
Activity Report of HUPO Human Liver Proteome Project (HLPP)

(July 1st, 2008 to August 30th, 2009)

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

(Originally prepared by Ms. Xue Gao and Mr. Junjie Zheng from Beijing Proteome Research Center)

Workshop in the Reporting Period

An HLPP workshop was held in conjunction with the 7th HUPO 2008 World Congress in Amsterdam, Netherland. The workshop was chaired by Laura Beretta (Fred Hutchinson Cancer Research Center) and Xiaohang Qian (Beijing Proteome Research Center). The workshop was organized around 3 sessions: Session 1- The human proteome project: the liver perspective, moderated by John Bergeron (McGill University, Canada); Session 2- The comparative analysis of the liver and plasma proteomes, moderated by Young-Ki Paik (Yonsei University, Korea) and Session 3- Opportunities for the study of liver diseases within the HLPP, moderated by Chantal Housset (INSERM, France). During the first session, Laura Beretta presented the current status of the genome coverage obtained from both a MS-based and an antibody-based analysis of the Liver Reference Samples. John Bergeron gave a presentation on a cell map approach to the study of the liver proteome and Bill Jordan presented the progress of the AOHUPO membrane proteome initiative. Finally, Paul Boersema described liver specificity (expression/quantitation) analysis applied to define protein complexes. In Session 2, dedicated to the comparative analysis of the liver and plasma proteomes, such study using mouse models was presented by Laura Beretta and the recently developed database “biomarker digger” was presented by Young-Ki Paik. In the final session dedicated to liver diseases, Chantal Housset presented an update on the biobanks established in France as well as the needs in clinical management of patients with liver diseases. Juan Falcon-Perez (CIC bioGUNE, Spain) presented some mouse models of liver diseases and Songfeng Wu

(Beijing Proteome Research Center) introduced repositories for HLPP expression and interaction data.

Major Accomplishments in the Reporting Period

With the latest progress, the focus of HLPP has been shifted gradually from the normal liver samples to diseased ones, from the whole liver tissues to the hepatocellular organelles and different sorts of liver cells, and from the static expressions to dynamic changes. Here is the brief summary of the major work of the HLPP since last July:

1. Protein Expression Profiling

A publicly available new website ‘Liverbase’ (<http://liverbase.hupo.org.cn>) was established, which integrates information on the human liver proteome, including the function, abundance, and organelle localization of proteins as well as associated disease information. The overall objective of the Liverbase is to provide a unique public resource for the liver community by providing comprehensive functional annotation of proteins implicated in liver development and disease. The central database features are manually annotated proteins localized in or functionally associated with human liver. In this first version of Liverbase, the associated data includes the liver proteome (6,788 identified proteins) and transcriptome (11,205 transcripts: 10,224 from CHIP and 5,422 from MPSS, respectively) from the Chinese Human Liver Proteome Project (CNHLPP). As a database made publicly available through the website, Liverbase provides browsing and searching capabilities and a compilation of external links to other databases and homepages. Liverbase enables (i) the establishment of liver GO slim with 51 nonredundant items; (ii) systematic searches of proteins within specific functional or metabolic pathways; (iii) systematic searches that aim to find the proteins that underlie common and rare liver diseases; and (iv) the integration of detailed protein annotation from the literature. Liverbase also contains an external links page with links to other biological databases and homepages, including GO, KEGG, pfam, SWISS-PROT, and GNF databases.

Liverbase users can utilize all these information to conduct systems biology research on liver.

A second major data set with over 2 million spectra obtained by Laura Beretta's group at the Fred Hutchinson Cancer Research on the French Liver Reference Sample has been searched through different pipelines at both the Fred Hutchinson Cancer Research Center and the Institute for Systems Biology in Seattle. A total of 474,617 spectra (23,886 unique peptides) were identified with very high confidence and corresponded to over 6,000 proteins. This liver proteome data set as well as the associated transcriptome will be made publicly available in 2009 following the HUPO annual meeting in Toronto.

Interest in hepatocytes as a biochemical research tool has led to develop the method of isolation this cell. A simple and economic isolation procedure is used for the hepatocytes and the purity can reach up to 95-99% with cell viability 85-90%. The protein expression profile of isolated hepatocytes is constructed by using 1DE-LC-ESI and RCX-LC-ESI: SDS-PAGE or RCX for proteins pre-fractionation and RPLC for separation of digested peptides followed with ESI-MS/MS (LTQ-FT). The first patch of 3,321 proteins with at least 2 peptides was identified with false discovery rate lower than 1%. This dataset will help understand the biological characteristic of the major cell in liver.

