

# A Strategy for Precise and Large Scale Identification of Core Fucosylated Glycoproteins\*<sup>S</sup>

Wei Jia†§¶, Zhuang Lu†¶||, Yan Fu¶||\*\*, Hai-Peng Wang\*\*, Le-Heng Wang\*\*, Hao Chi\*\*, Zuo-Fei Yuan\*\*, Zhao-Bin Zheng†, Li-Na Song†, Huan-Huan Han†, Yi-Min Liang†, Jing-Lan Wang†, Yun Cai†, Yu-Kui Zhang||, Yu-Lin Deng||, Wan-Tao Ying† ††, Si-Min He\*\*§§, and Xiao-Hong Qian†¶||

**Core fucosylation (CF) patterns of some glycoproteins are more sensitive and specific than evaluation of their total respective protein levels for diagnosis of many diseases, such as cancers. Global profiling and quantitative characterization of CF glycoproteins may reveal potent biomarkers for clinical applications. However, current techniques are unable to reveal CF glycoproteins precisely on a large scale. Here we developed a robust strategy that integrates molecular weight cutoff, neutral loss-dependent MS<sup>3</sup>, 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 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. *Molecular & Cellular Proteomics* 8:913–923, 2009.**

Glycoproteins are implicated in a wide range of biological processes such as fertilization, development, the immune response, cell signaling, and apoptosis. Altered glycosylation patterns can affect the conformations of glycoproteins and their functions and interactions with other molecules (1, 2).

From the †State Key Laboratory of Proteomics-Beijing Proteome Research Center-Beijing Institute of Radiation Medicine, No. 33 Life Science Park Road, Changping District, Beijing 102206, China, §Institute of Biophysics, Chinese Academy of Sciences, No. 15 Datun Road, Chaoyang District, Beijing 100101, China, ||Beijing Institute of Technology, No. 5 South Zhongguancun Street, Haidian District, Beijing 100081, China, and \*\*Institute of Computing Technology, Chinese Academy of Sciences, No. 6 Kexueyuan South Road, Beijing 100190, China

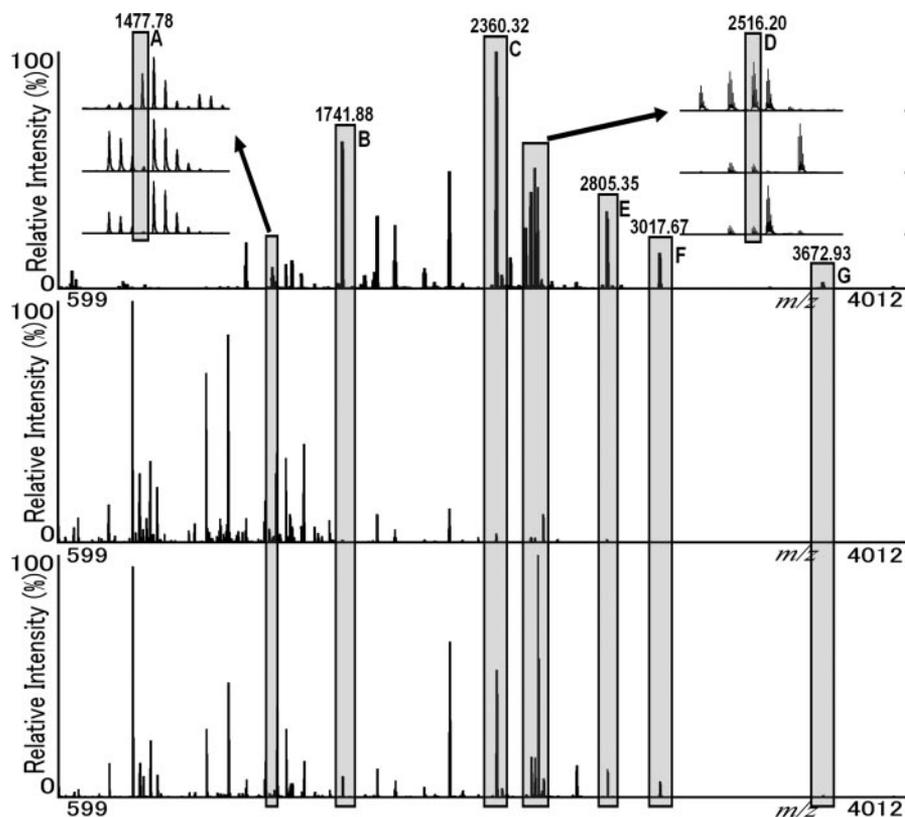
Received, November 5, 2008, and in revised form, January 7, 2009  
Published, MCP Papers in Press, January 12, 2009, DOI 10.1074/mcp.M800504-MCP200

