

# Challenges in deriving high-confidence protein identifications from data gathered by a HUPO plasma proteome collaborative study

David J States<sup>1</sup>, Gilbert S Omenn<sup>1</sup>, Thomas W Blackwell<sup>1</sup>, Damian Fermin<sup>1</sup>, Jimmy Eng<sup>2</sup>, David W Speicher<sup>3</sup> & Samir M Hanash<sup>1,2</sup>

The Human Proteome Organization (HUPO) recently completed the first large-scale collaborative study to characterize the human serum and plasma proteomes. The study was carried out in different locations and used diverse methods and instruments to compare and integrate tandem mass spectrometry (MS/MS) data on aliquots of pooled serum and plasma from healthy subjects. Liquid chromatography (LC)-MS/MS data sets from 18 laboratories were matched to the International Protein Index database, and an initial integration exercise resulted in 9,504 proteins identified with one or more peptides, and 3,020 proteins identified with two or more peptides. This article uses a rigorous statistical approach to take into account the length of coding regions in genes, and multiple hypothesis-testing techniques. On this basis, we now present a reduced set of 889 proteins identified with a confidence level of at least 95%. We also discuss the importance of such an integrated analysis in providing an accurate representation of a proteome as well as the value such data sets contain for the high-confidence identification of protein matches to novel exons, some of which may be localized in alternatively spliced forms of known plasma proteins and some in previously nonannotated gene sequences.

Proteomic technologies permit the extensive fractionation of proteins in biological specimens, analysis of peptides by MS and matching of mass spectra to peptide sequences in human genome and protein databases<sup>1-4</sup>. In practice, comprehensive proteomic analyses of specific fluids and tissues using these technologies may benefit from multi-laboratory, collaborative efforts because of the complexity of proteins and peptides that need to be analyzed. Different analytical platforms may target different

protein and peptide subsets and thus only integration of results from a diversity of platforms will give a full picture. In 2002, the international Human Proteome Organization (HUPO)<sup>5</sup> initiated a study of human serum and plasma proteins to evaluate the feasibility of such a collaborative approach to proteome analysis using diverse technologies applied by multiple laboratories to a common set of reference specimens (G.S.O. and collaborators<sup>6</sup>). The resulting publicly available data set (G.S.O. and collaborators<sup>7</sup>) is intended as an expandable resource for studies of serum and plasma to uncover variations in health and in disease, and for assessment of novel analytical tools.

## Deriving confidence measures for protein identifications

Eighteen laboratories, provided with aliquots of reference serum and plasma specimens, submitted 42,306 protein identifications from MS/MS using a variety of search engines and databases (Supplementary Notes and Supplementary Table 1 online). Lists with 18,098 distinct peptide sequences submitted constitute the 'Core Peptide Data set.' These peptides were matched to 15,710 entries (15,519 based on peptides with at least six amino acids) in the International Protein Index (IPI version 2.21, July 2003)<sup>8</sup>, which was chosen as the standard reference database for this project. An integration algorithm designed for this project<sup>9</sup> selected one representative protein among multiple proteins (isoforms and homologs) to which reported peptides gave 100% sequence matches. This integration process resulted in 9,504 proteins in the IPI database identified with one or more peptide sequences<sup>7</sup>. Subsets of 3,020 and 1,274 proteins were identified with two or more and three or more peptides, respectively. Protein lists are available at <http://www.bioinformatics.med.umich.edu/hupo/ppp> and <http://www.ebi.ac.uk/pride>.

A major issue in the analysis of MS data is assessment of the extent of false identifications. Several approaches have been used<sup>10-13</sup>, including probability-based evaluations of mass spectra<sup>14-16</sup>, reversed-sequence database searches<sup>17,18</sup> and Poisson analysis of the identifications by number of peptides matching<sup>9</sup>. Some of these approaches are available for one search engine type and only optimized for a small number of MS instruments<sup>14-16</sup>. The significance of peptide matches to proteins and the associated error rate also depend on the sequence length of the matched protein, given that longer sequences will present a greater

<sup>1</sup>University of Michigan, 100 Washtenaw Rd., Palmer Commons 2035B, Ann Arbor, Michigan 48109, USA. <sup>2</sup>Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N., PO Box 19024, Seattle, Washington 98109, USA. <sup>3</sup>The Wistar Institute, 3601 Spruce St., Philadelphia, Pennsylvania 19104, USA. Correspondence should be addressed to S.H. at [shanash@fhcrc.org](mailto:shanash@fhcrc.org).