2. Post-Translation Modification (PTM) Profiling

In order to overcome one of the major challenges on enriching the low-abundance proteins in the post-translational modifications, Profs. Xiaohong Qian and Yun Cai's groups developed an optimized technological platform for large-scale phosphoproteomic study and used to phosphopeptides enrichment and high-accuracy mass spectrometric identification of phosphoproteins of C57BL/6J mouse liver. Totally, 2,434 phosphorylation sites corresponding to 1,048 phosphoproteins were identified with a 1% FP via triplet experiments. Preliminary analysis for this dataset

has been performed with GO annotation and Motif analysis. This is the first dataset of C57BL/6J mouse liver phosphoproteome under physiological conditions, which provides a good foundation for mouse liver-related systematical phosphorylation research. Besides the phosphorylated proteome analysis, a global profiling and quantitative characterization of core fucosylated glycoproteome were also performed in their lab. Previous studies have revealed that core fucosylation (CF) patterns of some glycoproteins are more sensitive and more specific than evaluation of their total respective protein levels for diagnosis of many diseases, such as cancers. A robust strategy has firstly been developed in their lab, which involves a novel glycopeptides enrichment method based on molecular weight cutoff membrane, neutral loss-dependent MS³ scan, database-independent candidate spectrum filtering and optimization to effectively identify CF glycoproteins. The rationale for spectrum treatment was innovatively based on computation of the mass distribution in spectra of CF glycopeptides. The efficacy of this strategy was demonstrated by implementation for plasma from healthy subjects and the subjects with hepatocellular carcinoma. Over 100 CF glycoproteins and CF sites were identified, and over 10,000 mass spectra of CF glycopeptide were found. The scale of identification results indicates great progress for finding biomarkers with a particular and attractive prospect, and the candidate spectra will be a useful resource for the improvement of database searching methods for glycopeptides.

Laura Beretta's group analyzed changes occurring in the plasma N-glycoproteome during liver disease progression from fibrosis to cancer using a mouse model and characterized 569 N-glycosites. Similarly phosphopeptides and their changes with liver disease progression were identified in the same mouse model with 465 characterized phosphosites.

Prof. Young-Ki Paik's group has recently identified liver carboxylesterase 1 (hCE1) as biomarker candidate of HCC which is remarkably down-regulated in tumor tissues but highly overexpressed in plasma specimens from HCC patients compared to other

disease patient groups (e.g., liver cirrhosis, chronic hepatitis, cholangiocarcinoma, stomach cancer, and pancreatic cancer). From the receiver operating characteristic (ROC) analysis in HCC, both sensitivity and specificity were shown to be greater than 70.0% and 85.0%, respectively. Paik's lab has also developed an efficient online TiO₂-based 3DLC (SCX/TiO₂/C18)-MS³-linear ion trap system which provides fully automatic and highly efficient identification of phosphorylation sites in complex peptide mixtures. Using this system, low-abundance phosphopeptides were isolated from cell lines, plasma, and tissue of healthy and hepatocellular carcinoma (HCC) patients. Furthermore, the phosphorylation sites were identified and the differences in phosphorylation levels between healthy and HCC patient specimens were quantified by labeling the phosphopeptides with isotopic analogs of amino acids (SILAC for HepG2 cells) or water (H₂¹⁸O for tissues and plasma).

3. Protein-Protein Interaction Mapping

Systematical mapping of human protein-protein interactions have been conducted, but it is far from completed. One major goal of HLPP is to study the binary protein-protein interactions between liver proteins and the important liver protein complexes. Under the common effort of Prof. Xiaoming Yang's group and other 5 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 and a 5,000 gene array. More than 3,400 interactions were obtained and the quality of the dataset was validated by multiple independent experimental assays and bioinformatics analysis. More than 60% of 150 randomly selected pairs were verified by independent co-immunoprecipitation, GST-pull down or co-localization assays. Systematic integrating with liver gene and protein expression data, we identified some of liver specific functional modules such as metabolism, transport and transcription. The molecular mechanisms of specific liver physiological and pathological-related interactions were also investigated. The protein linkage map of human liver may be a

valuable source for human protein-protein interactome and the studies of liver diseases.