Abnormal glycosylation has been demonstrated in many pathological processes. Targeted glycosylation research is considered increasingly important as a way to find novel therapeutic approaches (2, 3), and core fucosylation (CF)<sup>1</sup> glycoproteomics has attracted particularly great attention (4, 5). Previous reports show that CF glycoproteins are involved in many important physiological processes, such as transforming growth factor- $\beta$ 1 (6) and epidermal growth factor signaling pathways (7). They also play key roles in many pathological processes, such as hepatocellular carcinoma (HCC) (8, 9), pancreatic cancer (10, 11), lung cancer (6, 12), ovarian cancer (13), and prostate cancer (14). Moreover the CF patterns of several glycoproteins have been reported to serve as more sensitive and specific biomarkers than their total respective protein levels (8, 9, 15, 16). The combination of a biomarker panel of CF glycoproteins is expected to serve as a more reliable diagnostic standard (13).

Glycoproteomics research has been conducted for several years and has led to the generation of many effective evaluation methods. Most of these methods use lectin or the chemical reagent hydrazide to enrich glycopeptides. The oligosaccharide chains are then completely released by treatment of the glycopeptides with peptide-*N*-glycosidase F. Finally the deglycosylated peptides and the deglycosylation sites are identified by tandem mass spectrometric analysis (17, 18). Although impressive results have been attained, this commonly used strategy is not an ideal choice for CF glycoproteins research. First, the enrichment specificity of lectin is not satisfactory (19) as hydrazide chemical reactions irreversibly destroy glycan structures, particularly fucose tags. Second, the deglycosylation site is determined by the 0.9840-Da mass shift caused by the asparagine to aspartic acid transfer; its confidence can be compromised by deamination of the Asn. Besides that, the CF site can no longer be distinguished from other glycosylation sites in the same glycoprotein. Thus, the ideal way to precisely identify CF glycoproteins on a large

<sup>1</sup> The abbreviations used are: CF, core fucosylation; HCC, hepatocellular carcinoma; rhEPO, recombinant human erythropoietin; RP, reversed phase; S<sub>2</sub>, symbol ion 2; S<sub>3</sub>, symbol ion 3; HS, Heremans-Schmid; SCX, strong cation exchange; LTQ, linear trap quadrupole.

**FIG. 1. The efficiency of the ultrafiltration method for enriching glycopeptide.** MS spectra from ultrafiltration experiments are shown with the retentate fraction (*top*), filtrate fraction (*middle*), and untreated fraction (*bottom*). Glycopeptide C<sup>#</sup>GLVPVLAENYN<sup>\*</sup>K (A) from apotransferrin only appeared in the retentate fraction. LC<sup>#</sup>PDC<sup>#</sup>PLLA-PLN<sup>\*</sup>DSR (B), VVHAVEVALATFNAESN<sup>\*</sup>-GSYLQLVEISR (F), and RPTGEVYDIEID-TLETTC<sup>#</sup>HVLDP TPLAN<sup>\*</sup>C<sup>#</sup>SVR (G) were from fetuin; GQALLVN<sup>\*</sup>SSQPWEPLQ-LHVDK (C) and EAEN<sup>\*</sup>ITGTG<sup>#</sup>AEHC<sup>#</sup>SLNEN<sup>\*</sup>ITVPDTK (E) were from rhEPO; QQQHLFGSN<sup>\*</sup>VTDC<sup>#</sup>SGNFC<sup>#</sup>LFR (D) was from apotransferrin. \*, annotated glycosite; #, carbamidomethylation.



scale is to provide direct evidence for the existence of CF modification. Traditional approaches, such as lectin blots, are not sufficiently powerful to meet this requirement. Instead recent advancements in high end MS-based techniques have ignited the hope to reach this challenging goal (20, 21).

