Analysis of human liver proteome using replicate shotgun strategy

ARTICLE *in* PROTEOMICS · JULY 2007

Impact Factor: 3.81 · DOI: 10.1002/pmic.200600338 · Source: PubMed

10 AUTHORS, INCLUDING:

Ming Chen
Thomas Jefferson University
26 PUBLICATIONS **214** CITATIONS
SEE PROFILE

Ying Jiang
Guilin Institute of Electronic Technology
73 PUBLICATIONS **900** CITATIONS
SEE PROFILE

Yun Cai
Shanghai Jiao Tong University
57 PUBLICATIONS **1,253** CITATIONS
SEE PROFILE

RESEARCH ARTICLE

Analysis of human liver proteome using replicate shotgun strategy

Ming Chen[1, 2], *Wantao Ying*[1, 2]*, *Yanping Song*[1, 2], *Xin Liu*[1, 2], *Bing Yang*[1, 2], *Songfeng Wu*[1, 2], *Ying Jiang*[1, 2], *Yun Cai*[1, 2], *Fuchu He*[1, 2] *and Xiaohong Qian*[1, 2]

[1] Department of Genomics and Proteomics, Beijing Institute of Radiation Medicine, Beijing, P. R. China
[2] Beijing Proteome Research Center, Beijing, P. R. China

In this study, a liquid-based shotgun strategy was used to comprehensively identify the expression of human liver proteome. Proteins were extracted from human liver tissue and digested in-solution. The tryptic digest mixture was desalted and separated by off-line strong cation exchange (SCX) chromatography with a 60-min elution. The MS/MS spectra were acquired in data-dependent mode after an RP chromatographic separation combined with linear IT MS analysis. To obtain the most comprehensive human liver proteome, each SCX fraction was run six times in RPLC MS/MS manner. Finally, more than 6 000 000 MS/MS spectra were collected. Using a relatively strict filter criteria, 24 311 proteins (48.42% of the predicted human proteome from human International Protein Index (IPI) protein database 3.07) corresponding to 13 150 nonredundant proteins were successfully identified, in which 7001 proteins (53.24%) were identified by two or more peptides, which could be considered as a high-confident dataset. Among the 6149 proteins (46.76%) identified by single peptide, 3812 proteins (61.99%) were detected more than twice in six repeated runs. Comparative analysis between different runs shows that the overlap of identified proteins between any two runs ranged from 25 to 44%. Of the nonredundant proteins identified, 8919 proteins (67.83%) were detected more than twice and 4231 proteins (32.17%) were detected only once in six RPLC MS/MS runs. The Gene Ontology annotation shows that the identified proteins come from various subcellular components. In addition, a large number of low abundant proteins were identified. The dynamic range of the approach reached at least nine orders of magnitude by estimating the concentration of proteins.

Received: May 4, 2006
Revised: Decemeber 27, 2006
Accepted: April 9, 2007

Keywords:
Human liver / Shotgun strategy / SEQUEST

1 Introduction

Liver is the biggest organ in the human body, which plays many key roles in life process, such as detoxification, metabolism of fat, protein and sugar, production of blood and bile, *etc.* [1]. The knowledge of the liver proteome could help us to explore not only the basic physiological function of liver, but also the potential mechanisms of liver diseases. Although the great achievements in the field of genomics disclose that there were approximate 25 000 genes in human [2], the proteome in the tissue is still hardly to be accurately determined because of the wide dynamic range and variable number of the proteins and their possible isoforms. Thus, our knowledge of the liver proteome is still limited.Recently, a large-scale proteome profile of human plasma has been announced [3, 4]. At the same time, other proteome projects

Correspondence: Dr. Xiaohong Qian, Department of Genomics and Proteomics, Beijing Institute of Radiation Medicine, 27 Taiping Road, Beijing 100850, P. R. China
E-mail: qianxh@nic.bmi.ac.cn
Fax: +86-10-8070-5155

Abbreviations: IAA, iodoacetamide; **IPI**, International Protein Index; **MudPIT**, multidimensional protein identification technology; **SCX**, strong cation exchange; **LTQ**, linear trap quadrupole

* Additional corresponding author: Dr. Wantao Ying
E-mail: yingwt@nic.bmi.ac.cn

have been carried out under the authorization of HUPO, such as the human liver proteome project (HLPP) and human brain proteome project (HBPP) [5]. The traditional approach to proteome expression profiling was 2-DE separation followed by MS or MS/MS analysis. However, the bias of 2-DE against proteins with extreme p*I* and molecular weight as well as the high hydrophobic property limited its wide application [6]. Therefore, analysis of complex protein mixture using 1-D SDS-PAGE combined with MS/MS was developed and successfully employed for comprehensively profiling proteome of several human and animal cell lines [7, 8]. Another approach to global analysis of proteome was one or 2-D LC separation followed by MS/MS analysis (LC-ESI-MS/MS, multidimensional protein identification technology (MudPIT)), which was first introduced by Yates and colleagues [9]. The successful applications of MudPIT strategy to proteome of complex system have been reported [10–13]. Such as Yu *et al.* [10] successfully identified 4542 proteins in mouse cortical neuron with MudPIT strategy. Washburn *et al.* [9] applied the same technology and identified 1574 proteins in human mammary epithelial cells. As we know, most biological samples have wide dynamic range in protein concentration (such as plasma), which limited the ability of MS to effectively monitor the low abundant species. How to increase the dynamic range of proteomic analytical strategy is a big challenge. Wolters *et al.* [14] reported that sample prefractionation both in protein or peptide level could extend the dynamic range by about four orders of magnitude. Shen *et al.* [15] performed a repetitive ultra-high-efficiency strong cation exchange (SCX)-LC-RPLC MS/MS analyses and identified human plasma proteins with a dynamic range of eight orders of magnitude. Theses results suggested that repetitive RPLC MS/MS runs could improve the dynamic range of analysis methodology by increasing the capture frequency of low abundant proteins. Liu *et al.* [16] studied the relationship between the abundance of protein and spectrum count, and found that peptide hits and spectrum counts were coincident with protein abundance, lower abundant proteins could be approached by increasing the replicating number of RPLC MS/MS runs performed on a sample.

