SHORT COMMUNICATION

# Human Plasma PeptideAtlas

*Eric W. Deutsch[1], Jimmy K. Eng[1], Hui Zhang[1]*, Nichole L. King[1], Alexey I. Nesvizhskii[1], Biaoyang Lin[1]*, Hookeun Lee[2]*, Eugene C. Yi[1], Reto Ossola[2]* and Ruedi Aebersold[1, 3]*

[1] Institute for Systems Biology, Seattle, WA, USA
[2] Institute for Molecular Systems Biology, ETH Zurich, Switzerland
[3] Institute for Molecular Systems Biology, ETH Zurich and Faculty of Natural Sciences, University of Zurich, Switzerland

Peptide identifications of high probability from 28 LC-MS/MS human serum and plasma experiments from eight different laboratories, carried out in the context of the HUPO Plasma Proteome Project, were combined and mapped to the EnsEMBL human genome. The 6929 distinct observed peptides were mapped to approximately 960 different proteins. The resulting compendium of peptides and their associated samples, proteins, and genes is made publicly available as a reference for future research on human plasma.

Received: March 16, 2005
Accepted: March 31, 2005

**Keywords:**
Databases / Human / Human Proteome Organisation / Plasma proteome / Tandem mass spectrometry

The protein content of human plasma is considered important for medical diagnosis and has the potential to provide a complete snapshot of the health of an individual. In addition to proteins that carry out their function within the circulatory system, plasma contains proteins that are secreted or leaked from cells and organs throughout the body. As a diagnostic tool, plasma is even more valuable by virtue of its accessibility, with millions of samples stored in clinical archives and even more obtained every year from patients.

Human plasma is thought to contain a large number of proteins, perhaps nearly all human proteins on account of low-level tissue leakage [1]. Further, human plasma also contains proteins from foreign organisms as well as millions of distinct immunoglobulins. However, a mere 22 proteins make up 99% of the mass of protein in human serum [2], and thus an investigation of the thousands of very low-abundance proteins is difficult.

**Correspondence:** Dr. Eric W. Deutsch, Institute for Systems Biology, 1441 N 34th Street, Seattle, WA 98103, USA
**E-mail:** edeutsch@systemsbiology.org
**Fax:** +1-206-732-1260

**Abbreviations: DAS**, Distributed Annotation Server; **ISB**, Institute for Systems Biology; **PPP**, Plasma Proteome Project; **SBEAMS**, Systems Biology Experiment Analysis Management System

Several recent studies have sought to provide a preliminary definition of the human plasma proteome [3–6]. Adkins *et al.* [3] performed LC-MS/MS experiments on immunoglobulin-depleted samples and reported 490 distinct proteins. Pieper *et al.* [4] identified 325 distinct proteins from samples with eight high-abundance proteins removed *via* immunoaffinity chromatography. Anderson *et al.* [5] provided a nonredundant list from four separate sources (previous literature and three other published experiments) of 1175 proteins. Chan *et al.* [6] published a list of 1444 distinct serum proteins from a large-scale LC-MS/MS experiment. A comparison of the data in these reports has shown limited overlap between studies and raised the question of how data from different plasma proteome studies could be evaluated and represented to facilitate meaningful comparisons.

HUPO has undertaken the Plasma Proteome Project (PPP), which aims to provide a comprehensive analysis of the proteins of human plasma and serum, including the analysis of variation within individuals as well as across individuals [7, 8]. As part of this project, various samples have been sent to over 40 laboratories for local analysis using a variety of protocols and platforms. Further information about this project can be found in other reports in this issue.

* Contributed unpublished data.

 

We previously developed the PeptideAtlas process [9] to create and make public a genome-mapped atlas of peptides observed in a set of LC-MS/MS proteomics experiments, initially for human and *Drosophila melanogaster*, with processing of data from additional organisms underway. Here, we present a PeptideAtlas build derived solely from human plasma (including serum) sample experiments, mostly generated for the HUPO PPP. Although the experiments were performed in different laboratories with varying protocols and platforms, the raw MS data have all been processed through the pipeline of tools developed at the Institute for Systems Biology with the goal of analyzing peptide MS/MS data consistently and with known error rates. The pipeline includes a step that assigns a probability of correctness for all putative peptide identifications. This uniform statistical validation ensures a consistent and high-quality set of peptide and protein identifications.