Our group has developed an innovative and systematic strategy for the precise and large scale identification of CF glycoproteins. Several steps were taken leading up to the development of our strategy. 1) We established a novel enrichment step for CF glycopeptides, combining the use of lectin for CF glycoprotein enrichment with ultrafiltration for further enrichment of glycopeptide. Glycopeptide enrichment by ultrafiltration based on molecular weight cutoff technology has the added merit of integrating enrichment, desalting, and concentration into a one-step operation. 2) We established a neutral loss-dependent MS<sup>3</sup> scan method that specifically captures partially deglycosylated CF glycopeptides (with fucosyl-*N*-acetylglucosamines residue retained). In MS<sup>3</sup>, the intensity distribution of the fragment peaks is much more homogeneous, and there are fewer theoretical fragment ions and interfering peaks than in MS<sup>2</sup>. 3) We established a novel database-independent candidate spectrum-filtering method for selecting partially deglycosylated CF glycopeptides and a spectrum optimization method. By introducing several strict and appropriate criteria into a scoring system, high quality candidate spectra can be selected before searching the database, which not only increases the database search efficiency but also improves the identification credibility. Further-

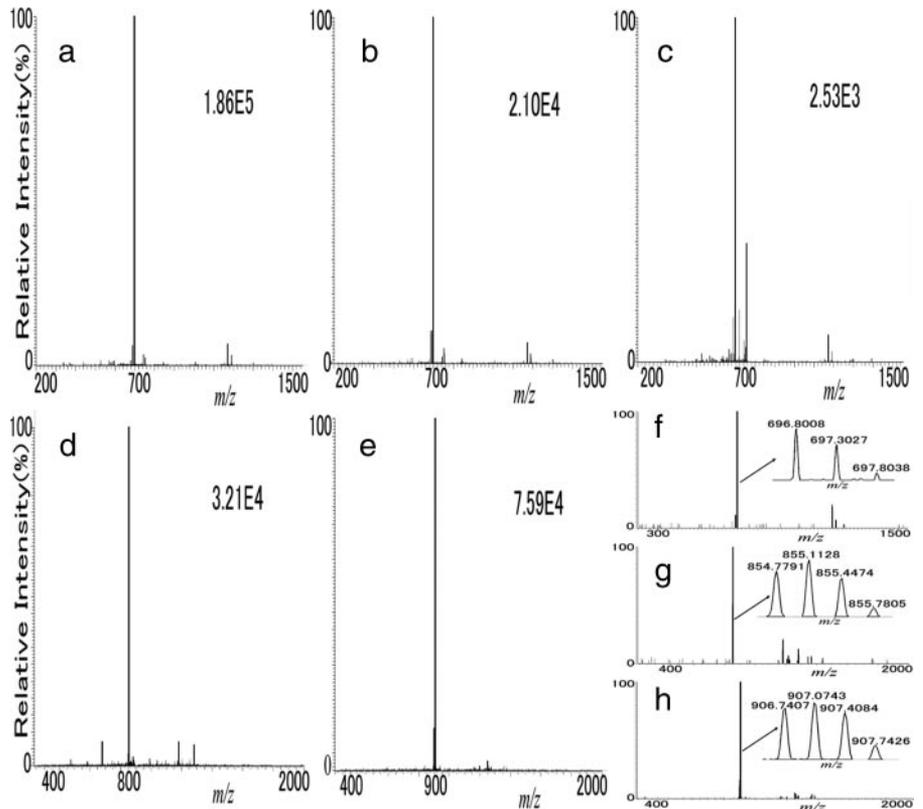
more by statistically analyzing candidate spectra, some important glycan-related fragmentation patterns were revealed. Based on these observations, many kinds of interfering peaks due to glycan fragmentation that are always very intensive and would decrease the accuracy of peptide scoring can be localized and removed from the spectra. This treatment can effectively increase the number of identifications through database searching or *de novo* analysis.

