A Dataset of Human Liver Proteins Identified by Protein Profiling Via Isotope-coded Affinity Tag (ICAT) and Tandem Mass Spectrometry*[S]

Wei Yan‡§, Hookeun Lee‡, Eric W. Deutsch‡, Catherine A. Lazaro¶, Weiliang Tang¶, Eric Chen‖, Nelson Fausto¶, Michael G. Katze‖, and Ruedi Aebersold‡

Proteins from human liver carcinoma Huh7 cells, representing transformed liver cells, and cultured primary human fetal hepatocytes (HFH) and human HH4 hepatocytes, representing nontransformed liver cells, were extracted and processed for proteome analysis. Proteins from stimulated cells (interferon-α treatment for the Huh7 and HFH cells and induction of hepatitis C virus [HCV] proteins for the HH4 cells) and corresponding control cells were labeled with light and heavy cleavable ICAT reagents, respectively. The labeled samples were combined, trypsinized, and subject to cation-exchange and avidin-affinity chromatographies. The resulting cysteine-containing peptides were analyzed by microcapillary LC-MS/MS. The MS/MS spectra were initially analyzed by searching the human International Protein Index database using the SEQUEST™ software (1). Subsequently, new statistical algorithms were applied to the collective SEQUEST search results of each experiment. First, the PeptideProphet™ software (2) was applied to discriminate true assignments of MS/MS spectra to peptide sequences from false assignments, to assign a probability value for each identified peptide, and to compute the sensitivity and error rate for the assignment of spectra to sequences in each experiment. Second, the ProteinProphet™ software (3) was used to infer the protein identifications and to compute probabilities that a protein had been correctly identified, based on the available peptide sequence evidence. The resulting protein lists were filtered by a ProteinProphet probability score $p \geq 0.5$, which corresponded to an error rate of less than 5%. A total of 1,296, 1,430, and 1,476 proteins or related protein groups were identified in three subdatasets from the Huh7, HFH, and HH4 cells, respectively. In total, these subdatasets contained 2,486 unique protein identifications from human liver cells. An increase of the threshold to $p \geq 0.9$ (corresponding to an error rate of less than 1%) resulted in 2,159 unique protein identifications (1,146, 1,235, and 1,318 for the Huh7, HFH, and HH4 cells, respectively). ***Molecular & Cellular Proteomics 3:1039–1041, 2004.***

From the ‡Institute for Systems Biology, Seattle, WA; ¶Department of Pathology, University of Washington, Seattle, WA; and ‖Department of Microbiology, University of Washington, Seattle, WA

Received, June 14, 2004

Published, MCP Papers in Press, July 21, 2004, DOI 10.1074/mcp.D400001-MCP200

This human liver proteomic dataset consists of three subdatasets generated from three protein profiling experiments using the following samples: human liver carcinoma cells (Huh7), primary cultures of human fetal hepatocytes (HFH)[1] (4), and an immortalized cell line derived from human fetal hepatocytes (HH4).[2]

The Huh7 and HFH cells were selected to study the interferon response in transformed (Huh7) and nontransformed (HFH) human liver cells, respectively. About 2×10^7 cells were either interferon-α_{2b}- (400 IU/ml Intron-A; Schering-Plough Co., Kenilworth, NJ) or mock- treated for 16 h before harvest. The cells were lysed, and cell lysates were fractionated into cytosolic, membrane, and nuclear fractions by sequential differential centrifugation at 3,000 $\times$ *g* (nuclear fraction from the pellet) and 100,000 $\times$ *g* (cytosolic fraction from the supernatant and membrane fraction from the pellet). Proteins from each subcellular fraction were labeled with isotopically light- (^{12}C, for stimulated cells) or heavy- (^{13}C, for control cells) ICAT reagents following the manufacturer's protocol (Applied Biosystems, Foster City, CA). Corresponding isotopically light- and heavy-labeled samples were then combined and digested with trypsin (Promega, Madison, WI). The resulting peptides were separated by strong cation exchange chromatography, as previously described (5), and affinity purified by avidin cartridges following the manufacturer's protocol (Applied Biosystems), through which the cysteine (Cys)-containing peptides were enriched. The Cys-containing peptides from ~20 fractions purified above were then subjected to μLC-ESI-MS/MS using an LCQ-DECA-XP ion-trap mass spectrometer (ThermoFinnigan, San Jose, CA) as previously described (6, 7).

All observed MS/MS spectra were subsequently subjected to search against the human International Protein Index (IPI) database (www.ebi.ac.uk/IPI/IPIhelp.html) (v2.28) using the SEQUEST™ software. Search parameters for the cleavable ICAT-labeled samples used in this study were the following: +227.13 Da for static modification on cysteine residues labeled with cleavable ICAT, +9 Da for ^{13}C isotopic ICAT-

[1] The abbreviations used are: HFH, human fetal hepatocytes; Cys, cysteine; MS/MS, tandem mass spectrometry; IPI, International Protein Index; HCV, hepatitis C virus; ponA, ponasterone A.