Published online 8 March 2006; doi:10.1038/nbt1183

number of peptides to potentially match mass spectra. We applied a Poisson model to estimate the expected number of false matches. We postulated that a mass spectrum is derived from a given protein in the database and in addition, there may be a number of false matches with similar or higher scores occurring at random across the sequence database. The frequency of false peptide matches,  $\mu$ , is estimated using a Poisson model satisfying the constraint that the number of false matches predicted cannot exceed the number of total matches observed. In particular, if  $\mu > 0.00075$ , then the number of proteins containing a predicted false positive match will exceed the total number of protein matches observed project-wide. This value was chosen as an upper bound for  $\mu$ . The estimated number of false identifications increases as  $\mu$  increases so this choice of  $\mu$  yields a conservative limit for protein identification confidence. The mean number of matches  $\lambda$  expected at random for a protein of length  $L$  is  $\mu L$ . The protein length-specific probability,  $P_{rand}$ , that  $M$  or more matches will be observed is

$$P_{rand} = \sum_{i=M}^{\infty} \frac{\lambda^i}{i!} \exp(-\lambda)$$

We also take into account the size of the protein database used, which influences the probability of random/false matches. IPI version 2.21 consisted of 43,730 nonredundant entries, so the Bonferroni correction for multiple hypothesis testing divides the probability for a single match by this large number of entries. The expected number of matches,  $E$ , based on multiple hypothesis testing, is

$$E = N_{db} P_{rand}$$

where  $N_{db}$  is the number of sequences in the database. The confidence,  $C$ , that we have identified the one true database entry from which the spectral data were derived and not one of the  $E$  false positives is

$$C = \frac{1}{1 + E}$$

On the basis of this analysis, and after taking into consideration the issue of multiple hypothesis testing, we found that 889 of the set of 9,504 proteins have confidence at the 95% level (Fig. 1 and Supplementary Table 2 online). This is vastly more stringent than is the case if protein identifications were considered individually. The field 'expect\_1' gives the expected number of times a false-positive match with the same or more peptides would be found for any given protein, whereas 'expect\_db' gives the expected number of false positives in searching the full database (Supplementary Table 2 online). Therefore, it is likely that the 9,504-protein set with one or more peptide matches includes additional true positives that would reduce the rate,  $\mu$ , and correspondingly increase the confidence in other assignments.

Simulations were performed to determine the accrual of high-confidence identifications as a function of sampling (Supplementary Fig. 1 online). The number of identifications achieving statistical significance continued to rise linearly with the number of experimental observations throughout the simulation. Therefore, the number of high-confidence plasma protein identifications likely will continue to increase as additional experimental data are obtained.

To determine the extent to which additional proteins may be deduced from the original mass spectra with the use of uniform peptide and protein identification procedures, we carried out an independent analysis using PepMiner<sup>19</sup> and Sequest (J.E. and colleagues<sup>20</sup>), processed 5.6 million MS/MS spectra generated by participating laboratories and clustered them by their mutual similarity. A total of 2,895 proteins were identified with two or more peptides. The corresponding number for the same mass spectra is 2,868 proteins among the proteins in the 3,020-

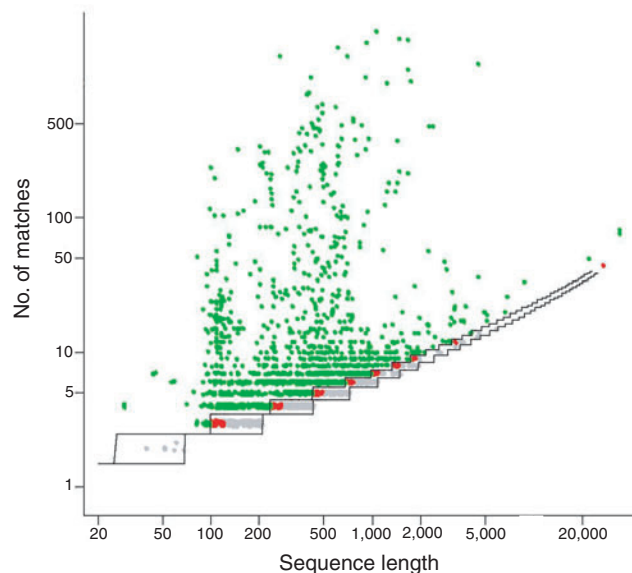
protein data set, with 1,051 proteins in common between the two data sets. An additional 700 of the 2,895 proteins occurred as single-peptide identifications in the set of 9,504 proteins. An independent analysis of the data using the X!Tandem search tool<sup>13</sup> identified 2,678 proteins with two or more distinct peptides. Of these, 218 are found in the 889-protein high-confidence set and 577 are in the set of 3,020 identifications. The striking differences in identifications with different search algorithms may result from differences in criteria for converting raw data to peak lists, differences in search parameters such as mass accuracy and the specific post-translational modifications considered during the database search.