The efficacy of this strategy was testified by implementing it on both healthy and HCC plasma. Respectively, 105 and 106 CF sites were identified from 72 and 79 glycoproteins, including 19 annotated potential glycosylation sites and 25 novel ones. This study holds promise for the large scale determination of core fucosylated biomarker panels from clinical samples, either body fluids or tissue biopsies.

#### EXPERIMENTAL PROCEDURES

**Materials**—The apotransferrin, fetuin, ribonuclease B, endoglycosidase F3, formic acid, TFA,  $\alpha$ -cyano-4-hydroxycinnamic acid, and *Lens culinaris* lectin (agarose conjugate, saline suspension) were purchased from Sigma, methyl- $\alpha$ -D-mannopyranoside was purchased from Fluka (St. Louis, MO), and sodium-3-[(2-methyl-2-undecyl-1,3-dioxolan-4-yl)methoxy]-1-propanesulfonate (RapiGest<sup>TM</sup> SF) was purchased from Waters. Sequencing grade porcine trypsin was purchased from Promega (Madison, WI); IgG was purified by use of a HiTrap Protein G HP column from GE Healthcare. The PD-10 desalting column was also from GE Healthcare. Deionized water was produced by a Milli-Q A10 system from Millipore (Bedford, MA). HPLC-grade quality ACN was purchased from J. T. Baker Inc. Iodoacetamide and DTT were obtained from ACROS. The Handee

**FIG. 2. The neutral loss peaks in MS<sup>2</sup> spectra of partially deglycosylated CF glycopeptides.** The intensities of the highest peaks are several times higher than that of the second most intense peak in all of these MS<sup>2</sup> spectra in the ion trap, resulting from loss of the fucose residue in CID. *a*, *b*, and *c* are MS<sup>2</sup> spectra from the same partially deglycosylated CF glycopeptide, EEQYJSTYR (from human IgG). Intensities of the base peaks were 1.86e5, 2.10e4, and 2.53e3, respectively. *d* and *e* are MS<sup>2</sup> spectra of simplified CF glycopeptides GQA-LLVJSSQPWEPLQLHVDK (intensity, 3.21e4; from rhEPO) and QQQLHFG-SJVTDC<sup>#</sup>SGNFC<sup>#</sup>LFR (intensity, 7.59e4; from apotransferrin). The MS<sup>2</sup> spectra in FT-ICR were collected to check the identities of the strongest peaks: *f* for IgG, *g* for *d*, and *h* for *e*. J, CF site; #, carbamidomethylation.



mini spin column kit was purchased from Pierce. The C<sub>18</sub> ZipTip and Microcon YM-3 were purchased from Millipore. Recombinant human erythropoietin (rhEPO) was a gift from the National Institute for the Control of Pharmaceutical and Biological Products. Healthy human plasma (0.8 ml for each experiment) was obtained from a healthy donor. Samples of hepatocellular carcinoma plasma were mixed from eight patients with 0.1 ml from each person.

**IgG Extraction**—Plasma was supplemented with IgG binding buffer (20 mM sodium phosphate, pH 7.0), and then IgG was depleted by trapping on a column of HiTrap Protein G. The unbound samples were desalted by a PD-10 column.

**Lectin Affinity**—Samples were supplemented with 1.6 ml of lectin binding buffer (20 mM Tris-buffered saline, 0.3 M NaCl, 1 mM MnCl<sub>2</sub>, 1 mM CaCl<sub>2</sub>, pH 7.4). The samples were incubated for 16 h at 4 °C with *L. culinaris* lectin in a spin column (about 300 μl of lectin-agarose and 400 μl of sample in each column). After unbound proteins were removed by washes with binding buffer, the CF glycoproteins were eluted with elution buffer (binding buffer supplemented with 200 mM α-D-methylmannoside), then desalted (by PD-10 column), and lyophilized.

**Reduction, Alkylation, and Trypsin Digestion**—Samples were dissolved in 200 μl of solution that contained 8 M urea and 5 mM DTT and were reduced at 37 °C for 4 h. Then iodoacetamide was added to the solution (final concentration, 15 mM), which was then further incubated for 1 h in darkness at room temperature. Afterward 50 mM NH<sub>4</sub>HCO<sub>3</sub> was added to reduce the concentration of urea below 1 M, and sequencing grade trypsin was added at a ratio of enzyme to protein of 1:50. The mixture was then vortexed and incubated at 37 °C overnight. 0.1% RapiGest SF was used instead of urea for protein denaturation in the repeat experiment of healthy and HCC plasma. TFA was added to the digested protein samples (final TFA concentration was 0.5%, pH < 2), and the samples were incubated at 37 °C for 45 min. Finally the acid-treated samples were centrifuged at 13,000 rpm for 10 min, and the supernatants were collected.

