Glycoproteomics Analysis of Human Liver Tissue by Combination of Multiple Enzyme Digestion and Hydrazide Chemistry

Rui Chen, Xinning Jiang, Deguang Sun, Guanghui Han, Fangjun Wang, Mingliang Ye, Liming Wang, and Hanfa Zou

J. Proteome Res., 2009, 8 (2), 651-661 • Publication Date (Web): 21 January 2009

Downloaded from http://pubs.acs.org on February 6, 2009

More About This Article

Additional resources and features associated with this article are available within the HTML version:

- Supporting Information
- Access to high resolution figures
- Links to articles and content related to this article
- Copyright permission to reproduce figures and/or text from this article

View the Full Text HTML
The study of protein glycosylation has lagged far behind the progress of current proteomics because of the enormous complexity, wide dynamic range distribution and low stoichiometric modification of glycoprotein. Solid phase extraction of tryptic N-glycopeptides by hydrazide chemistry is becoming a popular protocol for the analysis of N-glycoproteome. However, in silico digestion of proteins in human proteome database by trypsin indicates that a significant percentage of tryptic N-glycopeptides is not in the preferred detection mass range of shotgun proteomics approach, that is, from 800 to 3500 Da. And the quite big size of glycan groups may block trypsin to access the K, R residues near N-glycosites for digestion, which will result in generation of big glycopeptides. Thus many N-glycosites could not be localized if only trypsin was used to digest proteins. Herein, we describe a comprehensive way to analyze the N-glycoproteome of human liver tissue by combination of hydrazide chemistry method and multiple enzyme digestion. The lysate of human liver tissue was digested with three proteases, that is, trypsin, pepsin and thermolysin, with different specificities, separately. Use of trypsin alone resulted in identification of 622 N-glycosites, while using pepsin and thermolysin resulted in identification of 317 additional N-glycosites. Among the 317 additional N-glycosites, 98 (30.9%) could not be identified by trypsin in theory because the corresponding in silico tryptic peptides are either too small or too big to detect in mass spectrometer. This study clearly demonstrated that the coverage of N-glycosites could be significantly increased due to the adoption of multiple enzyme digestion. A total number of 939 N-glycosites were identified confidently, covering 523 noredundant glycoproteins from human liver tissue, which leads to the establishment of the largest data set of glycoproteome from human liver up to now.

Keywords: Glycoproteomics • Hydrazide chemistry • Multiple enzyme digestion • Human liver proteome project
analysis of N-glycoproteome. With high-abundance protein depletion and two-dimensional liquid chromatography MS/MS spectrometry, the hydrazide chemistry method resulted in the identification of 639 N-glycosylation sites in human serum. Nevertheless, the coverage of this method to study N-glycosylation in tissue is much lower. Only 445, 248 and 176 N-glycosites were identified from prostate cancer tissue, liver metastasis and bladder cancer tissue, respectively.

Although trypsin has the advantage of high specificity which cleaves exclusively at arginine (R) and lysine (K) residue, it is far from a perfect enzyme for studying postmodifications. Trypsin cleavage may be blocked by the attached carbohydrates sterically when trypsin cleavage site was located right after glycosylated asparagine. As a result, missed trypsin cleavage tends to occur more frequently when glycoprotein are digested. Moreover, some glycoproteins are not suitable for tryptic digestion either because they lack enough arginine and lysine residues in their sequences or they are not soluble at the optimal pH range of trypsin, for example, DQH sperm surface protein. This protein has extremely low solubility at neutral or slightly alkaline pH conditions. It is necessary to use V8 protein. This protein has extremely low solubility at neutral or optimal pH range of trypsin, for example, DQH sperm surface residues in their sequences or they are not soluble at the metastasis and bladder cancer tissue, respectively.

In addition, the digestion by protease with specificity complementary to trypsin can also generate more peptides within the preferable mass range for peptide identification and modification site location. With multiple enzyme digestion, it should be able to find N-glycosylation sites that would not be discovered with only trypsin digestion. To the best of our knowledge, the application of multiple enzyme digestion to study protein glycosylation was not conducted in proteome level up to now. The combination of this strategy with N-glycopeptide enrichment method such as hydrazide chemistry would be effective in the study of N-glycoproteomics.

The comprehensive analysis of human liver tissue N-glycoproteome by hydrazide chemistry in combination with multiple enzyme digestion strategy is presented in this work. Pepsin and thermolysin with broader specificity were used as complementary proteases to trypsin. It was found that this strategy can significantly improve N-glycosite coverage. A total number of 939 N-glycosites were identified with high confidence, covering 523 noredundant proteins. This is by far the largest data set obtained for human liver N-glycoproteome.

Materials and Methods

Materials and Chemicals. All the water used in this experiment was prepared using a Milli-Q system (Millipore, Bedford, MA); iodoacetamide (IAA) was purchased from AMERESCO (Solon, OH); dithiothreitol (DTT), ammonium bicarbonate, urea, trifluoroacetic acid (TFA), formic acid (FA) and acetonitrile (ACN) were all obtained from Aldrich (Milwaukee, WI). All the chemicals were of analytical grade except acetonitrile, which was of HPLC grade. 3-[(3-Cholamidopropyl)-dimethylammonio]-1-propane sulfonate (CHAPS) was from USB Corporation (Cleveland, OH). The Complete Mini protease inhibitor cocktail tablets were obtained from Roche Diagnostics (Mannheim, Germany). Trypsin (from bovine pancreas, TPCK-treated), pepsin (from porcine gastric mucosa) and thermolysin (from Bacillus thermoproteolyticus rokko) were obtained from Sigma (St. Louis, MO). Affi-Gel Hx Hydrazide Gel was purchased from Bio-Rad (Hercules, CA). PNGase F was from New England Biolabs (Ipswich, MA).