4. Protein Localization Mapping

Dynamic movements or/and transportation of proteins from one location to another site can often be followed by the physiological and pathological cell procedures, and the intracellular location can be viewed as a significant sign of protein functions. In the HLPP, construction of proteome localization map is essential to the elucidation of liver protein function. In order to examine the localization of proteins in mammalian cells on large-scale or global analyses, Prof. Xuemin Zhang's group developed a new system for comprehensive protein localization analysis, a cell-based high-throughput localization array which was integrated with laser scanning confocal microscope, image processing and other technologies. More than 1,000 proteins could be localized and screened at one time in subcellular level using this system. This system has been used to examine the co-localization of multi-proteins, discover the transforming factor closely-related to hepatic cell tumorigenesis, and screen the interaction between different proteins in mammalian cells, such as PIAS3 and PRB. Worth to mention, in screening substrates of ubiquitin-proteasomes pathway with this system, 112 proteins were tested and 24 candidate substrates, such as MAPKAPK3, NLK, and RhoGDI2, identified as novel target of ubiquitin. Although it still has some shortcomings, such as insufficient relationship with bioinformatics technology, the future of this system is promising. In the light of the next plan, the technology will be used as a key tool to validate the results from the other HLPP laboratories which are focused on the subcellular proteins expression profile and the protein-protein interactions.

Besides, an analytical method of protein co-localization and a microscope-based analytical system of protein interactions were set up. On this technology platform, the localization, translocalization, co-localization and interaction of proteins can be systematically analyzed.

5. Proteomic Analysis of Liver Diseases

Proteomic analysis of the liver diseases has been one of the most important directions for the HLPP. Furthermore, the comparison of the proteomic data of liver tissues with the blood samples needs to be carefully considered. A big special project on the proteomic analysis of liver diseases has been funded by Chinese Ministry of Science and Technology. Prof. Beretta is supported by grants from NIH to analyze the liver and plasma proteomes in disease.

Prof. Beretta's group has been analyzing the liver and plasma proteomes from 3 mouse models of liver cancer at different stages of disease progression from fibrosis, fatty liver to HCC. One of these models has been developed by Jose Mato and therefore this study establishes a basis of collaboration within HLPP between Profs Beretta and Mato. Prof Beretta's group has also been analyzing the plasma proteomes from patients with hepatitis C virus-associated liver cirrhosis and from patients with hepatitis C virus-associated liver cancer and identified protein changes occurring during the establishment of hepatitis C virus chronic infection.

Prof. Yinkun Liu's group studied the proteomic changes in liver tissues during the progression of the liver diseases, from the "normal liver" to hepatitis, hepatic fibrosis, hepatocirrhosis, HCC staging and its metastasis. In this study, they found 79 differential proteins which could be classified as 10 expression patterns by cluster analysis and 25 important staging biomarkers. Some of those potential biomarkers for liver cancer were validated, such as HSP27, CK19, HnRNA, Prx and Annexin. Meanwhile, 26 important different proteins in the serum, which were classified as 5 types, were identified to be related to different stages of liver diseases. It was found that haptoglobin may be a potential biomarker candidate in the early diagnosis of liver cancer. The concentrations of serum HP in HCC patients were higher than those in LC patients. Haptoglobin (AUC = 0.782) had greater accuracy than AFP (AUC = 0.733). The diagnostic values of HP were of 72.7% sensitivity at 70% specificity with the comparison of serum AFP (68.9% sensitivity at 70% specificity). As the combination

of Hp and AFP it could greatly improve the diagnostic accuracy (AUC = 0.838). Moreover, the serum Hp for AFP-negative patients between HCC and LC also had diagnostic value (AUC = 0.763). This observation implied that HP might be a potential candidate for assisting diagnosis, especially for the individuals with negative serum AFP. According to the results and other past work, they primarily established the multiple recognition on the progression of the liver diseases.

For cell surface N-glycan alteration, they developed a lectin microarray combined with glycosyltransferase oligochips and found the obviously change in cell surface of N-glycan between nonmetastatic HCC cell line (Hep 3B) and the HCC cell lines with different metastatic potential (MHCC97H and MHCCLM3), and the responsible glycosyltransferase.

The process of Non Alcoholic Fatty Liver Disease (NAFLD) includes steatosis, steatohepatitis and liver fibrosis. Prof. José Mato's group integrated metabolomic profiling to identify the NAFLD biomarkers. In a proteomic approach, differences in protein expression between patients with NAFLD and healthy controls were studied. Changes in protein expression of liver samples from each of the three groups of subjects, controls, non-alcoholic steatosis, and non-alcoholic steatohepatitis (NASH), were analyzed DIGE combined with MALDI TOF/TOF. Ten of the differentially expressed proteins were further analyzed by Western blot in tissue samples to confirm the observed changes. Following this approach one marker of steatosis, three NASH markers, and one marker common to both pathological stages were identified. The expression of these proteins was further validated by Western blot in serum samples of the three different cohorts of patients. Overall, three proteins whose level of expression can be correlated to a disease state in serum were found. NAFLD animal models were also established for biomarker discovery, including MAT1A deletion (inducing steatosis, NASH, and HCC), GNMT deletion (resulting in steatosis, fibrosis, and HCC) and CD81 transgenic mice (spontaneously developing steatohepatitis).