To accomplish a comprehensive identification of human liver proteome, a shotgun strategy was carried out with six replicating SCX-LC-RPLC MS/MS runs, which finally resulted in the identification of 24 311 proteins (48.42% of the predicted human proteome from human IPI protein database 3.07) corresponding to 13 150 unique protein/protein group. In this set, 7001 unique proteins (53.24%) were identified with two or more peptides matched. In 6149 proteins (46.76%) identified by single peptide, 3812 proteins were detected more than twice in the six RPLC MS/MS runs. In addition, a large number of low abundant proteins were identified. The dynamic range of the approach reached at least nine orders of magnitude by estimating the concentration of proteins. We conclude that replicate shotgun strategy is a high-throughput, sensitive analytical method for large-scale profiling of human liver proteome.

2 Materials and methods

2.1 Chemicals and reagents

HPLC-grade ACN was purchased from JTBaker (NJ, USA). HPLC-grade water was produced by a Milli-Q A10 system from Millipore (Billerica, MA, USA). Iodoacetamide (IAA) and DTT were obtained from ACROS (NJ, USA). Modified sequencing-grade trypsin was supplied by Promega (Madison, WI). PMSF, NaF and FA were purchased from Sigma (St. Louis, MO, USA).

2.2 Sample preparation

Human liver tissue, acquired from healthy donors, was collected and distributed from CNHLPP (China Human Liver Proteome Project) Sampling Committee. The use of these liver samples for proteomics research was approved by the local ethical committee. To reduce the individual variations, ten individual samples were pooled to create a uniform sample. Briefly, fresh liver samples were cleaned with ice-cold PBS at 4°C soon after receiving. Then, the samples were cut into pieces, weighed and preserved in liquid nitrogen until use. Frozen liver tissue (0.5 g) was ground into powder in liquid nitrogen and homogenized in 5 mL lysis buffer, containing 9.5 M urea, 1% DTT, 1 mM PMSF, 0.2 mM Na_2VO_3 and 1 mM NaF. The homogenization was sonicated on ice for 1 min with 1-s pulse-on and 1-s pulse-off. The mixture was incubated at room temperature for 30 min with repeated vortexing. The resultant suspension was centrifuged at $25\,000 \times g$, 20°C for 1 h. The supernatant protein concentration was measured by the RCDC protein assay method. The supernatant was then frozen in liquid nitrogen, and stored at −80°C for further analysis.

2.3 Protein digestion in solution

Human liver protein solution (100 μL), about 933 μg protein, containing 9.5 M urea and 6 5mM DTT, was diluted using 100 μL 50 mM NH_4HCO_3, the mixture was denatured and reduced at 37°C for 4 h, then the alkylation was performed by adding 33 μL 1 M IAA to the mixture and incubated at room temperature for 1 h in the dark. After that, 90 μL ACN and 577 μL 50 mM NH_4HCO_3 was added to the solution, then the sequencing-grade trypsin was added at the ratio of enzyme to protein 1:100, the mixture was vortexed and incubated at 37°Cfor 2 h, and another 1:100 trypsin solution was added to ensure complete digestion. The resulted mixture was continuously incubated for 24 h. After incubation, the tryptic peptide mixture was acidified with 5% formic acid and lyophilized by SpeedVac (Thermo Savant). The production was stored at −20°C for further analysis.

2.4 Peptide desalting by SPE and separated by off-line SCX

LC-18 SPE tube (Supelco, Beooefonte, PA, USA) was used to desalt the peptide mixture. The tube was first conditioned with 2 mL methanol, 2 mL of 0.1% TFA water solution, respectively. Then, 900 μL peptide mixture (pH 3.0) was loaded on the SPE tube and salt was removed by washing the tube with 1 mL of 0.1% TFA water solution. Finally, the desalted peptide was eluted from the tube with 1.5 mL of 80% ACN with 0.1% TFA, and dried under vacuum. Peptide mixture was separated by using the Hypersil SCX column (Thermo Keystone, 4.6 × 150mm) and Beckman Coulter PF-2D system (Beckman Coulter, Fullerton, CA, USA). Mobile phase A was 5 mM KH_2PO_4 containing 25% ACN adjusted to pH 3.0 with formic acid. Solvent B was solution A plus 350 mM KCl. The gradient elution profile used for solvent B was generally as follows: 0% B for 7 min; 0 to 15% B in 1 min; 15% B for 5 min; 15 to 25% B in 10 min; 25% B for 5 min; 25 to 60% B in 10 min; 60 to 100% B in 1 min; 100% B for 10 min; 100 to 0% B in 1 min; 0% B for 30 min before the next run. The flow rate was 0.7 mL/min, effluents were monitored at 214 nm, and fractions were collected every 1 min with an automated Gilson 215 liquid handler (Gilson, Middleton, WI, USA).The system was controlled by 32 Kraft software for SCX and Uniprot software (Beckman Coulter) for collector. All fractions were dried with a SpeedVac Concentrator (Thermo Savant) and stored at -20°C for further analysis.

2.5 Nano RPLC-ESI-MS/MS analysis

SCX fraction was analyzed by RPLC MS/MS using a Thermo Finnigan™ linear IT mass spectrometer (LTQ) equipped with an ESI source. Initial nanoLC run was performed on an LC-Packing system, which was configured with Famous, Swithos and Ultimate, and controlled by Dionex chromatography software. Two RP-C18 trap columns were connected with the 10-port valve, which allows one trap column loaded and the other eluted to MS through a PicoFrit™ tip column (BioBasic®C18, 5 μm, 75 μm id x 10 cm, 15-μm id spray tip, New Objective, Woburn, MA, USA) by Ultimate pump. In this study, about 19 μL SCX fraction was injected into one of trap columns by Famous loading pump with a flow rate of 10 μLl/min under the UserProg injection mode. The dissolved peptides were desalted on the trap column and eluted through the PicoFrit™ to MS instrument by Ultimate pump using a gradient of 2–40% B (80% ACN, 0.1% TFA in water) in 9 0min and 40–100% B in 15 min. The LTQ-MS was operated in data-dependent mode using normalized collision energy of 35%. The temperature of the ion transfer tube was set at 200°C and the spray voltage was set at 1.8 kV. The MS analysis was performed with one full MS scan followed by five MS/MS scans on the five most intense ions from the MS spectrum with the dynamic exclusion settings: repeat count 1, repeat duration 30 s and exclusion duration 30 s.