Sample Preparation. The human liver tissue was the noncancerous liver tissues ≥2 cm outside the hepatic cancer nodules removed by surgical operation from patients at the Second Affiliated Hospital of Dalian Medical University (Dalian, China). The noncancerous liver tissue has been verified by histopathological examination which excluded the presence of invading or microscopic metastatic cancer cells. The utilization of human tissues complied with guidelines of Ethics Committee of the Hospital. The isolated human liver tissue was placed in an ice-cold homogenization buffer consisting of 8 M urea, 4% CHAPS (w/v), 65 mM DTT, a mixture of protease inhibitor (Complete Mini protease inhibitor cocktail tablets, 1 tablet for 10 mL homogenization buffer) and 100 mM NH4HCO3 at pH 7.8. The tissues were homogenized in a Potter-Elvejem homogenizer with a Teflon piston, using 10 mL of the homogenization buffer per 2 g of tissue. The suspension was homogenized for approximately 1 min, sonicated for 100 W × 30 s and centrifuged at 25 000g for 1 h. The supernatant contained the total tissue proteins. The concentration of proteins was determined by BCA method, which was 7.8 mg protein/mL lysate. Appropriate volumes of protein sample were precipitated with 5 vol of 50% acetone, 50% alcohol and 0.1% acetic acid (v/v/v), lyophilized to dryness, and dissolved in reducing solution (8 M urea, 100 mM NH4HCO3). The sample was reduced by 10 mM DTT and carboxymidomethylated in 20 mM iodoacetamide.

Protein Digestion. Samples containing 5 mg proteins were digested with three different proteases separately using different
protocols. **Trypsin digestion**: Protein sample was diluted with 100 mM PBS buffer (pH = 8.0) to a urea concentration of 1 M. A total of 200 μg of TPK-treated trypsin was added and incubated for 24 h at 37 °C. The reaction was quenched by adding 10% TFA (v/v) to pH = 3. **Pepsin digestion**: Sample containing the same amount of proteins was diluted with 0.1% TFA (v/v) 4-fold and adjusted to a pH between 2 and 3 with 10% TFA (v/v). Pepsin was added at an enzyme-to-substrate ratio of 1:25 (w/w) and incubated for 24 h at 37 °C. The reaction was stopped by adding 10% TFA (v/v) to pH = 3. Each digest was stored at −80 °C until analysis.

**N-linked Glycopeptide Capture.** The N-linked glycopeptides in each digest were captured with hydrazide chemistry as described in a published protocol with minor modification.20 Digested sample solution was first desalted with homemade C18 solid phase extraction column. After it was dried in a Speed Vac (Thermo, San Jose, CA), it was reconstituted in 450 µL of oxidation buffer (100 mM NaAc, 150 mM NaCl, pH = 5.5); then, 50 µL of 100 mM sodium periodate was added, and the reaction was incubated at room temperature in darkness for 1 h. The oxidized sample was desalted with C18 column again and the eluted solution was directly added to 100 µL of Bio-Rad Affi-Gel H2 Hydrazide Gel (bed volume). The coupling reaction was left overnight at room temperature with shaking. The gel with captured N-linked glycopeptides was then washed with 1 mL of 1.5 M NaCl, methanol, and 100 mM NH4HCO3 sequentially three times to remove nonspecific adsorptions. After addition of 250 µL of fresh 100 mM NH4HCO3 and 10 µL of PNGase F (500 units per µL) to the gel, the enzymatic release of peptide moieties was carried out at 37 °C overnight. The end product for this procedure is the isolation of deglycosylated peptides (termed as N-glycopeptide in this study) that originally contain N-linked carbohydrates. The supernatant was collected by gentle centrifugation and combined with the supernatant of 100 mM NH4HCO3 wash fraction. The peptide solution was desalted by C18 column, dried and reconstituted in 25 µL of 0.1% formic acid as N-glycopeptide sample for LC-MS/MS analysis.

**Analysis of Enriched Glycopeptides.** The analysis of the N-glycopeptide samples was carried out by 2D nano-LC/MS/MS according to our previous method with minor modification.21 The 25 µL reconstituted glycopeptide sample was loaded onto a phosphate SCX monolithic column (150 μm × 7 cm i.d.) by air pressure. After that, the SCX column with loaded glycopeptides was manually connected to a C18 column (75 μm × 10 cm i.d.) in tandem by a union. The glycopeptides were eluted onto the analytical reversed phase column using step gradients generated with 1000 mM ammonium acetate in 0.1% formic acid. The salt concentrations of the buffer used for 10 step gradients were 50, 100, 150, 200, 250, 300, 350, 400, 500 and 1000 mM, respectively. Each salt step lasts 10 min with a following equilibrium by 0.1% formic acid for additional 10 min. After each elution step, a subsequent RPLC-MS/MS was executed with a 120 min gradient from 5% to 35% acetonitrile. The RPLC-MS/MS was performed on a nano-RPLC-MS/MS system. A Finnigan surveyor MS pump (Thermo Finnigan, San Jose, CA) was used to deliver mobile phase. For the C18 capillary separation column, one end of the fused-silica capillary was manually pulled to a fine point of ∼5 μm with a flame torch. The columns were in-house packed with C18 AQ beads (5 μm, 120 Å) from Michrom BioResources (Auburn, CA) using a pneumatic pump. The nano-RPLC column was directly coupled with a LTQ linear ion trap mass spectrometer from Thermo Finnigan (San Jose, CA) with a nanospray source. The mobile phase consisted of mobile phase A containing 0.1% formic acid (v/v) in water, and mobile phase B containing 0.1% (v/v) formic acid in acetonitrile. A Finngan LTQ linear ion trap mass spectrometer equipped with an ESI nanospray source was used for the MS experiment. A voltage of 1.8 kV was applied to the cross before the C18 capillary column. The LTQ instrument was operated at positive ion mode. Normalized collision energy was 35.0%. One microscan was set for each MS and MS/MS scan. All MS and MS/MS spectra were acquired in the data-dependent mode. The mass spectrometer was set that one full MS scan was followed by six MS/MS scans on the six most intense ions. The dynamic exclusion function was set as follows: repeat count 2, repeat duration 30 s, and exclusion duration 90 s. System control and data collection were done by Xcalibur software version 1.4 (Thermo). The scan range was set from m/z 400 to m/z 2000.

