see (3)). I thank Eric Deutsch and Terry Farrah of the Institute for Systems Biology for the current data from the PeptideAtlas Human Plasma Proteome build and for review of the manuscript.

⁶Ensuring intended comparisons of results across pairs or multiples of large complex experimental projects has been frustrating and remains an important goal. Analysis for differences requires replicates to demonstrate the extent of congruence of findings upon repeat analysis of the same specimen.

References

1. Omenn GS, States DJ, Adamski MR, Blackwell TW, Menon R, Hermjakob H, et al. Overview of the HUPO plasma proteome project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. *Proteomics*. 2005; 5:3226–3245. [PubMed: 16104056]
2. Omenn, GS., editor. Exploring the human plasma proteome. Wiley-Liss; Weinheim: 2006. p. 372
3. Adamski M, Blackwell T, Menon R, Martens L, Hermjakob H, Taylor C, et al. Data management and preliminary data analysis in the pilot phase of the HUPO Plasma Proteome Project. *Proteomics*. 2005; 5:3246–3261. [PubMed: 16104057]
4. Carr S, Aebersold R, Baldwin M, Burlingame A, Clauser K, Nesvizhskii A. The need for guidelines in publication of peptide and protein identification data: Working Group on Publication Guidelines for Peptide and Protein Identification Data. *Mol Cell Proteomics*. 2004; 3:531–533. [PubMed: 15075378]
5. Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem*. 2003; 75:4646–4658. [PubMed: 14632076]
6. Kersey PJ, Duarte J, Williams A, Karavidopoulou Y, Birney E, Apweiler R. The International Protein Index: an integrated database for proteomics experiments. *Proteomics*. 2004; 4:1985–1988. [PubMed: 15221759]
7. Hamacher M, Apweiler R, Arnold G, Becker A, Blüggel M, Carrette O, et al. HUPO brain proteome project: summary of the pilot phase and introduction of a comprehensive data reprocessing strategy. *Proteomics*. 2006; 6:4890–4898. [PubMed: 16927433]
8. States DJ, Omenn GS, Blackwell TW, Fermin D, Eng J, Speicher DW, Hanash SM. Challenges in deriving high-confidence protein identifications from data gathered by HUPO plasma proteome collaborative study. *Nat Biotech*. 2006; 24:333–338.
9. Kapp EA, Schütz F, Connolly LM, Chakel JA, Meza JE, Miller CA, et al. An evaluation, comparison and accurate benchmarking of several publicly-available MS/MS search algorithms: sensitivity and specificity analysis. *Proteomics*. 2005; 5:3475–3490. [PubMed: 16047398]
10. Haab BB, Geierstanger BH, Michailidis G, Vitzthum F, Forrester S, Okon R, et al. Immunoassay and antibody microarray analysis of the HUPO PPP reference specimens: systematic variation between sample types and calibration of mass spectrometry data. *Proteomics*. 2005; 5:3278–3291. [PubMed: 16038022]
11. Yan W, Apweiler R, Balgley BM, Boontheung P, Bundy JL, Cargile BJ, et al. Systematic comparison of the human saliva and plasma proteomes. *Proteomics Clin Appl*. 2009; 3:116–134. [PubMed: 19898684]
12. Deutsch EW. The PeptideAtlas Project. *Methods Mol Biol*. 2010; 604:285–296. [PubMed: 20013378]
13. Deutsch EW, Lam H, Aebersold R. PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep*. 2008; 9:429–434. [PubMed: 18451766]
14. Deutsch EW, Eng JK, Zhang H, King NL, Nesvizhskii AI, Lin B, et al. Human Plasma PeptideAtlas. *Proteomics*. 2005; 5:3497–3500. [PubMed: 16052627]
15. Lam H, Deutsch EW, Eddes JS, Eng JK, King N, Stein SE, Aebersold R. Development and validation of a spectral library searching method for peptide identification from MS/ MS. *Proteomics*. 2007; 7:655–667. [PubMed: 17295354]
16. Omenn GS, Aebersold R, Paik YK. HUPO plasma proteome project 2007 workshop report. *Mol Cell Proteomics*. 2007; 6:2252–2253.
17. Omenn, GS., Menon, R., Adamski, M., Blackwell, T., Haab, BB., Gao, W., States, DJ. The human plasma and serum proteome. In: Thongboonkerd, V., editor. *Proteomics of human body fluids: principles, methods, and applications*. Humana Press; Totowa, NJ: 2007. p. 195-224.
18. Omenn GS, Aebersold R, Paik YK. 7th HUPO world congress of proteomics: launching the second phase of the HUPO plasma proteome project (PPP-2) 16-20 August 2008, Amsterdam, The Netherlands. *Proteomics*. 2009; 9:4–6. [PubMed: 19053147]
19. HUPO – the Human Proteome Organisation. A Gene-centric Human Proteome Project. *Mol Cell Proteomics*. 9:427–429.

20. Gelfand, C., Omenn, GS. Pre-analytical variables for plasma and serum proteome analyses. In: Ivanov, A., Lazarev, A., editors. Sample preparation in biological mass spectrometry. Springer; NY: 2010. (in press)
21. Rai AJ, Gelfand CA, Haywood BC, Warunek DJ, Yi J, Schuchard MD, et al. HUPO plasma proteome project specimen collection and handling: towards the standardization of parameters for plasma proteome samples. *Proteomics*. 2005; 5:3262–3277. [PubMed: 16052621]

Fig. 1. Distribution of MSMS and FTICR/MS protein identifications as a function of the number of peptides detected per protein (from Fig. 4 from ref. (3)). The dark portion of each bar represents the percentage confirmed in at least one additional laboratory.

Fig. 2. Alternative protein identification lists with different inclusion criteria from the HUPO Plasma Proteome Project (from Fig. 5 of ref. (3)).

Fig. 3. Scheme for Proteome Xchange, involving EBI/PRIDE, UM Tranche/ProteomeCommons.org, and ISB/PeptideAtlas, with further distribution of dataset files to the interested proteomics and bioinformatics community.