2.6 Database searching

All MS/MS spectra were searched against the human IPI protein database (version 3.07) using the SEQUEST algorithm on Bioworks software (version 3.1) for peptide and protein identifications. All results were combined together using BuildSummary software [17], and filtered by the parameters of X_{corr} scores $\geq$1.9 for singly charged, X_{corr} scores $\geq$2.2 for doubly charged, and X_{corr} scores $\geq$3.75 for triply charged. ΔCn cutoff value of $\geq$0.1 and RSP $\leq$4 were used. Gene ontology analysis was completed using the GOfact in the web site: www.hupo.org.cn [18].

3 Results and discussion

3.1 Human liver proteins identified by off-line SCX and on-line RPLC-ESI-MS/MS

The high complexity of biological sample makes it very difficult to be analyzed directly by MS on large-scale proteome profiling. Therefore, peptides produced by tryptic digestion were first separated by SCX and then by RPLC before introduction into a linear IT mass spectrometer. Usually, off-line or on-line SCX as the first dimension combined with RPLC as the second dimension was used to separate peptide mixture, which has been proved a very efficient strategy as the prefractionation strategy in peptide level for the complex sample. As described by Li *et al.* [4], although sample losing could be avoided and the process could be automated in the peptide elution with step salt gradient in on-line SCX, the low resolution of the method was still lethal for MS/MS analysis. Whereas, in off-line SCX approach, peptides could be eluted with linear salt gradient instead of steps in on-line SCX, which performs excellent separation efficiency and can collect more fractions [4, 10, 12]. In our study, the SCX condition was optimized first to get best resolution between the salt concentrations of 0–350 mM before human liver sample experiment, then proteins extracted from human liver tissue were digested in solution and the resulted tryptic peptides were separated by off-line SCX. The same separation was performed three times and 60 fractions were collected during each elution and combined with three collections. Because a good reproducibility is essential for our parallel analysis, it was assessed by signal intensity, retention time, as well as the chromatographic profile. The chromatograms of three independent injections are shown in Supporting Information (Fig. S1). The patterns of the peaks and the elution times of three runs were very similar and no obvious changes in chromatographic profiles were observed. After separation, collected fraction was subjected to duplicate LTQ analysis through a PicoFrit™ column. Therefore, six replicating runs were performed in sum from three times of off-line SCX separation. Because few chromatogram peaks were observed in the first 16 min during the whole 60-min SCX elution, the first 16 fractions were combined together in

every 4 fractions. A high sensitive LTQ mass spectrometer was utilized to analyze each fraction after SCX separation. More than 6 000 000 MS/MS spectra were obtained from six MudPIT runs. To acquire high-confident identification proteins, we strictly filtered the identified peptide by using the SCX fraction 38 and 49 as an example (Fig. 1) and found that the most (98%) peptides had +2 and +3 charge state while only about 2.02% peptides had +1 charge. For $X_{corr} \geq 1.5$, peptides (single charged) were identified with nearly 0% false-positive rate, for $X_{corr} \geq 2.2$ and 2.8, for +2 and +3 charge peptides, the false-positive rate was less than 10%. Similar as the results in SCX fraction 38 and 49, in all identified 328 988 peptides, 19 644 (5.97%) were identified as $[M+H]^+$, 299 808 (91.13%) were $[M+2H]^{2+}$, and 9536 (2.89%) were $[M+3H]^{3+}$ (data not shown), suggesting that more confident results could be obtained by increasing the X_{corr} value of +2 charged peptide. Recently, Wang *et al.* [19] have established the expression profile of mouse brain proteome. They used the cutoff $X_{corr} \geq 1.6$, $X_{corr} \geq 2.4$ and $X_{corr} \geq 3.2$ for +1, +2 and +3 charge state peptide, respectively, and finally identified 7792 proteins with a false-positive rate of less than 5%. Qian *et al.* [20] reported that the false-positive rate of peptide assignment is directly dependent on the nature of the sample. In this study, we used $X_{corr} \geq 1.9$, $X_{corr} \geq 2.2$, $X_{corr} \geq 3.75$, for +1, +2 and +3 charge state peptides, respectively, and obtained the false-positive rate of less than 10%. Finally, 35 658 unique peptides and 24 311 proteins were identified. To obtain the nonredundant protein/protein groups, DTA and OUT file from all six runs were combined by using BuildSummary software, which finally presented 13 150 unique protein groups. Among them, 7001 proteins (53.24%) were identified by two or more peptide hits (Fig. 2A). In summary, two grades of datasets under different confidence level were obtained, one dataset being proteins identified by one or more peptide hits, and the other being proteins identified by two or more peptide hits.

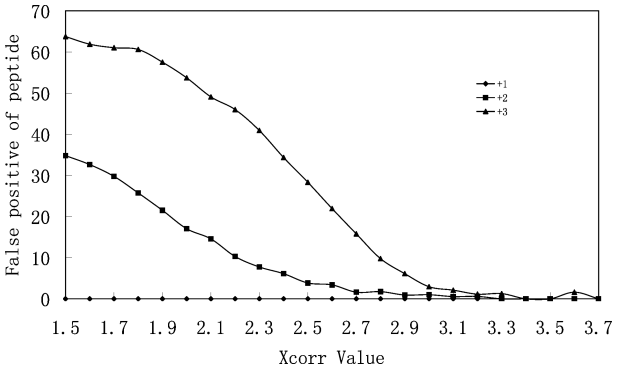


Figure 1. Evaluation of different X_{corr} value cutoff on the positive ratio of peptide identified level at +1, +2 and +3 charge states, and all with ≥Cn≥0.1. The data were obtained by searching the MS/MS data of SCX fraction 38 and 49 against a target-decoy IPI human database (version 3.07). The false-positive rate was calculated by doubling the number of peptides found from the reverse database and dividing the result by the total number of identified peptides from both databases.