### Characteristics of high-confidence protein identifications

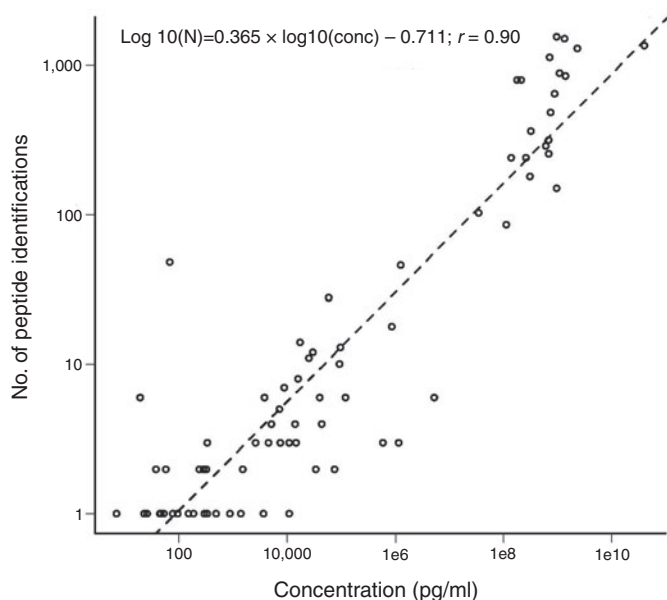
Quantitative immunoassays of the specimens used for MS were performed by four separate methods to evaluate the success in identifying proteins as a function of their abundance<sup>21</sup>. Among the 49 proteins matched to the 3,020-protein data set,  $\alpha$ -fetoprotein, CD30, Fas ligand, platelet-derived growth factor receptor  $\alpha$ , leukemia inhibitory factor receptor, stem cell receptor, matrix metalloproteinase 2/gelatinase, epidermal growth factor receptor, tissue inhibitor of metalloproteinase 1 (TIMP-1), insulin-like growth factor binding protein 2 (IGFBP-2), activated leukocyte adhesion molecule and selectin L were identified with measured concentrations from 200 pg/ml to 20 ng/ml. We observed a linear correlation between the logarithm of the number of peptides detected ( $N$ ) and the logarithm of concentration ( $C$ ) of a given protein (Fig. 2):

$$\log_{10}(N) = 0.365 \times \log_{10}(C) - 0.711$$

The correlation coefficient for this regression is  $r = 0.86$  for the 3,020-protein data set. Thus, the above equation can be used to roughly estimate



**Figure 1** Distribution of protein identifications. Protein length in amino acids is plotted versus number of peptide spectra matches to delineate low-confidence from high-confidence identifications after adjustment for multiple hypotheses testing of matching to the 43,370 entries in the IPI version 2.21 database. 889 identifications achieved high confidence (green dots >0.95) and 397 identifications had lower confidence (red dots 0.9–0.95 and gray dots 0.5–0.9); the remaining 8,218 reported identifications with less than 50% confidence by stringent criteria are not shown.



**Figure 2** Number of peptides identified as a function of protein concentration. Relationship between aggregate number of peptides identified by MS/MS and measured concentration of 76 proteins by quantitative immunoassays is plotted. The straight line shows a linear regression fit to the data ( $\log_{10}(N) = 0.365 \times \log_{10}(\text{conc}) - 0.711$ ), with correlation coefficient 0.90. The striking outlier at 67 pg/ml with 39 identifications is mucin-16.

protein concentration in the project data set, even when quantitative immunoassay data are not available, as has been recently reported<sup>22</sup>.

A striking anomaly is mucin 16 (also designated CA-125), a large protein with 22,152 amino acids and 2.35 MDa before glycosylation<sup>23</sup>, measured by immunoassay at 67 pg/ml, yet identified by multiple laboratories, with as many as 15 peptides by two laboratories. Identified peptides were spread throughout the length of the protein. Mucin 16 remained in the high-confidence subset, even after taking into account the size of the protein. The discrepancy between measured concentration by immunoassay and peptide identification by MS may be related to the fact that the two laboratories that identified a large set of peptides for Mucin 16 enriched for glycoproteins before analysis.