**Database Searching and Data Processing.** The MS/MS spectra were searched using SEQUEST22 (version 2.7) against a composite database including both original and reversed human protein database of International Protein Index (ipi.human.3.17.fasta, including 60 234 entries, http://www.ebi.ac.uk/IPI/IPIhuman.html). Enzyme was selected as the protease used for protein digestion. The cleavage site was set according to manufacturer’s guide, trypsin, KR/P; pepsin, FLE/VAG; thermolysin, LFIYMA/P. Enzyme limits was set fully enzymatic, cleaving at both ends. A maximum of two missed cleavage was allowed. Carboxamidomethylation (+57) was set as static modification and the following dynamic modifications were used: oxidation of methionine (+16), and a PNGaseF catalyzed conversion of asparagines to aspartic acid at N-glycosites (+1). The database search results of each protease were filtered with the criteria optimized by SFOER (version2.3).23 A false detection rate (FDR) of less than 2% was set for all peptide identifications. The majority of the identified peptides containing NXS/T-motif and only ~15% of identified peptides did not contain this motif. These peptide identifications probably resulted from nonspecific isolation of N-linked glycopeptides or random assignment in database search. This percentage of is lower than that in a similar work (~20%) for isolation of N-glycopeptides,11 which indicated the high specificity of N-glycopeptide isolation in this study. To reduce the false positive rate of identified N-glycosites, the peptides without such motifs were removed. After the filtration, few peptides that contained NXS/T were identified from the reversed database search (FDR < 0.1% for identification of N-glycopeptides was obtained), which suggested the high identification confidence. The identified glycoproteins from human liver tissue were categorized by Gene Ontology (GO) with component, function and process terms extracted from text-based annotation files downloaded from the European Bioinformatics Institute ftp site (ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/HUMAN/).

**Results**

**Glycoproteomic Analysis with Trypsin.** The proteins extracted from liver tissue were first digested by trypsin, and then the N-linked glycopeptides in the resulting digest was immobilized onto the hydrazide beads. After thoroughly washing, the peptide moieties of the captured N-linked glycopeptides were digested in solution and subjected to two-dimensional nano-LC/MS/MS. The N-linked glycopeptides were separated based on their molecular weights and charge states. The data were then processed by Xcalibur software and searched against the human protein database using the SEQUEST searching algorithm. The identified glycoproteins were further analyzed with Gene Ontology (GO) to gain insights into their biological functions. The results showed that a significant number of glycoproteins were identified, and the majority of the identified peptides contained NXS/T-motif and only ~15% of identified peptides did not contain this motif. These peptide identifications probably resulted from nonspecific isolation of N-linked glycopeptides or random assignment in database search. This percentage of is lower than that in a similar work (~20%) for isolation of N-glycopeptides,11 which indicated the high specificity of N-glycopeptide isolation in this study. To reduce the false positive rate of identified N-glycosites, the peptides without such motifs were removed. After the filtration, few peptides that contained NXS/T were identified from the reversed database search (FDR < 0.1% for identification of N-glycopeptides was obtained), which suggested the high identification confidence. The identified glycoproteins from human liver tissue were categorized by Gene Ontology (GO) with component, function and process terms extracted from text-based annotation files downloaded from the European Bioinformatics Institute ftp site (ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/HUMAN/).
were released from the beads by deglycosylation using PNGase F. The deglycosylated peptides (N-glycopeptides) were finally analyzed by 2D nano-LC-MS/MS. Three 2D nano-LC-MS/MS runs were performed in this study, which resulted in the identification of 840 unique tryptic N-glycopeptides from 384 glycoproteins. The identified N-glycopeptides and glycoproteins were given in the Supporting Information. On the basis of the NXS/T-motif, 622 N-glycosites were identified from human liver tissue.

The acquisition settings for MS/MS were in the range of 400–2000 Da in this study, thus, the peptide mass detection range should be 400–6000 Da considering three charge states of peptides (+1, +2, +3). However, the mass range of identified N-glycopeptides is much narrower than this range as shown in Figure 1A. Most of the N-glycopeptides locate themselves in a mass range of 800–3500 Da. For comparison, the MW distribution of nonglycopeptides identified by directly analyzing the tryptic peptides from the same sample with 1D nanoLC-MS/MS is also given in Figure 1B. It was found that the two distributions are very similar. The similar mass ranges of peptides identified by shotgun proteomics were well-documented in the literature. The small or big peptides cannot be identified through database search probably because these MS/MS spectra do not contain enough fragment information. The above facts indicate that the preferable mass range for peptides identified by shotgun proteomics is from 800 to 3500 Da.