**Enrichment, Desalting, and Concentration of Glycopeptides**—Tryptic digests were pipetted into Microcon YM-3 centrifugal filter devices. The absolute amount of glycoprotein in the digests was between 200 and 300 μg for each filter device, and the sample volume was diluted to 500 μl for each filter device. The samples were centrifuged at 8000 × *g* to reduce the sample volume from 500 μl to about 20 μl; this required about 3 h. Then 450 μl of deionized water were added to the reservoir and centrifuged at 8000 × *g* for 3 h; this was repeated twice. After that, the retentate fraction was transferred to a vial, and the reservoir was thrice washed with 20% ACN. All of the retentate fractions and wash solutions were pooled and lyophilized.

**Endoglycosidase F3 Digestion**—Glycopeptides were resuspended in 100 μl of sodium acetate solution (50 mM, pH 4.5) and then incubated with endoglycosidase F3 overnight at 37 °C. Ammonium acetate (50 mM, pH 4.5) was used instead of the sodium acetate in the repeat experiments of healthy and HCC plasma.

**Strong Cation Exchange (SCX) Peptide Fractionation**—10% enriched samples were directly analyzed with RP HPLC-MS two times. Other enriched CF glycopeptides were reconstituted with 300 μl of 5 mM ammonium chloride, pH 3.0, 25% acetonitrile and fractionated by SCX chromatography on a BioBasic SCX 250 × 4.6-mm column (Thermo Fisher). The particle size of the column was 5 μm and pore size was 300 Å. The separations were performed at a flow rate of 0.5 ml/min using the Elite HPLC system, and mobile phases consisted of 5 mM ammonium chloride, pH 3.0, 25% acetonitrile (A) and 500 mM ammonium chloride, pH 3.0, 25% acetonitrile (B). After loading 300 μl of sample onto the column, the gradient was maintained at 100% A for 10 min. Peptides were then separated using a gradient of 0–15% B over 1 min followed by a gradient of 15–50% B over 49 min. Then the gradient was changed to 50–100% over 5 min. The gradient was then held at 100% B for 5 min. A total of 15 fractions were collected, and each fraction was dried under vacuum.