Some 817 of the 889 proteins can be mapped to SwissProt through database cross references, and 244 of these proteins are annotated as having signal peptides in SwissProt release 45, although many of these are qualified as 'predicted' or 'possible'. SignalP 3.0 was applied to the set of 889 proteins to search for secreted signal peptides using the neural network algorithm and a hidden Markov model<sup>24</sup>. A subset of 271 of the 889 proteins was predicted to have a signal peptide by both models. Approximately half of the 889 proteins was not previously described as occurring in serum or plasma based on literature searches and queries of publicly available databases. Their corresponding genes were expressed in a wide variety of tissues. Surprisingly, a few are known to be expressed predominantly in the central nervous system, including G-protein receptor 85 (GPR85), human progenitor cell antigen (HPCA),  $\alpha$ 1,3-fucosyltransferase IX (FUT9), glial fibrillary acidic protein (GFAP) and inwardly rectifying  $K^+$  channel 2 (IRK 2) (**Supplementary Table 2** online).

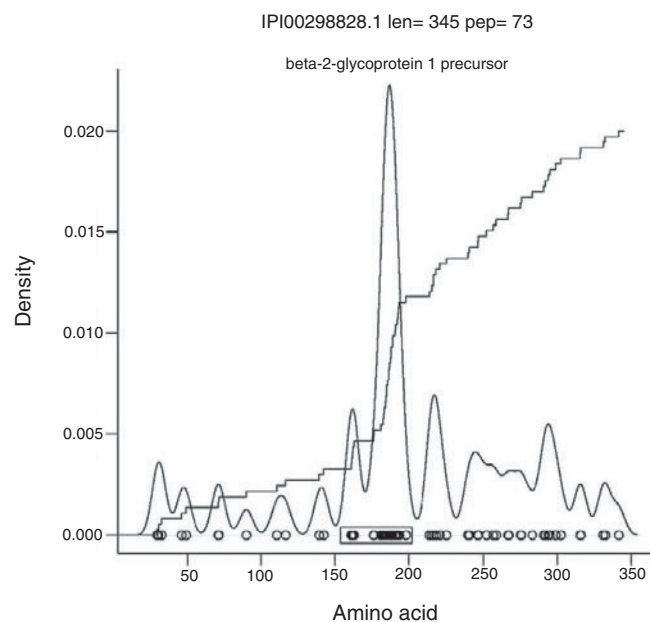
To test whether the species detected among the set of 889 proteins were full-length molecules or fragments, we examined clustering of peptide matches along the length of the primary translation product for a group of 97 proteins with more than 25 distinct peptide matches (further described in **Supplementary Notes** online). Highly significant clustering of peptide matches was observed for 11 proteins in this group. This find-

ing is consistent with the presence of circulating fragments that are more abundant than the full-length protein, as illustrated for  $\beta$ -2-glycoprotein 1 (**Fig. 3**), a plasma protein that interacts with several other proteins and with epitopes that are the targets of autoantibodies in the antiphospholipid syndrome<sup>25</sup>. A large excess of peptide matches is observed in the region from amino acids 174 to 200 (**Fig. 3**). The molecule is known to be cleaved biologically, but the high abundance of fragments containing the 174 to 200 interval has not previously been described.

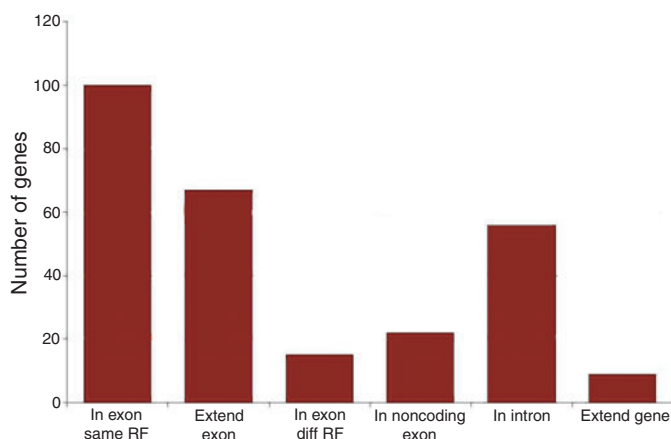
### Specimen and technology platform variables

The sets of four specimens (serum and plasma anticoagulated with EDTA, citrate or heparin) yielded rather similar numbers of proteins when analyzed by the same techniques (**Supplementary Table 1** online). The greatest yield of identified proteins by a single laboratory with a single specimen consisted of 1,168 proteins, 427 of which were confirmed in at least one other laboratory. The approach used by this laboratory (**Supplementary Table 1** online, lab 34, B1-serum) comprised depletion of six abundant proteins, extensive fractionation and the use of a ThermoFinnigan LTQ linear ion trap mass spectrometer. It was particularly interesting that the LTQ-based serum analysis results included multiple peptide identifications for 57 proteins identified with only one peptide using a very similar separation and analysis strategy for a plasma sample using an LCQ XP+ ion trap mass spectrometer. Low-abundance proteins that were confirmed with multiple peptide matches using the more sensitive LTQ included at least four proteins known to be present in the serum sample at low ng/ml levels (vascular/endothelial cadherin, insulin-like growth factor binding protein 3, TIMP1 and lymphocyte adhesion molecule 1)<sup>26</sup>. Analyses of the additional serum samples (B2, B3 and C1) contributed 604 nonduplicated proteins to the overall serum list (**Supplementary Table 1** online).