3.2 Physical chemical characteristics of the identified proteins

To further analyze the characteristics of identified proteins, annotation was performed to those proteins identified by two or more peptide hits. The proteins were classified based on their molecular weight (MW) and p*I*. Among the 7001 unique proteins, there were 32 (0.46%) proteins with MW<10 kDa and 1791 (25.58%) proteins with MW>100 kDa. Regarding the p*I* distribution, a total of 6552 (93.59%) proteins distributed within the p*I* range of 4–10, only 4 proteins had p*I*<4, and 445 (6.36%) proteins had p*I*>10, among them, 15 proteins had p*I*>12. However, only 565 (8.07%) proteins were found with the p*I* value in 7–8. The MW and p*I* distributions of identified proteins by our

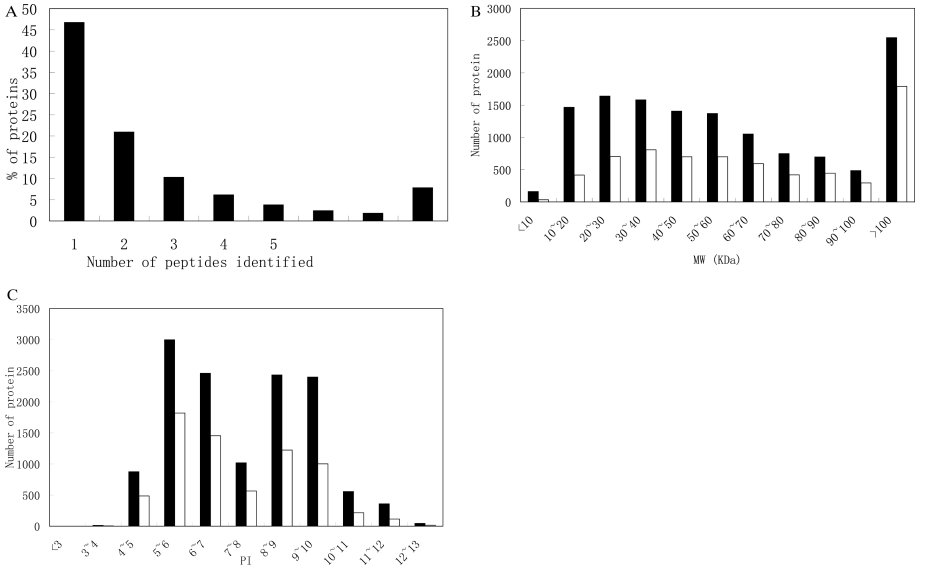


Figure 2. Histogram illustrating the characteristics of identified human liver proteins. (A) Distribution of the 13 150 identified proteins in relation to the number of the matching peptides used for identity assignment. (B) Distribution of the identified proteins in relation to their molecule weight, and p*I* (C) (■, one or more peptides hits; □, two or more peptides hits).

repetitive shotgun strategy are shown in Figs. 2B and C. It is interesting that the MW and p*I* distributions of identified proteins by single or more peptide hits were similar to those proteins by two or more peptide hits (Figs. 2B and C).

Traditionally, 2-DE coupled with MS is chosen as a mature technique in proteomics analysis. However, the discrimination of this method against proteins with extreme p*I* and MW limits its wide application in large-scale proteome expression profile analysis. Replacing 2-DE gel with 1-D gel could lead to an increase of identified proteins from solution protein pools [21]. However, the recovery of the protein/peptide after in-gel digestion and peptide extraction is still an issue. The data of this study reveal that a large number of proteins with extreme p*I* and MW, which probably be lost in 2-DE or 1-DE strategies, were successfully identified. Among these proteins, 750 (10.71%) proteins had MW higher than 160 kDa, and 32 (0.46%) proteins had MW lower than 10 kDa; the lowest and highest MW of proteins were 6.04 and 3816.22 kDa, respectively. On the other hand, the highest p*I* of the identified protein reached 12.66, and 347 (4.96%) proteins had p*I* value above 10. Figure 2 shows clearly that many extreme characteristics proteins (about 10%) have been identified, although most proteins (90%) were in the range of MW 20–160 kDa and p*I* value 3–10, demonstrating that compared with gel-based method, the solution digestion combined with 2-D LC could supply a more efficient separation for the peptides and increased their chances of being captured and identified by LTQ mass spectrometer. The same result was also reported by Shen and Barnidge [15, 22].

3.3 Gene ontology annotation

To further understand the functions of the proteins we identified, gene ontology (GO) annotation was performed. Among 7001 proteins, 4440 (63.40%) proteins were assigned to molecular functional classes, 3798 (54.20%) were mapped to biological process, and 3358 (48.00%) were ascribed to different subcellular locations. Eleven subcategories were divided for molecular function ontology, in which the catalytic proteins accounted for 42.70% (1898). More important, some putative low abundant proteins were detected, such as transcription factor (130) and signal transducer (444). For biological process, 2461 (64.80%) proteins were identified as metabolism related, 197 (5.20%) were cell cycle related and 50 (1.30%) reproduction related proteins (Supporting Information Fig. S2). For the subcellular location of identified proteins, most of them were localized on membrane, nucleus, cytoplasm and extracellular matrix, in which 1080 (32.20%) proteins were annotated as membrane proteins, 932 (27.80%) proteins were noted as nucleic proteins, and the other proteins came from subcellular organelle, such as mitochondrion (8.00%), Golgi (3.90%), ER (6.50%) and so on (Supporting Information Fig. S2). Altogether, during the global sample preparation, no effort was made to target any specific subset of proteins, thus, we believe that the proteins identified by our strategy could well cover the proteome expression profile of human liver.

3.4 Liver special proteins identified in replicate MudPIT runs

Metabolism (including detoxification) and reproduction are two main functions of human liver. Accordingly, 64.80% annotated proteins were metabolism related, 42.70% proteins had catalytic activity and 50 (1.30%) proteins were reproduction related (Supporting Information Fig. S2), *e.g.* bone morphogenetic protein 15 precursor, MTL5 protein, pregnancy-specific beta-1 glycoprotein and transcription factor MafF, *etc.* In addition, some proteins proved to be liver specific were also identified in our experiment, such as fatty acid-binding protein, 7α-hydroxylas, insulin-like growth factor family and their binding proteins and so on. The 7α-hydroxylas, a rate-limiting enzyme in the bile acid synthetic pathway, is considered as the production of a liver-specific gene [23]. Other proteins, *e.g.* biliverdin reductase and heme oxygenase (which catalyze reduction of biliverdin to produce bilirubin in liver) and Cytochrome p450 (which is known to metabolize a variety of compounds including nicotine, cotinine, butadiene, *etc.*) were also identified. GST are a family of enzymes involved in the binding, transport, and detoxification of a wide variety of endogenous and exogenous components. The mean total GST-α in human liver cytosols was 25.16 mg/mL [24]. However, interleukin-12 was reported to have a normal concentration of 77 pg/mL in liver [25]. These two proteins were found in our dataset, indicating that the replicated shotgun strategy used in this study could reach a dynamic range of nine or more orders of magnitude in protein concentration.