We assembled 28 MS/MS experiments, collectively representing 1001 LC-MS/MS runs, as summarized in Table 1. Of these experiments, 20 were the analysis of HUPO PPP standard samples, which are described elsewhere in this issue. The other eight are unpublished serum experiments, mostly performed at the Institute for Systems Biology (ISB) as part of other work. Nearly all the ISB data employ the glycopeptide capture technique [10] to mitigate the effects of the extremely abundant proteins.

**Table 1.** Summary of the contribution to the Plasma PeptideAtlas from each experiment

| Search ID | Sample tag | HUPO laboratories | No. of $p \geq 0.90$ spectra | No. of distinct peptides | No. of new distinct peptides | Is HUPO? |
|---|---|---|---|---|---|---|
| 411 | b1-CIT_glyco_lcq | 2 | 5832 | 740 | 740 | Y |
| 412 | NIBSC_glyco_lcq | 2 | 10054 | 1190 | 726 | Y |
| 414 | b1-CIT_glyco_qstar | 2 | 1379 | 306 | 61 | Y |
| 453 | HUPO12_run31 | 12 | 731 | 235 | 187 | Y |
| 454 | HUPO12_run32 | 12 | 1014 | 191 | 68 | Y |
| 455 | HUPO12_run33 | 12 | 1037 | 293 | 149 | Y |
| 456 | HUPO12_run34 | 12 | 810 | 169 | 40 | Y |
| 436 | HUPO22_M_CA_S | 22 | 9078 | 1578 | 1434 | Y |
| 399 | HUPO28_b1-CIT | 28 | 386 | 230 | 76 | Y |
| 400 | HUPO28_b1-SERUM | 28 | 514 | 289 | 64 | Y |
| 401 | HUPO28_b2-CIT | 28 | 1922 | 470 | 98 | Y |
| 402 | HUPO28_b2-SERUM | 28 | 1604 | 385 | 29 | Y |
| 403 | HUPO28_b3-CIT | 28 | 558 | 326 | 24 | Y |
| 404 | HUPO28_b3-SERUM | 28 | 556 | 307 | 15 | Y |
| 408 | HUPO29_b1-CIT_1 | 29 | 417 | 88 | 15 | Y |
| 409 | HUPO29_b1-CIT_win1 | 29 | 3008 | 549 | 155 | Y |
| 410 | HUPO29_b1-CIT_win2 | 29 | 593 | 183 | 34 | Y |
| 407 | HUPO34_b1-HEP | 34 | 8805 | 1562 | 650 | Y |
| 413 | HUPO37_b1-HEP_2LCQ | 37 | 24 | 23 | 6 | Y |
| 422 | HUPO40 | 40 | 5645 | 697 | 190 | Y |
| 254 | Serum_peo_peptides | | 7154 | 1026 | 663 | N |
| 275 | Breakfast_qtof08 | | 334 | 117 | 10 | N |
| 278 | Caex_qtof08 | | 905 | 255 | 38 | N |
| 281 | cat_ex_qtof | | 3514 | 1040 | 300 | N |
| 283 | Cation_ex_lcq | | 15751 | 2238 | 963 | N |
| 368 | PID_serum | | 4861 | 557 | 187 | N |
| 405 | HUPO28_Ref-CIT | | 373 | 257 | 5 | N |
| 406 | HUPO28_Ref-SERUM | | 337 | 224 | 2 | N |

Columns 1 and 2 provide an internal SBEAMS search batch numeric identifier and a short name (tag) for each experiment, respectively. The sample tags include the official HUPO sample names (*e.g.*, b1-SERUM) if known. Column 3 provides the HUPO laboratories from which the data are derived. The last eight experiments are serum experiments not from HUPO-provided samples, although the last two were provided by HUPO laboratories. Columns 4–6 tabulate the number of spectra identified with PeptideProphet $p \geq 0.9$, number of distinct peptides therein, and number of new distinct peptides added to the cumulative total (as plotted in Fig. 1). Clearly, the early experiments (arbitrarily sorted by HUPO laboratories number here) will have the greatest contribution to the cumulative list as nearly every peptide is new. The final column indicates if the full experimental raw data are part of the official HUPO PPP.

 

The mass spectra were searched using SEQUEST [11], and then each possible top identification was assigned a probability of being the correct identification using the PeptideProphet software [12]. The results of this automated searching and validation with an error model were loaded into an instance of the SBEAMS – Proteomics database (http://www.sbeams.org/). All peptides with a PeptideProphet probability of being correct $p$ greater or equal to 0.90 were combined in the database to form a master list of observed peptides across all these experiments. This list of peptides was then mapped to the EnsEMBL human proteome and genome, and the results are loaded into the PeptideAtlas database [9].