Separations of intact proteins uncovers protein variation resulting from alternative splicing, cleavage and other types of post-transla-



**Figure 3** Distribution of peptides identified for  $\beta$ -2-glycoprotein 1. Shown in the plot are the location of the distinct peptide matches along the x axis, a line showing the smooth density of matches across the protein, a second line showing the cumulative number of matches and a box showing the maximally enriched segment. All calculations were performed using the statistical package "R" (<http://www.r-project.org/>).



**Figure 4** Bar plot of the distribution of ORF types by gene. The most common category were genes containing ORFs that fall within previously annotated exons and are translated in the same reading frame, confirming the genome annotation. For 67 genes, we identified novel ORFs containing peptides that overlapped and extended a previously annotated coding exon; for 15 genes, novel ORFs where the peptides were not translated in the annotated reading frame; for 22 genes, ORFs with peptides from exons annotated as noncoding; for 56 genes, novel ORFs that fell entirely within annotated introns; and for 9 genes, ORFs that overlapped the annotated gene boundary but contained peptides that fell outside of the annotated gene boundaries.

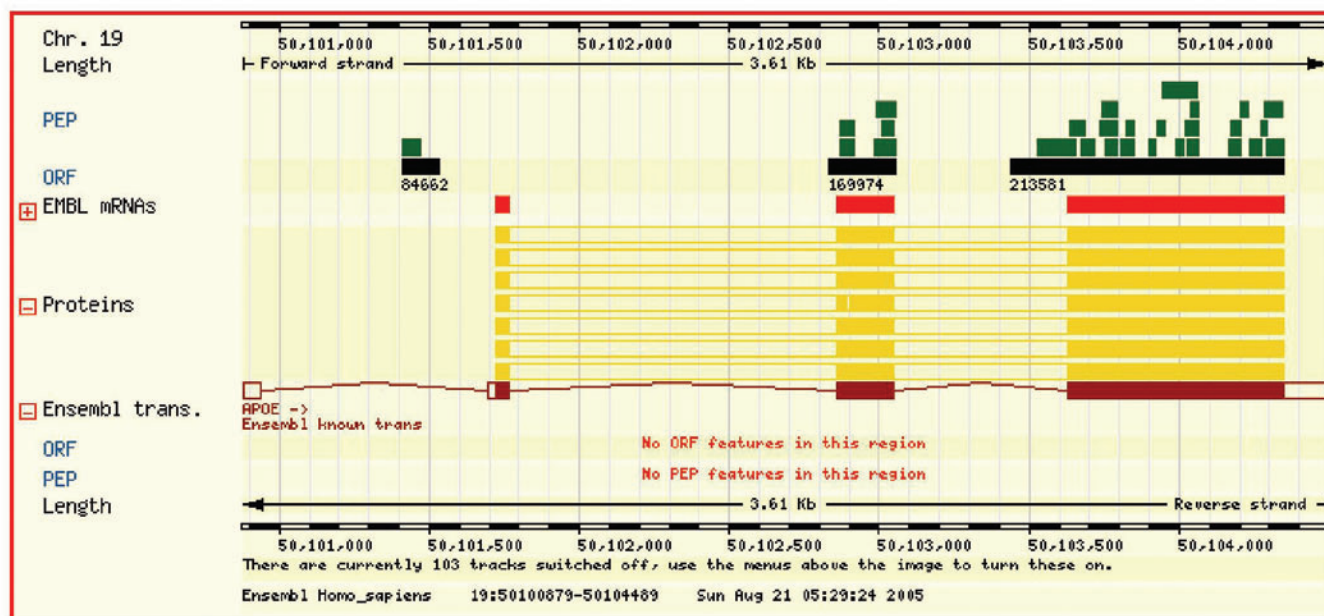
tional modification. A sampling of proteins that exhibited quantitative differences between serum and plasma, after separation of intact proteins based on charge, hydrophobicity and molecular weight, coupled with cyanine (Cy) dye labeling of serum (via Cy5), EDTA-plasma (via Cy3) and citrate-plasma (via Cy2)<sup>27</sup>, yielded 13 identified proteins that varied in band intensity by more than 100-fold. As expected, several differences in band intensities between serum and plasma were attribut-