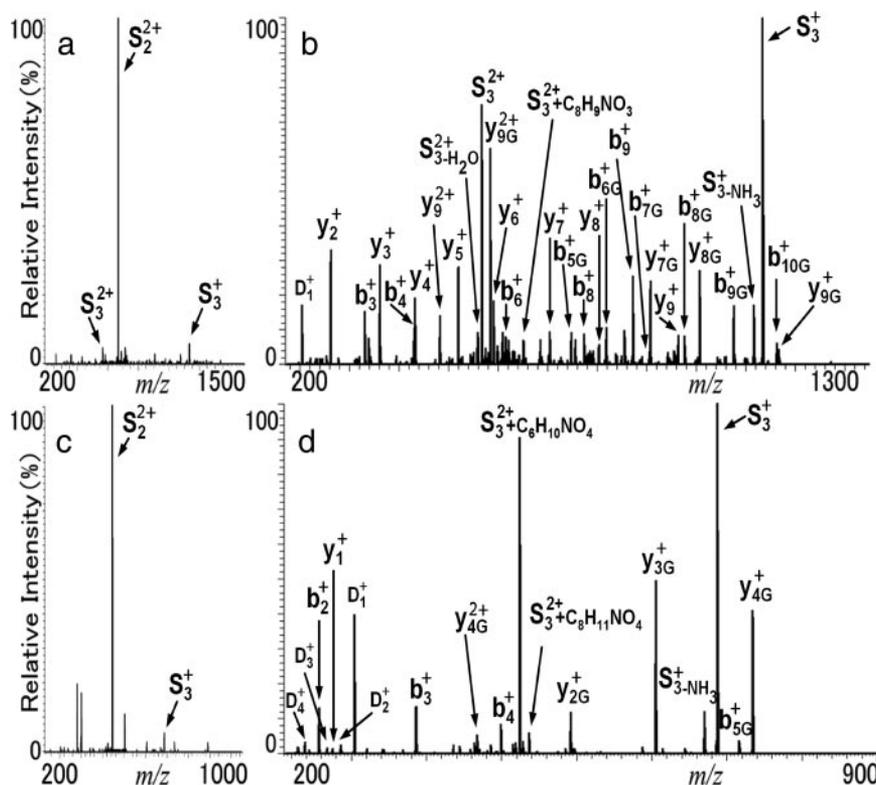


FIG. 3. **MS<sup>2</sup> and MS<sup>3</sup> spectra of fucosyl-GlcNAc-attached peptides.** The peak intensity distribution of the MS<sup>3</sup> spectrum is much more homogeneous than that of MS<sup>2</sup>, so better peptide sequence information can be obtained; the direct assignment of CF glycosites can be deduced from the b-type and y-type ion series attached with a GlcNAc residue in MS<sup>3</sup>. *a* and *b* are MS<sup>2</sup> and MS<sup>3</sup> spectra of GLC<sup>#</sup>VJASAVSR from insulin-like growth factor-binding protein 3, respectively. The peaks of b-type and y-type ions with or without GlcNAc residues appear synchronously and frequently, such as  $y_7^+$  and  $b_6^+$ . *c* and *d* are MS<sup>2</sup> and MS<sup>3</sup> spectra of a candidate that was analyzed *de novo*, respectively. The resulting *de novo* sequence GVEIJR (because the  $m/z$  of ion  $b_1$  is too low to detect, the sequence of the first two residues can also be “VG,” and “I” can also be “L” because of their same mass) was not found in the peptide database of tryptic digests (J located in the sequon NX(S/T/C) where X is any amino acid except proline).  $D_1$ ,  $C_8H_{14}NO_5$  (GlcNAc);  $D_2$ ,  $C_8H_{12}NO_4$ ;  $D_3$ ,  $C_8H_{10}NO_3$ ;  $D_4$ ,  $C_6H_{10}NO_3$ ;  $D_5$ ,  $C_7H_8NO_2$ . The  $y_{7G}^+$  identifies the GlcNAc residue with the same sequence as  $y_7^+$ . J, CF site; #, carbamidomethylation.

**RP HPLC-MS<sup>n</sup> Analysis**—RP HPLC-MS<sup>n</sup> experiments were performed on an LTQ-FT mass spectrometer (Thermo Fisher) equipped with a nanospray source and Agilent 1100 high performance liquid chromatography system (Agilent Technologies). Peptide mixes were separated on a fused silica microcapillary column with an internal diameter of 75  $\mu$ m and an in-house prepared needle tip with an internal diameter of  $\sim$ 15  $\mu$ m. Columns were packed to a length of 10 cm with a  $C_{18}$  reversed phase resin (GEAgel  $C_{18}$  SP-300-ODS-AP; particle size, 5  $\mu$ m; pore size, 300  $\text{\AA}$ ; Jinouya, Beijing, China). Separation was achieved using a mobile phase from 1.95% ACN, 97.95% H<sub>2</sub>O, 0.1% FA (phase A) and 79.95% ACN, 19.95% H<sub>2</sub>O, 0.1% FA (phase B), and the linear gradient was from 5 to 50% buffer B for 80 min at a flow rate of 300 nl/min. The LTQ-FT mass spectrometer was operated in the data-dependent mode. A full-scan survey MS experiment ( $m/z$  range from 400 to 2000; automatic gain control target, 5e5 ions; resolution at 400  $m/z$ , 100,000; maximum ion accumulation time, 750 ms) was acquired by the FT-ICR mass spectrometer, and the five most abundant ions detected in the full scan were analyzed by MS<sup>2</sup> scan events (automatic gain control target, 1e4 ions; maximum ion accumulation time, 200 ms). The scan model of MS<sup>2</sup> was set as the profile. An MS<sup>3</sup> spectrum was automatically collected when one of the three most intense peaks from the MS<sup>2</sup> spectrum corresponded to a neutral loss event of 73.0290  $m/z$ , 48.6860  $m/z$ , or 36.5145  $m/z$  (charges of parent ions were not collected). The normalized collision energy was 35.