Beginning with over 1.9 million spectra in 1001 MS runs, the PeptideProphet analysis yielded 87 209 spectra with a probability of $p \geq 0.90$. This resulted in 6929 distinct peptides with $p \geq 0.90$. By combining the error rates in all the individual experiments, we calculated an overall false positive rate of 14% for the 6929 distinct peptides. Of these, 6342 peptides were successfully mapped to the EnsEMBL 29.35b genome build. The remainder of the peptides were identified *via* SEQUEST searching against the IPI v2.21 database [13] with sequences that are not exactly in the EnsEMBL build. This has been observed in other PeptideAtlas builds [9].

This list of 6342 distinct peptides mapped to 1606 different EnsEMBL proteins and 1131 different EnsEMBL genes; however, in many cases a single peptide mapped ambiguously to several proteins. A simple strategy for reducing the multiple mappings [9] suggested that approximately 960 proteins have been identified in these samples. There were 666 distinct proteins to which a peptide was unambiguously mapped. See Table 2 for a summary of these statistics for both the HUPO PPP sample only build and the build for all 28 experiments.

**Table 2.** Summary of HUPO-only and all data Plasma PeptideAtlas builds

| HUPO only | ALL data | Statistic |
|---|---|---|
| 20 | 28 | Experiments (samples) included in build |
| 727 | 1001 | MS runs (mass spectrometer output files) |
| 1568528 | 1943440 | MS/MS spectra searched with SEQUEST |
| 53976 | 87209 | MS/MS spectra scored $p \geq 0.9$ by PeptideProphet |
| 4761 | 6929 | Distinct peptide sequences |
| 4416 | 6342 | Distinct peptides that mapped to EnsEMBL 29.35b |
| 1058 | 1606 | Possible proteins implicated in mapping |
| 755 | 1131 | Possible genes implicated in mapping |
| 622 | 960 | Simple reduced proteins (correction for ambiguous mappings) |
| 436 | 666 | Unambiguously mapped proteins (contain nondegenerate peptide) |

Columns 1 and 2 list the statistics for the HUPO-only data and all plasma/serum experiment PeptideAtlas builds, respectively. These statistics are discussed further in the text.

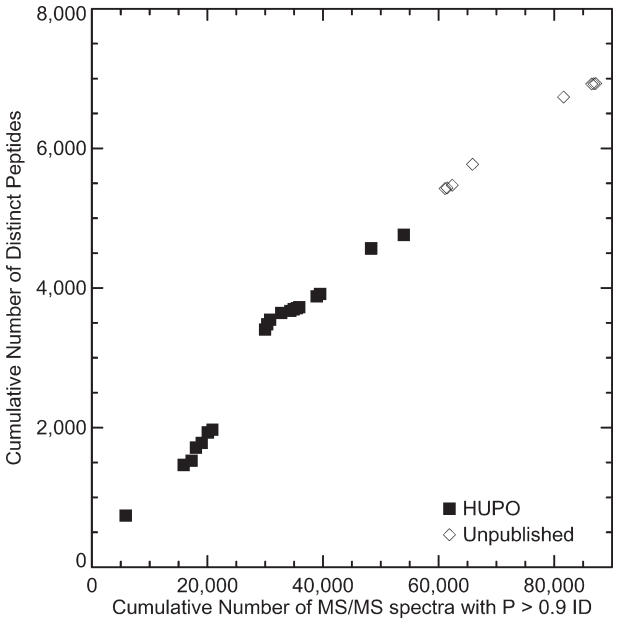


**Figure 1.** Cumulative number of distinct peptides as a function of the addition of more MS/MS spectra identified with $p \geq 0.9$. Eventually the pattern is expected to show saturation as most observable peptides are cataloged. However, at present, it still appears that ~65 new peptides are still cataloged *per* 1000 identified spectra added.

The accumulation of new distinct peptides as additional identified MS/MS spectra were added to the process is summarized in Fig. 1. Each point represents the addition of another experiment, arbitrarily sorted as shown in Table 1. The initial experiments contributed greatly to the cumulative numbers of distinct peptides, but the trend did become somewhat shallower as expected. The curve will asymptotically approach the total number of detectable peptides (with the used technologies and techniques). However, this level is far from being reached. At this point, approximately 65 new distinct peptides are being added for every 1000 new $p \geq 0.90$ spectra. This is a rate somewhat smaller than that observed in the main PeptideAtlas build [9].

