

Report from Human Liver Proteome Project

Activity Report of HUPO Human Liver Proteome Project (HLPP)

(July 1st, 2009 to August 30th, 2010)

Co-chairs: Laura Beretta, Fuchu He, José Mato

Workshop in the Reporting Period

The 11th HLPP Workshop was held on September 26th, 2009 during the HUPO 8th Congress in Toronto, Canada. More than fifty participants from fifteen countries attended the workshop which was organized around two sessions.

In Session One, five groups presented their individual progress reports. Dr. Fuchu He (BPRC, China) talked about the systematic analysis of liver proteome on Chinese human liver organelles. They identified 6311 proteins from liver organelles with high-confidence. They also predicted the localization of 1842 proteins in liver by bioinformatic approach and then verified them by biological experiments, such as GFP tag and image analysis. Further more, they annotated the function of liver organelle proteome systematically by modularized, including 151 novel proteins presented in liver. After Dr. He, Dr. Laura Beretta (FHRC, USA) reported the analysis on the profile of human adult liver reference sample and mouse model of liver cancer performed by her group. A total of 474,617 spectra (23,886 unique peptides) were identified with high confidence and corresponded to over 6,000 proteins obtained from the French Liver Reference Sample. Dr. Felix Elortza (CIC bioGUNE, Spain) introduced the progress on screening non-invasive biomarkers on liver diseases at CIC bioGUNE. They identified three proteins from serum which were correlated with the disease state in serum. Based on their screening result, they developed NAFLD animal models to explore the biological function of these molecules, including MAT1A deletion (inducing steatosis, NASH, and HCC) and GNMT deletion (resulting in steatosis, fibrosis, and HCC). Dr. Young-Ki Paik from Yonsei University (Korea) also introduced a simple method for quantitative analysis of N-linked glycoproteins in hepatocellular carcinoma specimens. The EPL-based approach could be used in the qualitative and quantitative analysis of N-linked glycopeptides without enrichment, rather than the detection of low abundance glycoproteins. At the end, Dr. Pengyuan Yang from Fudan University in China reported the

progress on human liver glyco-proteome and acetylo-proteome research. They identified 1637 glycosites with 853 non-redundant glycosylated proteins and 1200 acetylated peptides with 978 non-redundant acetylated proteins from the Chinese human liver sample. In session Two, all the participants discussed on four topics, focusing on “Technology strategies”, “Data Control, integration and comparison”, “The strategy of the proteomics analysis of liver diseases” and “Collaboration with other projects”.

The summary of development of proteomics-based technologies and strategies

At first, the scientific and technological frames and infrastructure established by HLPP team have greatly launched international human proteome initiatives and head the prosperity of proteomics in China, which were highly appreciated by Nature (*David Cyranoski. China takes centre stage for liver proteome. 425:41,2003*), Science (*Robert F. Service. Public projects gear up to chart the protein landscape. 302:1316,2003*) and Nat Biotech. (*Heping Jia. China pushes liver proteomics. 22:136, 2004*).

HLPP paid great attention on elaborating the strategies for proteomics analysis since the beginning. This was first proved in the preview project of HLPP, the human fetal liver proteome project, whereas we demonstrated the successful application of a CCPIT (comprehensive and complementary proteome identification technology) strategy, which comprises subcellular fractionation and multiple protein separation and identification technology (*MCP, 2006*). Further application of these techniques in the internationally collaborated Human Plasma Proteome Project (HPPP), also helped to define the complexity of the plasma/serum proteome (*Proteomics, 2005, HPPP Special Issue*). The HPPP efforts finally promoted the production of novel algorithms that could be used to derive high-confidence protein identification from massive mass spectra (*Nat Biotech, 2006*). The proteome strategies adopted in HLPP took full advantage of the knowledge accumulated during the earlier projects. In the initiative of HLPP, the importance of establishing SOPs from sample collection to protein/peptide separation, from mass spectrometer acquisition to data interpretation and presentation, was fully appreciated. The pre-publication (*Fuchu He, Proteomics, 2006*) on quality control of sample collection demonstrates the proteome profiles in the human liver are relatively stable under normal physiology and a pooled sample for proteomic analysis at the initial stage can represent the normal individual. These results are the important basis for liver protein expression profiles and comparative proteomics of liver disease,

which were highly appreciated by *JPR (The normal liver proteome. 5:3232, 2006)*. Under the guide of the SOPs, the technique strategies were optimized and combined. Our results were shown as follows: 1) Complementary identifications among different technique platforms did exist and thus guaranteed the in-depth coverage of complicated proteome; 2) The use of prefractionation techniques at protein level expanded the dynamic range of identified proteins; 3) The number of identifications tended to increase a lot and could reach a plateau after one strategy was run multiple time, i.e., 6 times for offline shotgun; 4) High quality datasets were obtained through consecutive filter of noisy data, by retaining only identifications with tandem mass spectra information, and proteins with at least two uniquely identified peptides.