3.5 Parallel analysis of replicated MudPIT

As discussed above, multidimensional chromatography was generally employed to simplify the complexity of peptide mixture. Although losing of proteins is inevitable during the perfractionation, the great success of this method in protein identification has been reported [4, 9, 10, 12, 17]. However, although the sample has been simplified by multiple dimension LC, the collection of MS/MS spectrum and protein identification process of mass spectrometer has been proved still less reproducible, especially for the low abundant proteins [14, 16]. To comprehensively identify human liver proteins, a replicated MudPIT strategy with six RPLC MS/MS runs was performed in this study. Identification of proteins in any two runs was compared and evaluated. Table 1 shows the number of peptides and protein groups derived from six runs. The relationship between cumulative number of identified proteins and the number of MudPIT runs indicated that six RPLC MS/MS runs were completely reasonable (Supporting Information Fig. S3). From Table 2, we can see that the overlap rates of protein groups between any two runs were in range from 25 to 44%. Although the chromato-

Table 1. Number of peptides, protein groups derived from six replicated shotgun runs

	Run1	Run 2	Run 3	Run 4	Run 5	Run 6	Total
Unique peptide	16 990	17 709	9679	8940	5485	5379	35 658
Unique/redundant protein group	6950	7370	5249	4991	3542	3507	13 150
	/8736	/9229	/6308	/5921	/4068	/4006	/24 311
Cumulative number of proteins	6950	10 137	11 459	12 340	12 777	13 150	13 150
Number of new additional proteins	/	3187/2021*[a)]	1322/984*	881/735*	437/382*	373/308*	6200/1876*

a) The numbers marked with "*" represent the numbers of new additional proteins detected only once with the increasing number of MudPIT runs.

Table 2. Reproducibility of non-redundant protein identification from replicated shotgun analysis

	Run1	Run 2	Run 3	Run 4	Run 5	Run 6	Total
Run 1	100%	41.6%	32.3%	29.8%	25.1%	25.0%	50.9%
Run 2		100%	33.9%	31.2%	26.2%	25.8%	54.0%
Run 3			100%	43.1%	34.6%	34.2%	38.5%
Run 4				100%	34.2%	33.8%	36.6%
Run 5					100%	44.4%	25.9%
Run 6						100%	25.7%
Total							100%

graphic separation process employing multidimensional separations has been shown fairly reproducible, the collection of MS/MS spectrum and the proteins identification process proved to be less reproducible, because there are many factors causing variation in MS analysis, such as the ionization efficiency of ESI, co-elution of the peptides from LC column, the randomicity of peptide captured and so on. To further investigate the reproducibility of replicated RPLC MS/MS, we analyzed the detection frequency of overlap proteins in any two independent runs. Figure 3 shows the linearity of frequency of identified unique proteins in all replicated RPLC MS/MS runs. The detection frequency for each protein was plotted; if each protein showed the same frequency in all the replicate experiment, the correlation coefficient would be 1. Note that the detection frequency of the protein identified in two independent RPLC MS/MS runs had perfect linear correlation, and R^2 values between 0.7442 and 0.9889 were observed (Supporting Information Fig. S4), demonstrating that the reproducibility of these RPLC MS/MS runs was acceptable. For example, the spectrum counts of 14-3-3 protein (zeta/delta) was 47, 78, 32, 54, 45 and 66 in run 1, run 2, run 3, run 4, run 5 and run 6, respectively, with CV about 30%. The proteins presented repeatedly could be validated reciprocally, and enhance the confidence level of identification results. Information about the protein abundance is very important in proteome research. In traditional 2-DE technique, this information could be deduced from the staining intensity, but what about it in shotgun strategy? Some researches reported that the number of identified peptides per protein (peptide hits), the ion counts of identified peptide (spectrum counts) or the total ion chromatography intensity, could be considered as the indicator for protein abundance [26, 27]. In fact, many publications have reported the relationship between the MS information of a protein and its abundance in cell, even the mathematical formula had been derived by statistical analysis [16, 27, 28]. In this study, we analyzed the detection frequency of the peptides by linear IT mass spectrometer in six runs, and found that 452 (1.27%) peptides were identified more than 100 times, 2152 (6.04%) peptides were detected 30~100 times, 2243 (6.29%) peptides were detected 20~30 times and 30 810 (86.40%) peptides were detected less than 10 times (Fig. 4A). We also analyzed the detection frequency (spectrum counts) of each unique protein in six RPLC MS/MS runs, and found that 8919 (67.83%) proteins were detected more than twice, in which, 2953 (22.46%) proteins were identified more than ten times, and 4231 (32.17%) proteins were detected only once in six runs (Fig. 4B). Partial list of the proteins identified less

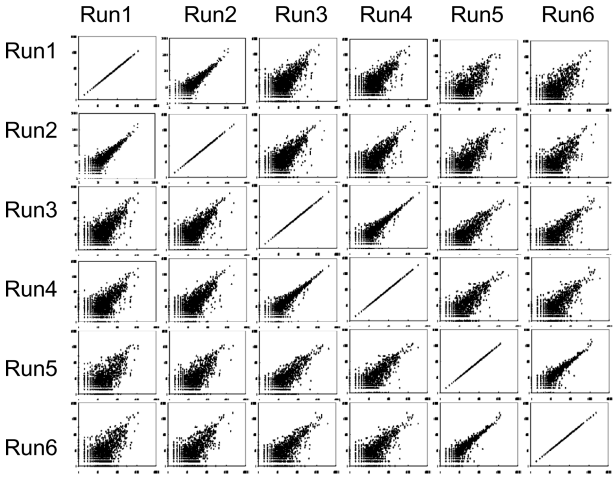


Figure 3. Linearity of frequency of overlap unique proteins distribution in difference runs. The x-axis and y-axis represent the detecting frequency of a protein in two independent LC MS/MS runs, respectively. The coordinate axis was converted into logarithmic calibration. Each dot in the graphs represents the detecting frequency of a protein identified in two independent RPLC MS/MS runs.