The N-glycosite is not likely to be identified by shotgun proteomics approach if the mass of corresponding N-glycopeptides is not in the range of 800–3500 Da. Therefore, it is of interest to see how many tryptic N-glycopeptides derived from the proteins in the human proteome are within this range. The mass distributions of tryptic peptides and tryptic N-glycopeptides generated by in silico digestion of all proteins in the human protein database (IPI Human 3.17) were given in Figure 2A. The in silico tryptic N-glycopeptides are tryptic peptides containing NXS/T motif. The percentage of tryptic peptides and tryptic N-glycopeptides in the mass range of 800–3500 Da were

Figure 1. Molecular weight distribution of peptides identified from tryptic digest of protein extracted from human liver tissue. (A) N-glycopeptides enriched by hydrazide chemistry; (B) nonglycopeptides identified by direct analysis.
70.0% and 68.7%, respectively. Since most proteins would generate dozens of peptides after trypsin digestion and proteins could be identified by any of these peptides with their masses in the range, the theoretical proteome coverage by tryptic peptides was still close to 100% even though 30.0% tryptic peptides are not in this range. However, the situation for the identification of protein modification is different. The glycosite could be identified only if a peptide containing this site is identified. As only 68.7% tryptic N-glycopeptides are in the preferred mass range, a significant amount of N-glycosites could not be identified if only trypsin was used. The above-mentioned in silico digestion did not include the missed cleavage sites, and the missed cleavage is inevitable in protein digestion especially for glycoproteins. We analyzed the sequences of the identified N-glycopeptides and nonglycopeptides from human liver tissue. It was found that among the 840 identified unique tryptic N-glycopeptides, 169 (20.1%) have missed cleavage site. While among the 1069 identified by 1D LC-MS/MS, 152 of 1069 (14.2%) have missed cleavage sites, respectively. These results indicate that missed cleavage tends to happen more frequently when glycoprotein is digested. To investigate how missed cleavage affects N-

![Figure 2](image-url)

**Figure 2.** Comparison of molecular weight distribution of peptides by in silico digestion of proteins in human protein database (IPI Human 3.17) by trypsin between all the peptides (in blue) and peptides with NXS/T motif (in red) with different missed cleavage sites. (A) No missed cleavage site; (B) one missed cleavage site; and (C) two missed cleavage sites. The average molecular weight of each distribution is listed in the inserted table.
glycosite identification, in silico digestion of proteins from human protein database with one missed cleavage was also conducted and the resulting peptide mass distributions are shown in Figure 2B. It is astonishing to find that N-glycopeptides with missed cleavage sites were much bigger than those of all peptides. With one missed cleavage site, the average MW for all peptides and N-glycopeptides was 2422.61 and 3756.37 Da, respectively. There are still 71.9% of peptides locate themselves in the mass range of 800–3500 Da for all the peptides; however, only 52.2% of tryptic N-glycopeptides were in the mass range of 800–3500 Da. When two missed cleavages were considered, this percentage descended to 33.5%. For all peptides, this percentage is still as high as 59.2% as shown in Figure 2C. These data indicate that much more N-glycopeptides could not be identified when trypsin cleavage sites are missed. These facts demonstrated that it was inadequate to use only trypsin for N-glycosite identification. To improve N-glycosite coverage in bottom up approach, enzymes with complementary cleavage specificity should be adopted to digest proteins.

**Glycoproteomic Analysis with Multiple Proteases Digestion.** We adopted multiple enzyme digestion to improve the coverage of N-glycosites for glycoproteome analysis. The aim of multiple enzyme digestion was once to promote sequence coverage in identification of amino acid sequence variations in shotgun analysis by the means of generating overlapped peptides. When it was applied to the protein identification from complex samples or tissue, the entire protein sequence of abundant proteins could be covered in the chances of identifying a modification like phosphorylation on a specific amino acid residue. Pepsin and thermolysin were used as complements to trypsin in this study. After digestion, the N-glycopeptides were enriched by the same hydrazide chemistry method and analyzed by 2D nano-LC-MS/MS. The number of identified N-glycosites and glycoproteins by these two proteases as well as by trypsin were given in Figure 3. The full lists of identified N-glycopeptides and glycoproteins were given in the Supporting Information. Among all 939 N-glycosites identified, only 20 N-glycosites were identified by all the three proteases and another 125 N-glycosites were identified by both proteases. These N-glycosites were identified by multiple peptides generated by different proteases; thus, the identification of these sites was very confident. The use of multiple proteases can reduce the ambiguity in mapping modifications and increase the possibility of obtaining high quality MS/MS spectrum to identify peptides.

However, the identified N-glycopeptides sharing the same N-glycosite account for a very small proportion in the total number of identified N-glycopeptides. Much more N-glycopeptides with different N-glycosites were identified using different proteases which indicated good complement in N-glycosite identification to trypsin and so the N-glycosite coverage could be significantly increased. The other two proteases used in this study led to identification of 317 (33.7%) additional N-glycosites and 139 (26.6%) additional glycoproteins. This complementary effect is illustrated in Figure 3. It can be clearly seen that the adoption of multiple enzyme digestion could both enhance N-glycosites coverage of the identified glycoproteins and identify the glycoproteins which cannot be identified by using trypsin alone.

Among the 317 additional N-glycosites, 98 (30.9%) could not be identified by trypsin in theory because the corresponding in silico tryptic peptides are not in the mass range of 800–3500 Da. For example, the tryptic peptide carrying N-glycosite at Asn426 of splice isoform Sap-mu-0 of Proactivator polypeptide precursor is N#STK, which is too small to be identified. However, this N-glycosite can be well-identified with a peptide of LEKN#STKQEIL (MW = 1303.73) by thermolysin. On the other hand, for proteins which lack lysine and arginine residues, trypsin digestion will generate peptides with longer sequence. These peptides cannot easily be fragmented in CID to give high quality mass spectra for sequence identification. For example, MAL2 Protein has only two trypsin cleavage sites in its sequence. Its tryptic peptide carrying N-glycosite is LQG-WVMFVSVTAFESLLFLGMFSGMVAQIDANWFIFLTDAYHTFV-FVFYFGAFELLEAAATLSHDLCN#TTITGQPLLSDNQYNINVA-SIAFMTTACYGCSLGLALR (MW = 12113.93); analysis of this huge peptide needs the top-down technique by high resolution mass spectrometry. Its pepsin generated peptide, HDDLHCN#TTITGQPLL (MW = 1161.82), is very easy to be analyzed using bottom-up method. Another example further demonstrated the necessity of using multiple enzyme digestion in glycosylation analysis as shown in Figure 4. Neural-cadherin is a calcium dependent cell–cell adhesion glycoprotein that functions during gastrulation and is required for establishment of left-right asymmetry. It has 10 NXS/T motifs in the sequence of this protein and 7 were annotated as potential N-glycosites in the Swiss-Prot database. However, in the N-glycosites database