**On-line Two-dimensional LC-MS<sup>n</sup>**—The autosampler was used to inject samples onto the SCX column (BioX-SCX, 5 cm) after which they were eluted onto a trap column using a stepwise gradient of 0, 20, 30, 40, 50, 60, 70, 80, 90, and 100% SCX-B. Peptides on the trap column were desalted and then eluted onto the RP column and into the mass spectrometer (the same method as RP HPLC-MS<sup>n</sup> analysis, but the linear gradient was from 5 to 50% buffer B for 120 min). Mobile phase buffer for SCX-A was 10 mM citric ammonia buffer, pH 3.0, and mobile phase buffer for SCX-B was 50 mM citric ammonia buffer, pH 8.5. Experiments of HCC samples were analyzed by this system (Eksigent NanoLC-2D) and repeated one time.

**Database Search and Analysis**—Dta files were generated by BioWorks 3.2 with default parameters and then treated by spectrum-filtering and spectrum optimization tools in pFind 2.1 Studio. The candidate spectra of MS<sup>3</sup> were searched against UniProt Knowledgebase Release 12.6 (human, 76,137 entries; UniProt Knowledgebase Release 12.6 consists of UniProtKB/Swiss-Prot Release 54.6 of December 4, 2007 and UniProtKB/TrEMBL Release 37.6 of December 4, 2007) using the pFind 2.1 search engine. The database was modified by substituting the letter N in glycosylation sequence NX(S/T/C) with J, which was defined to have the same mass as Asn (21), and then the target and reversed decoy database were combined for the search. Carbamidomethylation was considered for all Cys residues. Variable modifications contained oxidation of Met residues, carbamidomethylation and carbamylation (carbamylation was only considered as a

variable modification in experiments that used urea as the protein denature reagent) of peptide N-terminal and Lys residues, and a 203.0794-Da variable addition to J residues. At most, two missed tryptic cleavage sites were allowed. Tolerance of parent ions was  $\pm 20$  ppm, and tolerance of fragment ions was  $\pm 0.5$   $m/z$  for the primary search. The final identified results had a 1% false-positive rate (22), and the tolerance for parent ions was  $\pm 10$  ppm.

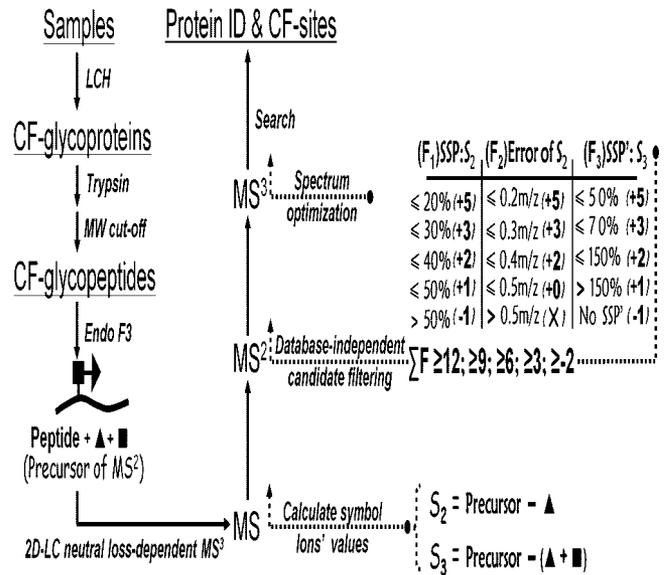
**MALDI-TOF MS Analysis**—After desalting with the C<sub>18</sub> ZipTip, all of the samples were mixed 1:9 with 5 mg/ml  $\alpha$ -cyano-4-hydroxycinnamic acid in 50% acetonitrile supplemented with 0.1% TFA, and 0.5  $\mu$ l of sample was applied to the MALDI target plate. The mass spectra were obtained using a 4800 Proteomics Analyzer MALDI-TOF/TOF instrument (Applied Biosystems). Prior to analysis, the mass spectrometer was externally calibrated with seven peptides obtained from tryptic digest of myoglobin. The  $m/z$  range of the MS scan was from 600 to 4000. Mass spectra were acquired in the positive reflector mode.