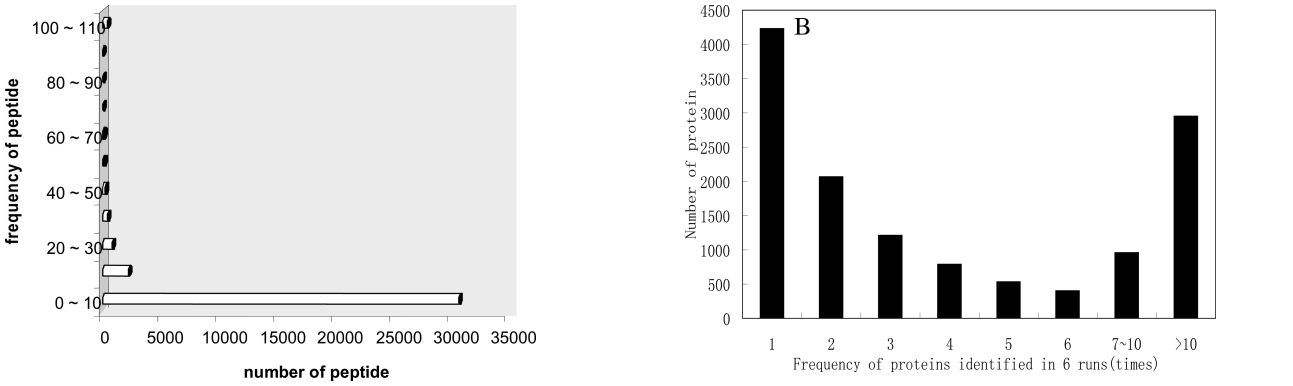
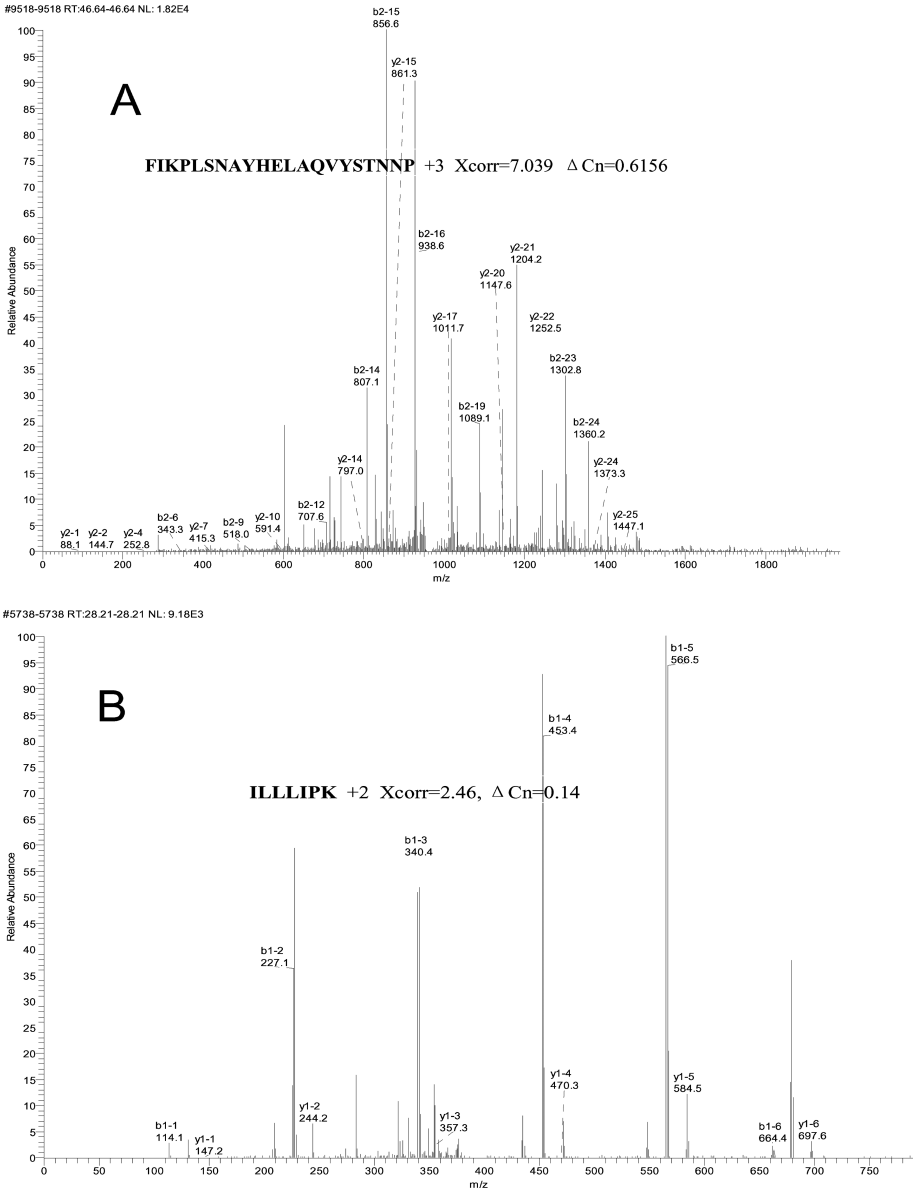


Figure 4. Distribution of the frequency of peptide (A) and frequency of protein identified in six LC-MS/MS runs (B).