able to proteins related to coagulation (fibrinogens  $\alpha$ ,  $\beta$  and  $\gamma$  as well as vitronectin)<sup>28</sup>. Interestingly, many proteins were identified in multiple fractions, exhibiting different pI, hydrophobicity or molecular mass values that reflected the expression of different isoforms or specific protein cleavage products, as illustrated by complement component 3 (Supplementary Fig. 2 online).

#### Novel protein coding sequences

Peptides may be identified by using MS/MS data to search whole genome sequences<sup>29,30</sup>. We searched for matches between mass spectra and the two-strand, three-reading frame translation of the National Center for Biotechnology Information (NCBI, Bethesda, MD) human genome sequence build 33, to uncover novel coding sequences. We then tested any matches against the NCBI nonredundant (NR) data set. The open source tool X!Tandem<sup>13</sup> was used to reanalyze 11 data sets. To accept a peptide identification, we required matches in at least two tandem mass spectra with each match having an X!Tandem hyperscore of 35 and representing the highest scoring match for the MS/MS spectrum. Furthermore, we required that the peptide have a unique location in the genome.

This procedure identified matches to 445 open reading frames (ORFs) within 180 known ENSEMBL-annotated genes (Fig. 4 and Supplementary Table 3 online) with ORF identification confidence > 95%, even after multiple hypothesis corrections for searching over 200 million ORFs. The largest group of novel ORFs were found within the boundaries of genes encoding known plasma proteins, consistent with these novel ORFs being novel, alternative-splice isoforms. Their presence in known plasma protein-encoding genes further enhances the confidence in their identification, as no constraints to blood proteins were imposed on the search. In addition, 99 genes contained ORFs that fall within annotated exons, thus confirming the genome annotation; the limited number of identifications in this class reflects the very stringent requirements used for accepting a match in this very large target database. Fifteen genes contained ORFs that fall within an annotated coding



**Figure 5** Novel ORFs in the *APOE* gene. A display generated by linking our novel ORF database to the ENSEMBL genome browser using distributed annotation system (DAS) tracks is shown with the location of novel ORFs (black rectangles) and peptides (dark green rectangles) in the *APOE* gene, demonstrating the presence of novel circulation splice isoforms of this protein in blood. These displays are hyperlinked to a public server providing peptide information, peptide sequence, accession numbers, laboratory identifier, sequence identifier and identification confidence.

exon but are translated in a different reading frame. Twenty-one genes contained ORFs that matched exons annotated as noncoding; 67 genes contained peptides in ORFs that overlap within an annotated exon; and 56 genes contained ORFs located in regions annotated as introns. Eight ORFs overlapped the annotated boundaries of a gene, consistent with the presence of a novel splice expanding the gene boundaries. The sum of these counts exceeds the total because, in many cases, genes contained both ORFs confirming annotated exons and additional ORFs not matching annotated exons.

The fact that we find evidence for additional splice isoforms in most of these well-known and well-studied genes is striking. Examples include a novel ORF and alternatively spliced forms in the apolipoprotein E (APO-E) gene and a circulating splice isoform of WNT10b (wingless related mouse mammary tumor virus integration site 10b) (Fig. 5). Our whole-genome ORF search also identified 218 putative novel ORFs that did not fall within any annotated gene in the human genome (Supplementary Table 3 online). Genscan finds gene models for 61% of these ORFs. Blast searches of these ORFs demonstrate similarity to a number of known plasma proteins, including immunoglobulins and haptoglobin. Two ORFs show strong similarity to the Alu transposon. These findings are consistent with identification of novel splice isoforms; alternative splicing as a result of Alu transposition has been previously described<sup>31</sup>.

## DISCUSSION

A data set of proteins identified in serum and plasma specimens has been developed through the collaborative HUPO Plasma Proteome Project. After a first round of analysis<sup>7</sup>, it became clear that to determine true-positive protein identifications, it would be necessary to take into account the probability of random matches as a function of protein length and to adjust for the number of entries in the database searched for protein assignments. We have developed several different subsets of proteins from the integrated data, including subsets of increased stringency. A major challenge in the analysis of complex peptide mixtures by MS is the inherently incomplete sampling of peptide ions in the mass spectrometer<sup>32–35</sup> with a bias toward acquisition of more abundant peptide ions. Because of this sampling process, repeat analyses of the same complex mixture on the same instrument led to overlapping but nonidentical sets of data<sup>35</sup>.