[2] W. Tang and N. Fausto, personal communication.

labeled cysteine, +16 Da for oxidized methionine; mass tolerance ± 3 Da; restriction on Cys-containing peptides; and no proteolytic enzyme specified. Accuracy of the SEQUEST assignments of MS/MS spectra to peptide sequences was estimated by the PeptideProphet™ software based on a statistical model (2). For each identified peptide, a probability score was computed on a scale of 0 (for "incorrect") to 1 (for "correct") based on match of the peptide sequence to the tandem mass spectra and the trypsin proteolytic pattern. These assigned peptides were then subjected to ProteinProphet™ analysis to assign a protein probability score for each identified protein or related protein group inferred from the peptide data (3). The protein probabilities, again on a scale of 0 to 1, discriminate correct ($p = 1$) from incorrect ($p = 0$) protein identifications. Validation of initial data base search results on the basis of statistical modeling allows the presentation of large-scale proteomics datasets with known sensitivity for positive identifications and error rates for false positive identifications.

In the Huh7 cells, 23,310 peptides, with a PeptideProphet probability score $p \geq 0.05$, were obtained and included for subsequent ProteinProphet analysis. The sequences of the assigned peptides, together with their IPI reference name, PeptideProphet probability, and calculated and measured mass, are presented as reference for future proteomics studies (Supplemental Table I*a*). Using ProteinProphet software, 1,146 proteins or related protein groups were identified with an arbitrary probability cut-off of $p \geq 0.9$ (Supplemental Table II*a*). This value corresponded to 87.5% sensitivity (*i.e.* 87.5% of all possible identifications were made) and a false positive error rate of 0.7% (Supplemental Fig. 1*A*). This type of analysis also allows the investigator to compute the implications of changing the probability value on sensitivity and false positive error rate. For example, a reduction of the protein probability from 0.9 to 0.5 in this subdataset increased the number of protein identifications to 1,296, increased the sensitivity to 95.7%, and also increased the error rate to 3.8%.

Similarly, a total of 31,641 peptides ($p \geq 0.05$) were obtained from the analysis of the HFH cells (Supplemental Table I*b*). These assigned peptides were used for subsequent ProteinProphet analysis and lead to the identification of 1,235 proteins and related protein groups ($p \geq 0.9$) (Supplemental Table II*b*), which corresponded to 86.5% sensitivity and 0.8% error rate (Supplemental Fig. 1*B*). A reduction of protein probability threshold to 0.5 resulted in 1,430 protein identifications with 96% sensitivity and 4.6% error rate.

The HH4 cells are immortalized human hepatocytes. Two HH4-based cell lines were constructed, based on an ecdysone-inducible expression system (8), to induce expression of the entire hepatitis C virus (HCV) ORF or green fluorescence protein, respectively.[3] The ecdysone-regulated gene expression system consists of a modified ecdysone receptor (a heterodimer of VgEcR and RXR) that binds to its recognition sequence (5xE/GRE) and associates with transcription corepressors to repress the downstream promoter. Upon induction by a plant-derived ecdysone analog ponasterone A (ponA), ponA binds to the VgEcR to release the corepressors and recruit cotransactivators to activate transcription of the downstream target genes. We performed two ICAT labeling experiments to investigate HCV-mediated protein expression profiles in human hepatocytes. The first compared total cell extracts from HCV ORF-induced cells (ponA+, light-labeled) with noninduced cells (ponA-, heavy-labeled). The second experiment compared total cell extracts from cells with induced HCV proteins (light-labeled) with cells carrying induced green fluorescence protein (heavy-labeled). The labeled samples were subjected to the same analyses as described above. From HH4 cells we obtained a total of 28,029 peptide