Table 3. Partial proteins identified less than ten times in six MudPIT runs

IPI	Protein name	Frequency	Function	Subcellular
IPI00011757	RNA polymerase II elongation factor ELL2	2	transcription regulator	Nucleus
IPI00026627	XRP2 protein	3	protein binding	Membrane
IPI00298182	DNA-binding protein RFX2	2	transcription regulator	Nucleus
IPI00555830	Poly(A) polymerase beta (testis specific)	1	nucleic acid binding	Nucleus
IPI00555830	Macrophage stimulating 1	2	ion binding	Unknown
IPI00020902	Zinc finger protein	2	ion binding	Nucleus
IPI00004379	H6 homeodomain protein	2	transcription regulator	Nucleus
IPI00549854	Cytoglobin	3	antioxidant activity	Cytoplasm
IPI00305626	HypotHetical protein LOC253982	1	catalytic activity	Membrane
IPI00292026	CCAAT/enhancer binding protein delta	2	nucleic acid binding	Nucleus
IPI00100192	Hypothetical protein DKFZp566B1046	4	catalytic activity	Unknown
IPI00154603	Fibroblast growth Factor 2	3	receptor binding	Extracellular
IPI00334197	Krueppel-related zinc finger protein 2	1	ion binding	Nucleus
IPI00065557	Zinc finger protein 488	1	ion binding	Nucleus
IPI00001863	Wnt inhibitory factor 1 precursor	1	kinase activity	Unknown
IPI00001563	Cyclin-dependent kinase 4 inhibitor B	1	enzyme regulator	Unknown
IPI00001564	Cyclin-dependent kinase 6 inhibitor	1	enzyme regulator	Nucleus
IPI00001661	RCC1 protein	1	enzyme regulator	Nucleus
IPI00002325	DNA polymerase mu	1	DNA binding	Nucleus
IPI00001649	DnaJ	1	protein binding	Unknown
IPI00001726	Ubiquitin carboxyl-terminal hydrolase 35	1	hydrolase	Unknown
IPI00001786	USP36 protein	1	hydrolase	Unknown
IPI00002276	Galactoside-binding soluble lectin 13	1	hydrolase	Unknown
IPI00001883	Sorting nexin 9	1	signal transducer	Unknown
IPI00002283	PREDICTED: KIAA1337 protein	1	receptor activity	Unknown
IPI00001738	Nuclear pore complex protein Nup88	1	transporter	Nucleus
IPI00002150	Adapter-related protein complex 4 beta 1 subunit	1	transporter	Cytoplasm
IPI00002372	ATP-binding cassette sub-family D, member 3	1	transporter	Cytoplasm

than ten times in all six runs by mass spectrometer are listed in Table 3. One can see that most of them were binding and regulator proteins, among which, some low abundant proteins could be found, for example, macrophage-stimulating 1 in nanogram level and fibroblast growth factor 2 in picogram level in human plasma [15]. Some proteins identified only once in six RPLC MS/MS runs are also shown in Table 3. Regarding their function, we found that most of them were signal molecular and regulator molecular. This result is consistent with their low abundance within the cell. As a comparison, partial list of the proteins identified more than 1000 times in six RPLC MS/MS runs was submitted as Supporting Information (Table S1), we found that most of those proteins were enzyme or come from cytoplasm, such as catalase, splice isoform 1 of carbamoyl-phosphate synthase and peroxiredoxin 6, and they were all high abundant proteins in



liver. As mentioned above, more than 90% confidence in peptide level was acquired in our study, but still 46.76% (6149) proteins were identified by single peptide hits (Fig. 2A). A further analysis of those proteins was performed and it was found that most of them were low abundant proteins and play important function in liver. Washburn *et al.* [9] also noted that abundant proteins were identified with multiple peptides and low abundant proteins by one or two. Interestingly, those proteins could be divided into two categories: (i) the single peptide was grasped only once by mass spectrometer (2337 proteins 38.01%); (ii) the single peptide was grasped twice or more times (3812 proteins 61.99%). Partial proteins identified by single peptide with detecting frequency more than 10 times are listed (Supporting Information Table S2). We manually inspected the MS/MS spectra of those peptides. Figure 5 shows the examples of peptide FIKPLSNAYHELAQVYSTNNP, ILLLIPK, GKDAMQSFMMPFYLMK and GGHPPAIQSLINLLADNR, corresponding to COP9 signalsome complex subunit 3, Interleukin 23 receptor, WW domain-binding protein 2 and nuclear receptor coactivator 5, respectively. As can be seen, these spectra have fairly good quality for positive matches.

4 Concluding remarks

The research here presents a comprehensive analysis of human liver proteome. The results showed that replicate RPLC MS/MS strategy was a high throughput, sensitive analytical method for large-scale profiling of human liver proteome.