Through a literature review combined with analysis of three experimental data sets<sup>36</sup>, we compiled a list of 1,175 proteins, only 46 of which were in all four sources; 990 of these proteins have IPI version 2.21 identifiers, of which 316 were found among the 3,020-protein IPI protein set with two peptide matches, and 471 matched the 9,504-protein data set. Shen *et al.*<sup>17</sup> reported that 800 or 1,682 proteins could be identified in a human serum specimen, depending on the filtering criteria<sup>17</sup>. These authors provided us with the data needed to reanalyze their findings using our statistical criteria. Reanalysis using the Sequest filter criteria and the application algorithm used in the HUPO study, yielded 1,073 proteins. Then, the length-dependent statistical analysis yielded 433 proteins with confidence >95%. Of these, 179 are in the 889-protein subset of highly confident identifications presented in this study. Another study<sup>37</sup> reported 1,444 proteins in serum, of which we mapped 1,019 with IPI identifiers; of these, 257 proteins matched with the 3,020-protein data set. Zhou *et al.*<sup>38</sup> have identified an aggregate of 210 low-molecular weight proteins or peptides after multiple immunoprecipitation steps with antibodies against 6 abundant proteins; of 148 proteins mapped to IPI, 62 were found in the 3,020-protein list.

Thus, across studies, partial sampling of peptides and variable identification of proteins is a dominant, but not surprising, feature, given the

heterogeneity in the approaches applied and the sampling nature of MS analysis. Nevertheless, a substantial depth of analysis has been achieved with combinations of depletion of highly abundant proteins, fractionation of intact proteins followed by digestion to peptides, fractionation of peptides and two or more MS/MS runs for each fraction.

Annotation of ORFs in genomes is most often performed computationally based on features in the DNA sequence. Proteomic data as obtained in this study provide a complementary method to annotate the genome<sup>39,40</sup>. We identified peptide matches for genes and sequences within genes not previously known to have protein products, as well as for splice isoforms and for known exons of well-annotated genes, including some encoding plasma proteins previously extensively characterized.

It is surprising that the number of protein identifications between the two systematic independent comparisons that were made showed <50% agreement between the two data sets. Therefore, it is prudent to apply multiple search tools to allow protein identifications with high confidence and to fully explore the repertoire of potential identifications in a given biological sample. There is also a clear need for improved computational tools for reliable analysis of large MS/MS data sets and improved standards for data analysis. The availability of such improved tools, together with improved MS instrumentation with increased scan speed, larger dynamic range and greater sensitivity should contribute substantially to improved mining of the human proteome.

*Note: Supplementary information is available on the Nature Biotechnology website.*

## ACKNOWLEDGMENTS

The collaborative HUPO Plasma Protein study and the data analysis presented here have been supported by a trans-National Institutes of Health grant supplement 84982 administered by the National Cancer Institute, by pharmaceutical and technology company sponsors and by voluntary efforts of collaborating laboratories.

## COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at <http://www.nature.com/naturebiotechnology/>

Reprints and permissions information is available online at <http://npg.nature.com/reprintsandpermissions/>

- Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003).
- Sadygov, R., Cociorva, D. & Yates, J.R. Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nat. Methods* **1**, 195–202 (2004).
- Olsen, J. & Mann, M. Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. *Proc. Natl. Acad. Sci. USA* **101**, 13417–13422 (2004).
- Orchard, S., Hermjakob, H. & Apweiler, R. Annotating the human proteome. *Mol. Cell. Proteomics* **4**, 435–440 (2005).
- Hanash, S. & Celis, J.E. The human proteome organization: a mission to advance proteome knowledge. *Mol. Cell. Proteomics* **1**, 413–414 (2002).
- Omenn, G.S. The Human Proteome Organization plasma proteome project pilot phase: reference specimens, technology platform comparisons, and standardized data submissions and analyses. *Proteomics* **4**, 1235–1240 (2004).
- Omenn, G.S. *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* **5**, 3226–3245 (2005).
- Kersey, P. *et al.* The International Protein Index: an integrated database for proteomics experiments. *Proteomics* **4**, 1985–1988 (2004).
- Adamski, M. *et al.* Data management and preliminary data analysis in the pilot phase of the HUPO Plasma Proteome Project. *Proteomics* **5**, 3246–3261 (2005).
- Carr, S. *et al.* The need for guidelines in publication of peptide and protein identification data. *Mol. Cell. Proteomics* **3**, 531–533 (2004).
- Cargile, B.J., Bundy, J.L. & Stephenson, J.L. Potential for false positive identifications from large databases through tandem mass spectrometry. *J. Proteome Res.* **3**, 1082–1085 (2004).
- Eriksson, J. & Fenyo, D. Protein identification in complex mixtures. *J. Proteome Res.* **4**, 387–393 (2005).
- Fenyo, D. & Beavis, R.C. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem.* **75**, 768–774 (2003).