![Figure 3. Complementary effect of multiple enzyme digestion in glycoproteomics analysis of human liver tissue. (A) N-glycosites and (B) glycoprotein.](image-url)
multiple enzyme digestion. The Asn402 and Asn692 were identified with trypsin digestion; the molecular weight of the tryptic glycopeptides, N-glycosites at Asn325 and Asn402 were found on the identified glycoproteins was further explored on Expasy (www.unipep.org) which is based on the analysis of tryptic glycopeptides, N-glycosites at Asn325 and Asn692 were found to be adjacent to the N-glycosite to be accessed by trypsin, and therefore, missed cleavage sites did increase the chance to identify these N-glycosites. It should be mentioned that this makes little contribution on the coverage of glycosite in proteome level since N-glycopeptides below the preferable detection mass range takes a very small portion of tryptic N-glycopeptides as shown in Figure 2A. Besides, these N-glycosites may be identified by other proteases with different cleavage specificities in the multiple enzyme digestion approach. The benefit of multiple enzyme digestion was also demonstrated in the study of protein phosphorylation.19 With the use of subtilisin and elastase as complementary proteases, additional sites of phosphorylation were discovered that some glycosites could not be identified with trypsin because the N-glycopeptides with missed cleavage sites are too big to be identified. However, in the N-glycopeptides generated with completely in silico trypsin digestion, 4.7% of them have a molecular weight less than 800. The corresponding glycosites cannot be identified because these N-glycopeptides are too small. Missed cleavage may increase the chance for the identification of these N-glycosites because of the generation of bigger N-glycopeptides. To investigate this possibility, in silico digestion of identified trypsin N-glycopeptides with missed cleavage and calculation of the mass of these N-glycopeptides with complete digestion were performed. It was found that of the 63 identified trypsin N-glycopeptides with missed cleavage sites, 6 of corresponding N-glycopeptides with complete digestion were too small to be detected. Missed cleavage sites did increase the chance to identify these N-glycosites. It should be mentioned that this makes little contribution on the coverage of glycosite in proteome level since N-glycopeptides below the preferable detection mass range takes a very small portion of tryptic N-glycopeptides as shown in Figure 2A. Besides, these N-glycosites may be identified by other proteases with different cleavage specificities in the multiple enzyme digestion approach. The benefit of multiple enzyme digestion was also demonstrated in the study of protein phosphorylation.19 With the use of subtilisin and elastase as complementary proteases, additional sites of phosphorylation were discovered in their study.

Categorization of Identified Glycoproteins. The identification of 939 N-glycosites and 523 glycoproteins from human liver tissue was achieved in this study. The full list of identified N-glycopeptides, N-glycosites and glycoproteins is given in Supporting Information. The 523 identified N-glycoproteins from the human liver tissue were categorized by Gene Ontology (GO) and the annotations of 399 were found. The diagrams of the classification were shown in Figure 5. These proteins were found to have various functions and be involved in many biological processes. The identified glycoproteins include coagulation and complement factors, receptors, enzymes, proteases and protease inhibitors. Functional information of the identified glycoproteins was further explored on Expasy (http://www.expasy.org). Some proteins related to liver disease were found to be N-glycosylated. Hepatocyte growth factor receptor (c-Met) was found to be N-glycosylated at Asn108 which was identified by N-glycopeptides generated by both trypsin and thermolysin. This glycoprotein is receptor for hepatocyte growth factor and scatter factor which has a tyrosine-protein

Figure 4. N-glycosites of Neural-cadherin identified by pepsin (in green), by thermolysin (in yellow), and by trypsin (in gray). The in silico trypptic glycopeptides are underlined with black line.
kinase activity. The c-Met has been found to be overexpressed in hepatocellular carcinoma (HCC) compared with nontumorous liver tissue. Scavenger receptor class B member 1, which is a receptor for different ligands such as phospholipids, cholesterol ester, lipoproteins, phosphatidylserine and apoptotic cells, plays a critical role in hepatitis C virus (HCV) attachment and/or cell entry by interacting with HCV E1/E2 glycoproteins heterodimer. It was identified by pepsin with N-glycosites at Asn102 and Asn330. Identification of these glycoproteins confirms that the adoption of multiple enzyme digestion is beneficial for glycoproteomics analysis.