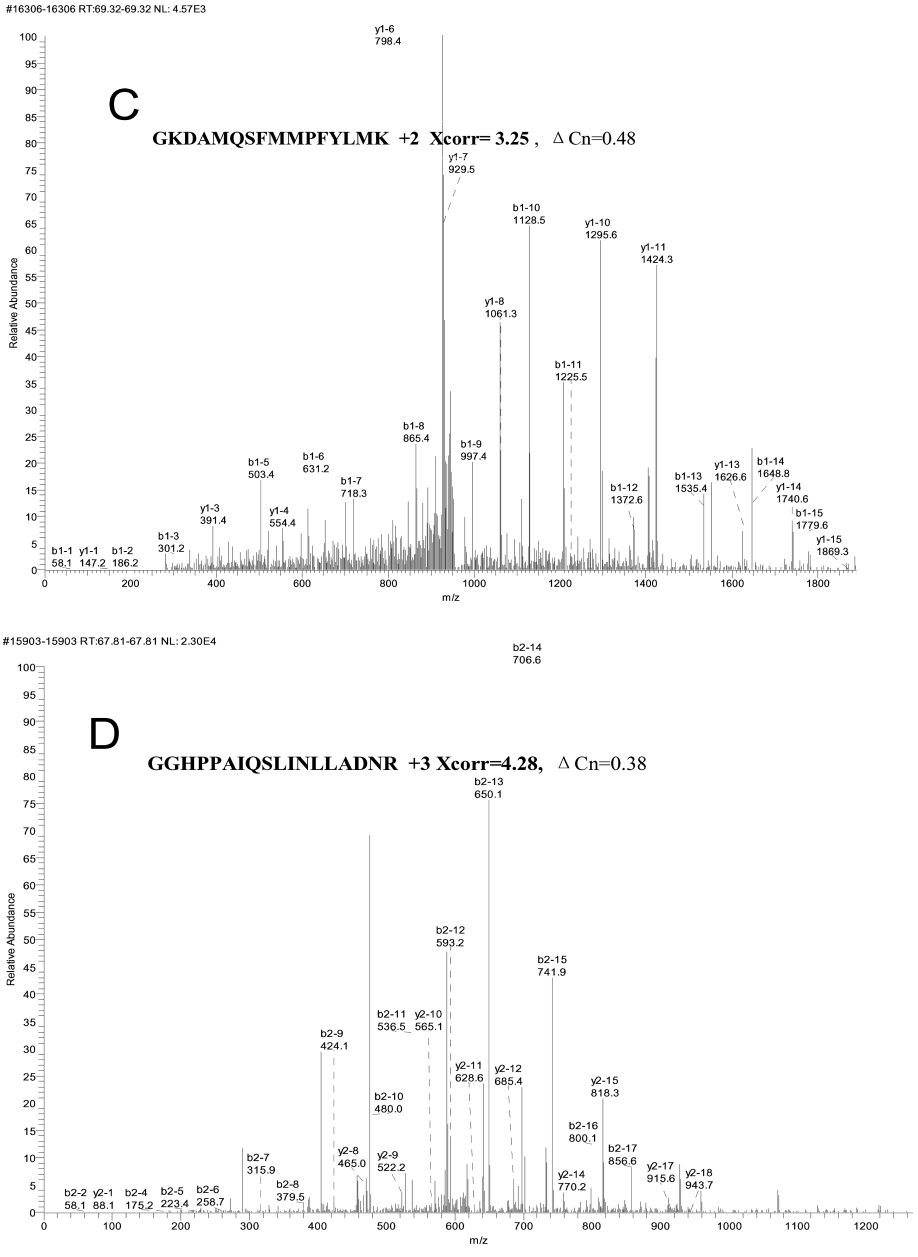


Figure 5. MS/MS spectra of peptide **FIKPLSNAYHELAQV-YSTNNP** (A), **ILLLIPK** (B), **GKDAMQSFMMPFYLMK** (C) and **GGHPPAIQSLINLLADNR** (D), corresponding to COP9 signalsome complex subunit 3, Interleukin 23 receptor, WW domain-binding protein 2 and Nuclear receptor coactivator 5, respectively.

More proteins, especially some low abundant proteins could be identified by increasing the number of LC-MS/MS runs performed on a sample. Furthermore, analyzing the counts of peptide or protein detected in different runs can give some information on their relative abundance, and the proteins identified in replicated runs could be validated reciprocally, to enhance the confidence level of identification results.

This work was supported by the National Key Basic Research Program of China (2001CB510201, 2004CB518707), the National Natural Science Foundation of China (30321003, 20405017, 20505018, and 20505019) and the Science and Technology Foundation of Beijing Municipality (H030230280190).

5 References

[1] He, F. C., *Mol. Cell. Proteomics* 2005, *4*, 1841–1848.

[2] Cagney, C., Amiri, S., Premawaradena, T., Lindo, M. *et al.*, *Proteome Sci.* 2003, *1*, 1–15.

[3] Omenn, G. S., States, D. J., Adamski, M., Blackwell, T. W. *et al.*, *Proteomics* 2005, *5*, 3226–3245.

[4] Li, X., Gong, Y., Wang, Y., Wu, S. *et al.*, *Proteomics* 2005, *5*, 3423–3441.

[5] Fountoulakis, M., Juranville, J. F., Dierssen, M., Lubec, G., *Proteomics* 2002, *2*, 1547–1576.

[6] Lambert, J. P., Ethier, M., Smith, J. C., Figeys, D., *Anal. Chem.* 2005, *77*, 3771–3788.

[7] Zhao, Y., Zhang, W., Kho, Y. J., Zhao, Y. M., *Anal. Chem.* 2004, *76*, 1817–1823.

[8] Chen, E. I., Hewel, J., Habermann-Felding, B., Yates III, J. R., *Mol. Cell. Proteomics* 2006, *5*, 53–56.

[9] Washburn, M. P., Wolfers, D., Yate III, J. R., *Nat. Botechnol.* 2001, *19*, 242–247.

[10] Yu, L., Conrads, T. P., Uo, T., Kinoshita, Y. *et al.*, *Mol. Cell. Proteomics* 2004, *3*, 896–907.

[11] Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J. *et al.*, *J. Proteome Res.* 2003, *2*, 43–50.

[12] Jacobs, J. M., Mottaz, H. M., Yu, L., Anderson, D. J. *et al.*, *J. Proteome Res.* 2004, *3*, 68–75.

[13] Lin, D., Tabb, D. L., Yates III, J. R., *Biochimica et Biophysica Acta* 2003, *1646*, 1–10.

[14] Wolters, D.A., Washburn, M. P., Yates, J. R. *Anal. Chem.* 2001, *73*, 5683–5690.

[15] Shen, Y., Jacobs, J. M., Camp, D. G., Fang, R. *et al.*, *Anal. Chem.* 2004, *76*, 1134–1144.

[16] Liu, H., Sadygov, R. G., Yates, J. R., *Anal. Chem.* 2004, *76*, 4193–4201.

[17] Jin, W. H., Dai, J., Li, S. J., Xia, Q. C. *et al.*, *J. Proteome Res.* 2005, *4*, 613–619.

[18] Li, D., Li, J., Ouyang, S., Wu, S. *et al.*, *Prog. Biochem. Biophys.* 2005, *32*, 1026–1029.

[19] Wang, H., Qian, W. J., Chin, M. H., Petyuk, V. A. *et al.*, *J. Proteome Res.* 2006, *5*, 361–369.

[20] Qian, W. J, Liu, T., Monroe, M. E., Strittmatter, E. F. *et al.*, *J. Proteome Res.* 2005, *4*, 53–62.

[21] Xiong, Y., Chalmers, M. J., Gao, F. P., Cross, T. A. *et al.*, *J. Proteome Res.* 2005, *4*, 855–861.

[22] Barnidge, D., Tschumper, R.C., Jelinek, D. F., Muddiman, D. C. *et al.*, *J. Chromatogr. B* 2005, *819*, 33–39.

[23] More, G. L., Drevon, C. A., Mechleder, D., Trawick, J. D. *et al.*, *Biochem. J.* 1997, *324*, 863–867.

[24] Mulder, T. P. J., Court, D. A., Peters, W. H. M., *Clin. Chem.* 1999, *45*, 355–359.

[25] Rotwein, P., *Proc. Natl. Acad. Sci. USA* 1986, *83*, 77–81.

[26] Pang, J. X., Ginanni, N., Dongre, A. R., Hefta, S. A. *et al.*, *J. Proteome Res.* 2002, *1*, 161–169.

[27] Ishihama, Y., Oda, Y., Tabata, T., Sato, T. *et al.*, *Mol. Cell. Proteomics* 2005, *4*, 1265–1272.

[28] States, D. J., Omenn, G. S., Blackwell, T. W., Fermin, D. *et al.*, *Nat. Biotechnol.* 2006, *24*, 333–338.