14. Keller, A., Nesvizhskii, A.I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74**, 5383–5392 (2002).
15. Nesvizhskii, A.I., Keller, A., Kolker, E. & Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **75**, 4646–4658 (2003).
16. Sadygov, R.G. & Yates, J.R. A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. *Anal. Chem.* **75**, 3792–3798 (2003).
17. Shen, Y. *et al.* Ultra-high-efficiency strong cation exchange LC/RPLC/MS/MS for high dynamic range characterization of the human plasma proteome. *Anal. Chem.* **76**, 1134–1144 (2004).
18. Perkins, D.N., Pappin, D.J., Creasy, D.M. & Cottrell, J.S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567 (1999).
19. Beer, I., Barnea, E., Ziv, T. & Admon, A. Improving large-scale proteomics by clustering of mass spectrometry data. *Proteomics* **4**, 950–960 (2004).
20. Eng, J.K., McCormack, A.L. & Yates, J.R.I. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994).
21. Haab, B.B. *et al.* Immunoassay and antibody microarray analysis of the HUPO reference specimens: systematic variation between sample types and calibration of mass spectrometry data. *Proteomics* **5**, 3278–3291 (2005).
22. Ishihama, Y. *et al.* Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol. Cell. Proteomics* **4**, 1265–1272 (2005).
23. O'Brien, T.J. *et al.* The CA 125 gene: an extracellular superstructure dominated by repeat sequences. *Tumour Biol.* **22**, 348–366 (2001).
24. Bendtsen, J.D., Nielsen, H., vonHeijne, G. & Brunak, S. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* **340**, 783–795 (2004).
25. Miyakis, S., Giannakopoulos, B. & Krilis, S.A. Beta 2 glycoprotein I-function in health and disease. *Thromb. Res.* **114**, 335–346 (2004).
26. Tang, H.Y. *et al.* A novel four-dimensional strategy combining protein and peptide separation methods enables detection of low-abundance proteins in human plasma and serum proteomes. *Proteomics* **5**, 3329–3342 (2005).
27. Wang, H. *et al.* Intact-protein based high-resolution three-dimensional quantitative analysis system for proteome profiling of biological fluids. *Mol. Cell. Proteomics* **4**, 618–625 (2005).
28. Misek, D.E. *et al.* A wide range of protein isoforms in serum and plasma uncovered by a quantitative Intact Protein Analysis System (IPAS). *Proteomics* **5**, 3343–3351 (2005).
29. Choudhary, J.S., Blackstock, W.P., Creasy, D.M. & Cottrell, J.S. Interrogating the human genome using uninterpreted mass spectrometry data. *Proteomics* **1**, 651–667 (2001).
30. Kuster, B., Mortensen, P., Andersen, J.S. & Mann, M. Mass spectrometry allows direct identification of proteins in large genomes. *Proteomics* **1**, 641–650 (2001).
31. Kneahling, J. & Graveley, B.R. The origins and implications of Alternative splicing. *Trends Genet.* **20**, 1–4 (2004).
32. Link, A.J. *et al.* Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **17**, 676–682 (1999).
33. Liu, H., Sadygov, R.G. & Yates, J.R. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* **76**, 4193–4201 (2004).
34. Washburn, M.P., Wolters, D. & Yates, J.R. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19**, 242–247 (2001).
35. Ghaemmaghami, S. *et al.* Global analysis of protein expression in yeast. *Nature* **425**, 737–741 (2003).
36. Anderson, N.L. *et al.* The human plasma proteome: a nonredundant list developed by combination of four separate sources. *Mol. Cell. Proteomics* **3**, 311–316 (2004).
37. Chan, K.C. *et al.* Analysis of the human serum proteome. *Clin. Proteomics* **1**, 101–225 (2004).
38. Zhou, M. *et al.* An investigation in the human serum “interactome”. *Electrophoresis* **25**, 1289–1298 (2004).
39. Jaffe, J.D., Berg, H.C. & Church, G.M. Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics* **4**, 59–77 (2004).
40. Oyama, M. *et al.* Analysis of small human proteins reveals the translation of upstream open reading frames of mRNAs. *Genome Res.* **14**, 2048–2052 (2004).