The optimized technique strategies implemented in HLPP also underwent further exam under a HUPO coordinated projects, in which a simulated protein mixture were delivered to 27 internationally renowned labs for testing the reliability of mass-spectrometry-based proteomics. BPRC held one position among the 7 labs initially reported all proteins correctly (*Nat Methods, 2009*).

Major Achievements in the Reporting Period

HLPP has made tremendous progresses since initiative. We set up a management infrastructure, identified reference labs, confirmed standard operating procedures (SOPs), initiated international collaborations and finally finished the protein expression profile of liver tissue and organelles, human liver transcriptome and protein-protein interaction map and achieved the first data set of the modification profile and proteome localization map. During this reporting period since last July, the major accomplishments of the HLPP are as follows:

The databanks

To popularize the valuable resources of HLPP, Chinese Human Liver Proteome Profiling Consortium has developed three databases named Liverbase (<http://liverbase.hupo.org.cn>), dbLEP (<http://dblep.hupo.org.cn>) and LiveMap (<http://livermap.hupo.org.cn>). The overall objective of the Liverbase is to provide a unique public resource for the liver research community by providing comprehensive functional annotation of proteins implicated in liver development and disease. The Liverbase integrates information on the human liver proteome, including the function, abundance and subcellular localization of proteins as well as associated disease information. The key features of Liverbase are manually annotated proteins localized in or functionally associated with human liver. In this first version of Liverbase, the data sets

include the human liver proteome (6,788 proteins) and transcriptome (11,205 heavily expressed genes: 10,224 from CHIP and 5,422 from MPSS, respectively) from the CNHLPP. As a database made publicly available through the web site, Liverbase provides browsing and searching capabilities and a compilation of external links to other databases and homepages. As a professional liver database, Liverbase includes comprehensive reviews and functional summarization for each individual protein.

The dbLEP (Database of Liver Proteome Expression Profile) aims to be an information centre for liver protein expression profile. So far, dbLEP holds three datasets of liver proteome expression profiles, i.e. human fetal liver, HLPP French liver with approximately 17,247 proteins and 36,990 peptides, and Chinese liver with 23,345 identified proteins. For each dataset, dbLEP provides all of the peptide and protein information, including none-redundant identified proteins, all identified possible proteins, peptides and their spectra. The detailed annotation as well as abundant links and flexible search functions are also provided for each identified protein.

LiverMap dataset is designed to store full description of the protein interactions, protein complexes and pathways in human liver. And now, this dataset contains 3484 high confident Y2H interactions between 2582 human liver proteins from CNHLPP. For each interaction, we presented the experimental data together with their detailed biological annotation from PRINCESS (Protein Interaction Confidence Evaluation System) of CNHLPP.

We also joined the international effort on proteome data by sharing with EBI (PRIDE database), Institute for Systems Biology, Seattle, USA (PeptideAtlas database), and Johns Hopkins University, Baltimore, USA (Mathivanan S, et al. Human Proteinpedia enables sharing of human protein data (*Nat Biotech.* 2008, Feb; 26(2):164-7). Based on the tremendous achievements and contributions in the project, we succeeded in pushing forward and widely popularizing the proteomics in China in the past several years.

As a result, proteomics got more attention and unprecedented support from the governments and most scientists. Not only has the proteomics research been placed as a core part in the first Major Scientific Research Program in the National Medium and Long Term S&T Development, but also the Pilot Hub Of ENcyclopedical proteomIX (PHOENIX), a national large scientific facilities for proteomics, has been approved to construct. The PHOENIX will be the first large scientific facilities specialized on life sciences among all of the Chinese national large scientific

facilities. It has attracted much attention from all over the world and was reported by *Science* after it got approval (*Science*, 2009, 323: 1417).