A distinguished result is the identification of 6 glycosyltransferases and 10 glycosidases. All the identified glycosyltransferases and glycosidases and their functions are listed in Table 1. The collaboration of these enzymes controls the synthesis and degradation of glycoproteins. Glycosyltransferases transfer sugar residues from an activated donor substrate to a growing carbohydrate group. Previous studies have found that these glycosyltransferases in mammalians are widely glycosylated, especially N-glycosylated, for example, bovine β1, 4-galactosyltransferase, which catalyzes the specific transfer of galactose from UDP-galactose to a nonreducing terminal N-acetyl-D-glucosamine, forming a β1,4 glycosidic linkage. Six glycosyltransferases were confirmed to be glycosylated in this study. It is of great importance to study the expression of glycosyltransferases in liver tissue since the structures of complex carbohydrates are altered in cancer and these changes are highly associated with invasion and metastasis of cancer cells. A good example is the overexpression of N-acetylgalactosaminyltransferases III (GnT-III) and V (GnT-V) in cancer progression and metastasis. Some of the identified glycosyltransferases are also disease related, for example, Protein O-mannosyl-transferase 1, which transfers mannosyl residues to the hydroxyl group of serine or threonine residues to form O-mannosyl glycans. It was found that mutations in POMT1 gene would deactivate this glycosyltransferase and cause Walker-Warburg syndrome. Furthermore, in some particular circumstances, glycosylation is necessary for the proper folding of these glycosyltransferases themselves and maintaining their activities. It is important to investigate the relationship between the function and modification of these glycosyltransferases, which is not so clear yet. N-glycosites were found in 10 glycosidases in this study. Glycosidases degrade glycoproteins, glycolipids and other glycoconjugates in lysosome, especially in liver. The deficiency of lysosomal enzyme may cause lysosomal storage disease which is fatal to human of all ages. For example, Schindler disease is an infantile neuroaxonal dystrophy resulting from the deficient activity of the lysosomal hydrolase, α-N-acetylgalactosaminidase (α-NAGA), which is identified with two N-glycosylation sites on Asn177 and Asn201 in this study.

Glycosylation is largely restricted to the addition of complex glycans to proteins or lipids localized on the cell surface or within luminal compartments of subcellular organelles, such as the endoplasmic reticulum (ER), Golgi apparatus, endosomes, or lysosomes. Therefore, the subcellular location of identified glycoprotein can be used to evaluate the reliability of the glycoproteome analysis. We used freely available prediction software SingalP 2.0 for the predication of signal peptides and TMHMM (version2.0) for the prediction of transmembrane regions as Zhang et al. did to predict the likely subcellular localization of all the identified glycoproteins. Of the 939 identified N-glycosites, 896 (95.4%) were from cell surface proteins, secreted proteins and transmembrane proteins, where glycosylation are expected to happen. This percentage is higher than that reported by Zhang et al. (93.4%), which indicated higher specificity for the identification of glycoproteins in this study. Low percentage of intracellular proteins was also observed in the GO annotation for the glycoproteins identified in this study. Only a few identified proteins were annotated as nuclear (9, 1.7%), cytoskeleton (5, 0.96%) and mitochondrion (4, 0.76%). Above subcellular location of glycoproteins demonstrated the reliability of glycoprotein identification, which was largely attributed to the high confidence of N-glycopeptide identification.

**Discussion**

Shotgun proteomics has become a powerful tool in comprehensive characterization of proteins in complex biological samples. It begins with the digestion of protein mixtures into peptides by using proteases, typically trypsin. The resulting
<table>
<thead>
<tr>
<th>IPI</th>
<th>protein name</th>
<th>function</th>
<th>glycosites</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPI00184851</td>
<td>Type 2 lactosamine alpha-2,3-sialyltransferase</td>
<td>Involved in the synthesis of sialyl-paragloboside, a precursor of sialyl-Lewis X determinant. Has an alpha-2,3-sialyltransferase activity toward Gal-beta1,4-GlcNAc on glycoproteins and glycolipids. Has a restricted substrate specificity; it utilizes Gal-beta1,4-GlcNAc on glycoproteins, and neolactotetraosylceramide and neolactohexaosylceramide, but not lactotetraosylceramide, lactosylceramide or asialo-GM1.</td>
<td>N308</td>
</tr>
<tr>
<td>IPI00298235</td>
<td>Splice Isoform 1 of Protein O-mannosyl-transferase 1</td>
<td>Transfers mannosyl residues to the hydroxyl group of serine or threonine residues. Coexpression of both POMT1 and POMT2 is necessary for enzyme activity; expression of either POMT1 or POMT2 alone is insufficient.</td>
<td>N471</td>
</tr>
<tr>
<td>IPI0032432</td>
<td>Splice Isoform 1 of N-acetylglucosamine-1-phosphotransferase subunits alpha/beta precursor</td>
<td>Catalyzes the formation of mannose 6-phosphate (M6P) markers on high mannose type oligosaccharides in the Golgi apparatus. M6P residues are required to bind to the M6P receptors (MPR), which mediate the vesicular transport of lysosomal enzymes to the endosomal/prelysosomal compartment.</td>
<td>N699</td>
</tr>
<tr>
<td>IPI00058192</td>
<td>Splice Isoform 1 of GDP-fucose protein O-fucosyltransferase 1 precursor</td>
<td>Catalyzes the reaction that attaches fucose through an O-glycosidic linkage to a conserved serine or threonine residue in EGF domains. Plays a crucial role in Notch signaling.</td>
<td>N160</td>
</tr>
<tr>
<td>IPI0025674</td>
<td>Dolichyl-diphosphooligosaccharide-protein glycosyltransferase 67 kDa subunit precursor</td>
<td>Essential subunit of N-oligosaccharyl transferase enzyme which catalyzes the transfer of a high mannose oligosaccharide from a lipid-linked oligosaccharide donor to an asparagine residue within an Asn-X-Ser/Thr consensus motif in nascent polypeptide chains.</td>
<td>N338</td>
</tr>
<tr>
<td>IPI0001387</td>
<td>CMP-N-acetylneuraminic acid-beta-galactosaminide-alpha-2,6-sialyltransferase</td>
<td>Transfers sialic acid from the donor of substrate CMP-sialic acid to galactose containing acceptor substrates.</td>
<td>N149, N161</td>
</tr>
<tr>
<td>IPI0012440</td>
<td>Plasma alpha-1-fucosidase precursor</td>
<td>alpha-1-Fucosidase is responsible for hydrolyzing the alpha-1,6-linked fucose joined to the reducing-end N-acetylgalactosamine of the carbohydrate moieties of glycoproteins.</td>
<td>N239</td>
</tr>
<tr>
<td>IPI0009410</td>
<td>ER degradation-enhancing alpha-mannosidase-like 3</td>
<td>Involved in endoplasmic reticulum-associated degradation (ERAD). Accelerates the glycoprotein ERAD by proteasomes. This process depends on mannose-trimming from the N-glycans. Seems to have alpha-1,2-mannosidase activity.</td>
<td>N195</td>
</tr>
<tr>
<td>IPI0012585</td>
<td>beta-Hexosaminidase beta chain precursor</td>
<td>Responsible for the degradation of GM2 gangliosides, and a variety of other molecules containing terminal N-acetyl hexosamines, in the brain and other tissues.</td>
<td>N323, N327, N84</td>
</tr>
<tr>
<td>IPI0041344</td>
<td>beta-Galactosidase precursor</td>
<td>Cleaves beta-linked terminal galactosyl residues from gangliosides, glycoproteins, and glycosaminoglycans of the brain.</td>
<td>N464, N555</td>
</tr>
<tr>
<td>IPI0041490</td>
<td>alpha-N-Acetylgalactosaminidase precursor</td>
<td>Hydrolysis of terminal nonreducing N-acetyl-d-galactosaminidase residues in N-acetyl-alpha-d-galactosaminidases.</td>
<td>N177, N201</td>
</tr>
<tr>
<td>IPI0032488</td>
<td>Splice Isoform 1 of Epididymis-specific alpha-mannosidase precursor</td>
<td>Cleaves beta-linked terminal galactosyl residues from gangliosides, glycoproteins, and glycosaminoglycans.</td>
<td>N516</td>
</tr>
<tr>
<td>IPI00298235</td>
<td>Lysosomal alpha-glucosidase precursor</td>
<td>Hydrolysis of terminal, nonreducing 1,4-linked alpha-d-glucose residues with release of alpha-d-glucose.</td>
<td>N140, N470, N882, N925</td>
</tr>
<tr>
<td>IPI0011454</td>
<td>Splice Isoform 2 of Neutral alpha-glucosidase AB precursor</td>
<td>Cleaves sequentially the 2 innermost alpha-1,3-linked glucose residues from the Glc(2)Man(9)GlcNAc(2) oligosaccharide precursor of immature glycoproteins.</td>
<td>N193</td>
</tr>
<tr>
<td>IPI0012789</td>
<td>mannosidase, alpha, class 2B, member 1 precursor</td>
<td>Necessary for the catabolism of N-linked carbohydrates released during glycoprotein turnover. Cleaves all known types of alpha-mannosidic linkages.</td>
<td>N367, N766</td>
</tr>
<tr>
<td>IPI0025869</td>
<td>alpha-Galactosidase A precursor</td>
<td>Hydrolysis of terminal, nonreducing alpha-d-galactose residues in alpha-d-galactosides.</td>
<td>N215</td>
</tr>
</tbody>
</table>