We compared the results of the Plasma PeptideAtlas build with the compendium of plasma proteins of Anderson *et al.* [5] derived from four other sources. We mapped the proteins in that source to EnsEMBL proteins and then determined which of those proteins are in the Plasma PeptideAtlas. Some proteins from Anderson *et al.* did not map to EnsEMBL readily with the accession numbers given, and were excluded for the purpose of this comparison. Of the proteins found in all the four sources, all are found in the Plasma PeptideAtlas. For the proteins found in at least three, two, and one sources, we find in the Plasma PeptideAtlas 96, 76, and 27%, respectively.

The collaborative analysis of all the HUPO samples obtained from 18 laboratories yielded a total of 3020 proteins for which at least two peptides were reported in two different

 

analyses [14]. We compared this set to the results of our Plasma PeptideAtlas build based purely on the HUPO samples with $p \geq 0.90$ (false positive rate ~14%), and found that our build contains 479 of the 3020 proteins.

We have set up the Plasma PeptideAtlas data as a DAS source that can be browsed using the EnsEMBL genome browser. Instructions on configuring the EnsEMBL browser to view these data can be found on the PeptideAtlas website.

The compendium of peptides, derived from this large set of LC-MS/MS experiments on human plasma and serum samples, is publicly available for future studies. As part of the PeptideAtlas project, we will continue to accept submission of raw MS data derived from human plasma samples and publicly release new builds of the Human Plasma PeptideAtlas at our website http://www.peptideatlas.org/ with an increasing set of experiments. In addition to the build results, the raw mass spectrometer output for all published or otherwise public datasets are downloadable in mzXML [15] format from our repository.

*This work has been funded in part with federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, under contract no. N01-HV-28179. We gratefully acknowledge HUPO laboratories 12, 22, 28, 29, 34, 37, and 40 for allowing us to use these data in the Plasma PeptideAtlas.*

## References

[1] Anderson, N. L., Anderson, N. G., *Mol. Cell. Proteomics* 2002, *1*, 845–867.

[2] Tirumalai, R. S., Chan, K. C., Prieto, D. A., Issaq, H. J., Conrads, T. P., Veenstra, T. D., *Mol. Cell Proteomics* 2003, *2*, 1096–1103.

[3] Adkins, J. N., Varnum, S. M., Auberry, K. J., Moore, R. J., Angell, N. H., Smith, R. D., Springer, D. L., Pounds, J. G., *Mol. Cell. Proteomics* 2002, *1*, 947–955.

[4] Pieper, R., Gatlin, C. L., Makusky, A. J., Russo, P. S., Schatz, C. R., Miller, S. S, Su, Q. *et al.*, *Proteomics* 2003, *3*, 1345–1364.

[5] Anderson, N. L., Polanski, M., Pieper, R., Gatlin, T., Tirumalai, R. S., Conrads, T. P., Veenstra, T. D. *et al.*, *Mol. Cell. Proteomics* 2004, *3*, 311–326.

[6] Chan, K. C., Lucas, D. A., Hise, D. *et al.*, *Clin. Proteomics* 2004, *1*, 101–225.

[7] Omenn, G. S., *Dis. Markers* 2004, *20*, 131–134.

[8] Omenn, G. S., *Proteomics* 2004, *4*, 1235–1240.

[9] Desiere, F., Deutsch, E. W., Nesvizhskii, A. I., Mallick, P., King, N., Eng, J., Aderem, A. *et al.*, *Genome Biol.* 2004, *6*, R9.

[10] Zhang, H., Yi, E. C., Li, X.-J., Mallick, P., Kelly-Spratt, K. S., Masselon, C. D., Camp, I. D. G. *et al.*, *Mol. Cell. Proteomics* 2005, *4*, 144–155.

[11] Eng, J., McCormack, A. L., Yates, J. R., *J. Am. Soc. Mass Spectrom.* 1994, *5*, 976–989.

[12] Keller, A., Nesvizhskii, A. I., Kolker, E., Aebersold, R., *Anal. Chem.* 2002, *74*, 5383–5392.

[13] Kersey, P. J., Duarte, J., Williams, A., Karavidopoulou, Y., Birney, E., Apweiler, R., *Proteomics* 2004, *4*, 1985–1988.

[14] Omenn, G. S., States, D. J., Adamski, M., Blackwell, T. W., Menon, R., Hermjakob, H., Apweiler, R. *et al.*, *Proteomics* 2005, *5*, DOI: 10.1002/pmic.2005–00358.

[15] Pedrioli, P. G. A., Eng, J. K., Hubley, R., Vogelzang, M., Deutsch, E. W., Pratt, B., Nilsson, E. *et al.*, *Nat. Biotechnol.* 2004 *22*, 1459–1466.

 