Expression and modification profiles

By utilizing the power of quantitative proteomics, one of the HLPP group in Fudan University (Shanghai, China) revealed that protein acetylation is a prevalent modification in enzymes that catalyze intermediate metabolism. Virtually, every enzyme in glycolysis, gluconeogenesis, the tricarboxylic acid (TCA) cycle, the urea cycle, fatty acid metabolism, and glycogen metabolism was found to be acetylated in human liver tissue. The reversible acetylation of metabolic enzymes ensure that cells respond environmental changes via promptly sensing cellular energy status and flexibly altering reaction rates or directions. The concentration of metabolic fuels, such as glucose, amino acids, and fatty acids, influenced the acetylation status of metabolic enzymes. The study suggests that acetylation plays a major role in metabolic regulation. The results have been published in *Science* in 2009.

In an effort led by Laura Beretta at the Fred Hutchinson Cancer Research Center, a total of 2,352,706 ms² spectra were processed using the Trans Proteomic pipeline (TPP) and PeptideAtlas at the Institute for Systems Biology (ISB, Seattle). A total of 44,710 unique peptides and 6,913 proteins were identified. Among them, 1,934 proteins were specifically expressed in liver and have never been detected in any of the other 229 human samples included in the human PeptideAtlas. Qualitative and semi-quantitative comparisons were performed with the corresponding transcriptomes. Similar proteomic profiling and data analysis was performed on the liver tissue and corresponding plasma of mouse models of hepatocellular carcinoma at different stages of disease progression. The comparative analysis of these data sets identified novel proteins as well as pathways and protein variants in liver disease.

Interaction and localization maps

To better understand the functional organization of the human liver proteome, under the common effort of Prof. Xiaoming Yang's group and other five labs, a high throughput Y2H platform was constructed and optimized for the large scale and massively parallel screen of protein-protein interaction in the liver library. 3,484 interactions between 2,582 human liver proteins were identified in BPRC, and a database (LiverMap) was designed to store the full description of these interactions. From them, 58 potential liver disease-associated proteins and 268 potential regulators of NF- κ B pathway were identified, of which GIT2 was characterized as a novel

modulator of the NF- κ B signaling pathway. It implicates that this organ specific PPI map could be valuable resource for lots of biological research fields.

Protein Localization maps

Proteins localization to different organelles is a central feature of cellular organization. We report more than 6311 proteins of the human liver organelle proteome and provide spatial information based on their relative abundance in the nuclei, plasma membranes, cytosol, mitochondria and endoplasmic reticulum. Of the 6311 proteins, 2527 had no subcellular localization deduced from Swiss-Prot. According to the protein abundance range, KNN (K-nearest neighbor) and Bayesian model, which combined KNN and other five sub-cellular localization prediction methods (pTAGET, Proteome Analyst, WoLFPSORT, TargetP and NUCLEO), were used respectively to newly map 1,858 proteins to five organelles in human liver. To validate our subcellular assignments, SEFC (seamless enzyme free cloning)-a high throughput cloning method, was employed to construct green fluorescent protein (GFP)-tagging vectors. And their subcellular localization was assessed by confocal microscopy. This method showed clear subcellular localization of 14/19 positive controls. We then tested 84 proteins that lacked prior experimental support of subcellular localization, 53% (45/84) of the results was in agreement with our prediction model results. The consistency is much higher than the reported success rate (32%, 131/404) from the first large-scale GFP microscopy of mitochondrial localization (2008, CELL), suggesting the high accuracy of our prediction and stability of SEFC. And the validation on more proteins localization is going on.

Based on some previous report and our own multi-organelle localization experiments, human liver proteome localization (HLP-Loc) database (version 1.0) has been constructed and will be provided as a public resource. As the first large scale human liver proteome localization database, HLP-Loc is the first step towards the goal to understand some new proteins function. The integration of the localization datasets with other functional data will ultimately provide a biological atlas of function and make the user to conduct systems biology research on liver.

Proteomic Analysis of Liver Diseases

Liver diseases are vital diseases influencing national welfare and people's livelihood. Proteomic analysis of the liver diseases is one of the most important directions in HLPP. The followings are some of the examples we have made during study:

Phase-specific proteomic patterns in the progression of nonalcoholic fatty liver disease.