*a The functions of proteins are obtained from Expasy (http://www.expasy.org).*
peptides are separated by liquid chromatography and then analyzed by tandem mass spectrometry.\textsuperscript{38,39} However, the entire sequence of each identified protein could not be covered since only a small portion of peptides will be detected by mass spectrometry. Some of trypsin generated peptides cannot be identified either because they are too small or too hydrophilic to retain on reversed phase LC column. In an analysis of hemoglobin proteins from human blood samples, the short His-containing peptides, GHKG (at position α 57–60) and AHKG (at position β 62–65), were most likely washed off the LC column during loading sample and not detected by MS.\textsuperscript{25} On the other hand, large peptides could not be identified by most types of ion trap mass spectrometers because of the limitation in instrument resolution. In fact, peptides of a suitable mass range would be preferred for analysis in shotgun proteomics. According to MW distribution of the identified peptides in this study as shown in Figure 1, peptides smaller than 800 Da or bigger than 3500 Da are not likely to be identified by the shotgun proteomics approach. Although we do not need the complete sequence to identify a protein, this mass preference would lead to the missing of modification sites in glycoproteomics analysis when the modified peptides are out of the mass range. In silico digestion of proteins in human proteome database by trypsin indicated that many N-glycopeptides are not in the mass range of 800–3500 Da, which means that these N-glycosites could not be located if only trypsin is used.

To circumvent this problem, a multiple enzyme digestion strategy was integrated with the hydrazide capture protocol to analyze the N-glycosylation in human liver tissue in this study. Besides trypsin, two other proteases, that is, pepsin and thermolysin, with broader specificity were used. The combination of using these three proteases resulted in the identification of 939 N-glycosites; among them, 317 (33.7%) N-glycosites were only identified by pepsin and thermolysin. Among the 317 additional N-glycosites, 98 (30.9%) could not be identified by trypsin in theory because the corresponding in silico trypptic peptides are either too small or too big to detect in mass spectrometry. This study clearly demonstrated that the coverage of N-glycosylation sites could be significantly increased due to the adoption of multiple enzyme digestion.