The analysis of urinary sub-proteomes such as urinary vesicles as an alternative for

biomarker discovery has been also explored. Urinary vesicles present in rat and mouse urine samples have been analyzed by proteomic methods, as an approach for identifying potential biomarkers for diseases before attempting human trials. Biochemical and proteomic characterization of highly purified exosome-like urinary vesicles was performed and 28 proteins previously unreported in these vesicles have been reported together with many others that have been previously associated with diseases. In addition, by using two experimental models one for acute and other for chronic liver injury, several exosomal proteins have been shown to be altered and subsequently they could be considered as interesting candidate biomarkers to be included in future studies.

In another study, Prof. Chantal Housset's collaborative group compares the proteome of the tumor and non-tumor areas of liver samples, which were isolated by using laser microdissection (LM). Differential protein profiling exhibited noticeable differences between tumor and non-tumor: 30% of the protein spots with deregulated expression in tumorous LM-samples did not display any modification in homogenates; conversely 15% of proteins altered in tumorous homogenates were not impaired in LM-hepatocytes. In France, eight liver centres have joined their effort to collect, preserve and distribute tissue samples from hepatocellular carcinomas, according to the recommendations for Biological Resources Centres (BRC) released by the Organisation for Economic Cooperation and Development (OECD), which were recently implemented in France. This led to the creation of a Liver cancer biobanks network. Samples are collected and stored locally, whereas related annotations are remotely collected within a central database. More than 1,400 tumor samples and adjacent liver samples, are made available to research groups following acceptance of research projects submitted to a scientific committee. Samples from the Liver cancer biobanks network were used in recent works aiming at identifying key genomic alterations occurring in the development of hepatocellular carcinomas, as well as new set of biomarkers, including transcriptomic and proteomic signatures, and candidate

targets for anti-cancer drugs (Clément et al. Cancer Lett 209). Assurance quality of both tissue samples and annotations, as well as access to samples from other participating centres are major added values. The network has been opened to the International Cancer Genome Consortium (ICGC) which aims at generating comprehensive catalogues of genomic abnormalities in tumours from 50 different cancer types and/or subtypes, including hepatocellular carcinoma

6. HLPP databank

With the rapid progress of HLPP, massive data have been generated since 2004, of which, two kinds of dataset are really amazing. One is for the proteins expression profile, and the other is for the protein-protein interaction.

To manage the valuable resource effectively and present it for researchers with different interests to carry out analysis conveniently, a web-based database of liver proteome expression profile named dbLEP has been developed. The researchers especially those interested in large scale proteome or specific protein function could get great benefit from it. Currently dbLEP holds three datasets, human fetal liver, HLPP French liver with approximately 17,247 proteins and 36,990 peptides, Chinese liver with 607,851 identifications corresponding to 62,117 peptides containing 45,781 peptides with tandem MS spectra led to the identification of 23,345 proteins with a 95% confidence level.

In Chinese liver dataset, after removal of duplicates, the number of identified proteins was reduced to 12,951. By eliminating the proteins with only one peptide match, the number of identified proteins was further reduced to 6,788. Further analysis of the detection frequency for each protein showed that 82% of the proteins were identified with more than three times. Besides the normal 95% confidence data set, the higher confidence (99%) data set was also provided to confirm the biological conclusion. At the 99% confidence, 5,454 unique proteins were identified, of which 3,013 were identified with two or more unique peptides.

Some other datasets, such as HLPP Chinese liver organelles, organelles of C57 mouse liver will be online soon. Both non-redundant proteins and all possible proteins are presented so that users could understand each dataset comprehensively. Besides the complete identification, dbLEP provides related information like MS for users to verify the confidence. Plenty of protein annotation, such as description of function, family, GO classification etc. and lots of cross links to the other related databases, such as UniPort, Kegg and HPRD *etc.*, are provided as well. dbLEP is accessible at **<http://dblep.hupo.org.cn>**.

The databanks for protein-protein interaction, localization and the PTM profile have being built, and some data have been generated and integrated. For instance, a databank was established for protein-protein interactions found from the HLPP. Moreover, to expand interaction network and provide valuable annotation and evidence for protein and interaction of HLPP dataset, we also collect and demonstrate protein-protein interactome identified in experiment and other renowned databases between human and mammals, including IntAct, HPRD, Bind, Mint and Dip. Additionally, abundant gene annotation is available as well.