Nonalcoholic fatty liver disease (NAFLD) has emerged as a common public health problem that can progress to end-stage liver disease. For the first time, the down-regulation of ECHS1 in the liver tissue was confirmed by immunohistochemistry in NAFLD patients, suggesting its potential as protein marker of NAFLD. Our results may help clarify the pathogenesis of NAFLD.

Main players in Hepatitis C virus (HCV) and Hepatitis B virus (HBV) infection. HCV and HBV infection are major health concerns worldwide. The group of Laura Beretta at the Fred Hutchinson Cancer Research Center has identified using a proteomic approach that HSP70 is associated with HCV particles and that this interaction is necessary for HCV life cycle. These data have been published in the past year in *Hepatology*. Subsequent studies performed in China have found that interaction between HSP90 and HSP70/HSP60 contributes to the HBV life cycle by forming a multi-chaperone machine and that the block of this interaction by a small chemical could stop HBV infection, suggesting that it might be saved as therapeutic targets for HBV-associated diseases.

Protein markers for HepatoCellular Carcinoma (HCC). HCC is a highly malignant tumor, and chronic infection with hepatitis B virus is one of the major risk factors for this disease. The finding and validation of the overexpression of Hsp70/Hsp90-organizing protein and heterogeneous nuclear ribonucleoproteins C1/C2 in multiple HCC tissues indicated their potential as protein markers for HCC. The analysis on Tissue interstitial fluid (TIF) from tumor and nontumor tissues of a hepatocellular carcinoma patient indicates the potential for biomarker discovery in TIF and it is the first analysis of the liver TIF proteome and provides a foundation for further application of TIF in liver disease biomarker discovery.

A major program on HCC biomarkers, funded by NIH and directed by Laura Beretta at the Fred Hutchinson Cancer Research Center is ongoing.

Biomarkers for prediction of HCC metastasis. Metastasis is the main cause for treatment failure and high fatality of HCC. Six dysregulated proteins have been validated in our study. Interestingly, the up-regulation of solute carrier family 12 member 2 (SLC 12A2) and protein disulfide-isomerase A4 (PDIA4) were further confirmed in the culture supernatants by Western blotting and in the sera of HCC patients with different metastatic potentials by ELISA. Our study provided not only the valuable insights into the HCC metastasis mechanisms but also the candidate biomarkers for prediction of HCC metastasis.

Using animal models for hepatic disease research and biomarker discovery. This is another active line within HLPP at CIC bioGUNE (Bilbao, Spain) which they are focusing on. The proteome of urinary vesicles present in urine samples obtained from experimental models for the study of liver injury was performed by this group with the goal to identifying potential biomarkers for hepatic disease. The proteomic analysis of highly purified exosome-like vesicles from urine samples was released. In total, 134 proteins were detected, including metabolic enzymes, solute transporters, peptidases and proteins involved in cell signaling and in cytoskeleton organization. Many of these proteins have previously been associated with diseases and a further characterization of them will increase the number of disease urinary biomarkers. The presence of different vesicle populations with a size smaller than 220nm was also demonstrated based on their protein composition. Moreover, the proteomic characterization of highly purified exosome-like urinary vesicle identified 28 proteins previously unreported in these vesicles, and many that have been previously associated with diseases, such as the prion-related protein. Furthermore, in urine samples from D-galactosamine-treated rats, a well-characterized experimental model for acute liver injury, a severe reduction in some proteins that normally are clearly detected in urinary vesicles was observed. Finally, differential protein content on urinary vesicles from a mouse model for chronic liver injury has been also identified. Our results argue positively that urinary vesicles could be a source for identifying non-invasive biomarkers of liver injury. Some proteins such as Cd26, Cd81, Slc3A1 and Cd10 have been found to be differentially expressed in urinary vesicles from some of the analyzed models suggesting their potential role as biomarkers for liver injury.

Mouse models of liver cancer have been also extensively characterized by Laura Beretta's group at the Fred Hutchinson Cancer Research Center. In-depth profiling has been performed in liver

tissue and plasma at different of disease progression. This analysis identified early markers of HCC.

Interaction with other HUPO initiatives

A major accomplishment has been the data integration and cross-analysis of the data sets of the liver reference sample generated by the Beretta group at the Fred Hutchinson, the plasma data set collected by Gil Omenn and Ruedi Aebersold and the kidney and urine data sets generated by Tadashi Yamamoto and colleagues.

Finally, a special issue on liver proteomics (editor Laura Beretta) has been published earlier this year in *Proteomics-Clinical Applications*.