The liver is a vital organ present in human body and is the largest one. It plays a major role in metabolism and has a variety of functions in the body, including glycogen storage, decomposition of red blood cells, plasma protein synthesis, and detoxification. Liver diseases, such as viral hepatitis, and hepatocellular carcinoma, are vital to human life. There are over 350 million hepatitis virus carriers in the world; over a million deaths per year (about 10% of all deaths in the adult over 35). More than 200 million people suffer from hepatitis virus carriers in the world; over a million deaths each year are caused by liver cancer.\textsuperscript{7,40,41} The Human Liver Proteome Project (HLPP) is one of the initiatives launched by the Human Proteome Organization (HUPO).\textsuperscript{7} Its global scientific objectives are to reveal the human liver proteome, expression profiles, modification profiles, a protein linkage map, and a proteome localization map. Post-translational modification (PTMs) profile of proteins in human liver is one of global scientific objectives of HLPP. The study of protein glycosylation in human liver is of great importance since many liver diseases are correlated with glycoproteins. For example, Deficiency of α-1-antitrypsin, an N-linked glycoprotein, is the most common genetic cause of liver disease in children. It also causes chronic liver injury and hepatocellular carcinoma in adults.\textsuperscript{42} N-acetylglucosaminyltransferases III (GnT-III) and V (GnT-V), play a pivotal role in the processing of N-linked glycoproteins, and are highly involved in cancer progression and metastasis.\textsuperscript{32} Although phosphoproteomics analysis of human liver tissue has been recently conducted in our laboratory,\textsuperscript{43} few works have been done for the analysis of N-glycosylation in human liver tissues. Totally, 523 glycoproteins with 939 N-glycosites were identified with high confidence in this study. These proteins were found to have various functions and be involved in many biological processes. The identification of glycosyltransferases and glycosidase may provide an alternative way to judge liver function by the analysis of the relative expressions of these proteins. Moreover, this large-scale analysis could contribute new N-glycosites to the database of human N-glycosites (www.unipep.org), which is a valuable resource for biomarker discovery.

**Conclusion**

This is the first large-scale analysis of human liver N-glycoproteome by shotgun proteomics approach. A combination of hydrazide capture method and multiple enzyme digestion strategy was applied for the comprehensive analysis which resulted in the identification of 939 N-glycosites and 523 glycoproteins from human liver tissue. This data set could provide valuable information for the study of liver disease and biomarker discovery. It is shown that protease with broader specificity gives good complement to that of trypsin. Use of trypsin alone resulted in identification of 622 N-glycosites, while use of pepsin and thermolysin resulted in identification of additional 317 N-glycosites. Among the 317 additional N-glycosites, 98 (30.9%) could not be identified by trypsin in theory because the corresponding in silico tryptic peptides are either too small or too big. This study clearly demonstrated that the coverage of N-glycosylation sites could be significantly increased due to the adoption of multiple enzyme digestion.

**Acknowledgment.** Financial supports from the National Natural Sciences Foundation of China (No. 20675081, 20735004), the China State Key Basic Research Program Grant (2005CB522701, 2007CB914104), the China High Technology Research Program Grant (2006AA02A309), the Knowledge Innovation program of CAS (KJCX2.YW.H09, KSCX2.YW.079) and the Knowledge Innovation program of DICP to H.Z. and National Natural Sciences Foundation of China (No. 20605022, 90713017) to M.Y. are gratefully acknowledged.

**Supporting Information Available:** Supporting table includes all the identified glycopeptides, N-glycosites and glycoproteins identified by trypsin, thermolysin and pepsin, respectively. This material is available free of charge via the Internet at http://pubs.acs.org.

**References**


\textsuperscript{2} Varki, A. Biological roles of oligosaccharides—all of the theories are correct. \textit{Glyobiology} 1993, 3 (2), 97–130.


\textsuperscript{5} Zhang, H.; Yi, E. C.; Li, X. J.; Mallick, P.; Kelly-Spratt, K. S.; Masselon, C. D.; Camp, D. G.; Smith, R. D.; Kemp, C. J.; Aebersold, R. Liver Proteome Project (HLPP) is one of the initiatives launched by the Human Proteome Organization (HUPO). Its global scientific objectives are to reveal the human liver proteome, expression profiles, modification profiles, a protein linkage map, and a proteome localization map. Post-translational modification (PTMs) profile of proteins in human liver is one of global scientific objectives of HLPP. The study of protein glycosylation in human liver is of great importance since many liver diseases are correlated with glycoproteins. For example, Deficiency of α-1-antitrypsin, an N-linked glycoprotein, is the most common genetic cause of liver disease in children. It also causes chronic liver injury and hepatocellular carcinoma in adults. N-acetylglucosaminyltransferases III (GnT-III) and V (GnT-V), play a pivotal role in the processing of N-linked glycoproteins, and are highly involved in cancer progression and metastasis. Although phosphoproteomics analysis of human liver tissue has been recently conducted in our laboratory, few works have been done for the analysis of N-glycosylation in human liver tissues. Totally, 523 glycoproteins with 939 N-glycosites were identified with high confidence in this study. These proteins were found to have various functions and be involved in many biological processes. The identification of glycosyltransferases and glycosidase may provide an alternative way to judge liver function by the analysis of the relative expressions of these proteins. Moreover, this large-scale analysis could contribute new N-glycosites to the database of human N-glycosites (www.unipep.org), which is a valuable resource for biomarker discovery.

**Conclusion**

This is the first large-scale analysis of human liver N-glycoproteome by shotgun proteomics approach. A combination of hydrazide capture method and multiple enzyme digestion strategy was applied for the comprehensive analysis which resulted in the identification of 939 N-glycosites and 523 glycoproteins from human liver tissue. This data set could provide valuable information for the study of liver disease and biomarker discovery. It is shown that protease with broader specificity gives good complement to that of trypsin. Use of trypsin alone resulted in identification of 622 N-glycosites, while use of pepsin and thermolysin resulted in identification of additional 317 N-glycosites. Among the 317 additional N-glycosites, 98 (30.9%) could not be identified by trypsin in theory because the corresponding in silico tryptic peptides are either too small or too big. This study clearly demonstrated that the coverage of N-glycosylation sites could be significantly increased due to the adoption of multiple enzyme digestion.
Glycoproteomics Analysis of Human Liver Tissue


PR8008012