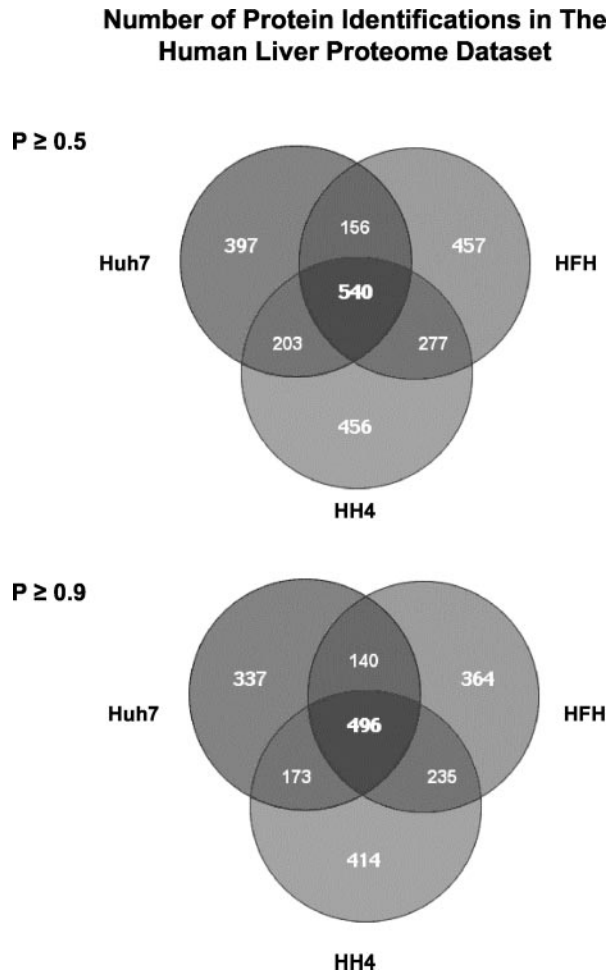

FIG. 1. **Comparison of the protein identifications of the three subdatasets from the Huh7, HFH, and HH4 cells.** Identified protein or related protein groups from the three proteomics subdatasets (Supplemental Table II*a* for Huh7, II*b* for HFH, and II*c* for HH4) were compared based on the IPI identification of each entry. The results were displayed by Venn Diagram (www.venndiagram.com) at the protein probability threshold of 0.9 and 0.5, respectively.

[3] W. Tang and N. Fausto, unpublished data.

assignments with $p \geq 0.05$ (Supplemental Table I*c*), which contributed to identification of 1,318 ($p \geq 0.9$) or 1,476 proteins and related protein groups ($p \geq 0.5$) (Supplemental Table II*c* and Supplemental Fig. 1*C*).

Taken together, proteomics analyses of the three human liver cells of both transformed and nontransformed cells lead to a total of 2,159 ($p \geq 0.9$) or 2,486 ($p \geq 0.5$) unique protein identifications. Among them, 496 ($p \geq 0.9$) or 540 ($p \geq 0.5$) were found in all three liver cells, while 337/397, 364/457, and 414/456 proteins and related protein groups were uniquely observed in the Huh7, HFH, and HH4 cells, respectively ($p \geq 0.9/p \geq 0.5$). Comparison of the three proteomics subdatasets from human liver cells (Supplemental Table III*a* for $p \geq 0.9$ and III*b* for $p \geq 0.5$) are also shown as a Venn Diagram (Fig. 1) using the on-line Create-A-Venn system at www.venndiagram.com. This human liver proteomics datasets with more than 2,000 protein identifications, presented in a statistically validated and transparent way, describes a possible mechanism for publishing large-scale protein identification datasets in the literature and for data comparison from different experiments.

* This work was supported in part by grants from the National Heart, Lung, and Blood Institute Proteomics Center at the Institute for Systems Biology (N01-HV-28179) and the National Institute on Drug Abuse (1P30DA01562501).The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

S The on-line version of this manuscript (available at http://www.mcponline.org) contains supplemental material.

§ To whom correspondence should be addressed: Institute for Systems Biology, 1441 N. 34th St., Seattle, WA 98103. Tel.: 206-732-1305; Fax: 206-732-1299; E-mail: wyan@systemsbiology.org.

REFERENCES

1. Eng, J., McCormack, A., and Yates, J. (1994) 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
2. Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74,** 5383–5392
3. Nesvizhskii, A. I., Keller, A., Kolker, E., and Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. Establishment, characterization, and long-term maintenance of cultures of human fetal hepatocytes. *Anal. Chem.* **75,** 4646–4658
4. Lazaro, C. A., Croager, E. J., Mitchell, C., Campbell, J. S., Yu, C., Foraker, J., Rhim, J. A., Yeoh, G. C., and Fausto, N. (2003) *Hepatology* **38,** 1095–1106
5. Han, D. K., Eng, J., Zhou, H., and Aebersold, R. (2001) Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nat. Biotechnol.* **19,** 946–951
6. Lee, H., Yi, E. C., Wen, B., Reily, T. P., Pohl, L., Nelson, S., Aebersold, R., and Goodlett, D. R. (2004) Optimization of reversed-phase microcapillary liquid chromatography for quantitative proteomics. *J. Chromatogr. B. Analyt. Technol. Biomed. Life Sci.* **803,** 101–110
7. Von Haller, P. D., Yi, E., Donohoe, S., Vaughn, K., Keller, A., Nesvizhskii, A. I., Eng, J., Li, X. J., Goodlett, D. R., Aebersold, R., and Watts, J. D. (2003) The application of new software tools to quantitative protein profiling via isotope-coded affinity tag (ICAT) and tandem mass spectrometry: II. Evaluation of tandem mass spectrometry methodologies for large-scale protein analysis, and the application of statistical tools for data analysis and interpretation. *Mol. Cell. Proteomics* **2,** 428–442
8. No, D., Yao, T. P., and Evans, R. M. (1996) Ecdysone-inducible gene expression in mammalian cells and transgenic mice. *Proc. Natl. Acad. Sci. U. S. A.* **93,** 3346–3351