RESULTS AND DISCUSSION

**Core-fucosylated Glycopeptide Enrichment from Plasma**—Robust and convenient operation procedures were established to obtain partially deglycosylated CF glycopeptides. After IgG depletion, plasma proteins were mixed with *L. culinaris* lectin to enrich for the CF glycoproteins. Binding proteins were digested by trypsin, and the resulting glycopeptides were enriched through a molecular weight cutoff technique. *N*-Linked glycopeptides usually have larger molecular weights than non-glycopeptides (19, 23); therefore, an ultrafiltration membrane with a molecular mass limit of 3000 Da was utilized to enrich for glycopeptides. This step integrates enrichment, desalting, and concentration into one operation. Glycopeptides were then treated with endoglycosidase F3, which specifically cleaves the glycosidic bond between the two proximal *N*-acetylglucosamines (GlcNAc) and leaves the fucosyl-GlcNAc residues on the peptides. Endoglycosidase F3 was chosen here for treating CF glycoprotein because a large number of the glycans of plasma glycoproteins have biantennary structure, which is a more efficient substrate for endoglycosidase F3 (24). For other structures, such as tetraantennary and other bulky glycans, the reactivity of endoglycosidase F3 is poor, so there may need to be additional evaluation to choose the proper glycosidase for other kinds of samples like tissue biopsies.

A tryptic peptide mixture from four standard glycoproteins, apotransferrin, fetuin, rhEPO, and ribonuclease B, was used to illustrate the efficiency of the ultrafiltration method (Fig. 1). Half of this tryptic peptide mixture was directly treated with peptide-*N*-glycosidase F (untreated sample); the other half was separated by ultrafiltration into a retentate fraction (high molecular weight) and a filtrate fraction (low molecular weight), and then both fractions were treated with peptide-*N*-glycosidase F. The deglycosylated glycopeptides were detected by the +0.984-Da mass drift on Asn to Asp.

In total, eight *N*-glycopeptides were reported for four glycoproteins. Six of these glycopeptides were directly found in untreated samples by MALDI-TOF MS. However, in addition to these six glycopeptides, one more glycopeptide



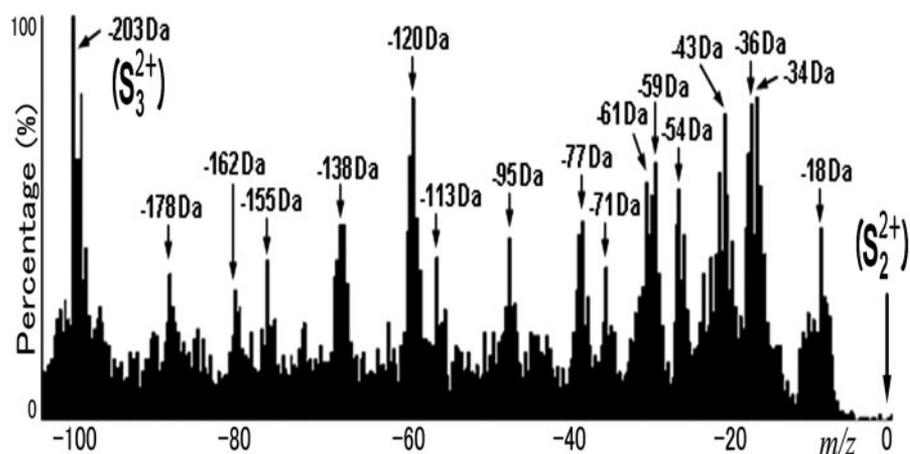
**FIG. 4. The process of the strategy for CF glycoprotein identification.** CF glycoprotein identification was achieved through enrichment of CF glycopeptides, partial deglycosylation of CF glycopeptides, HPLC neutral loss-dependent MS<sup>3</sup>, candidate spectrum filtering, spectrum optimization, and database searching. *F*<sub>1</sub> identifies the intensity ratio of the second strongest peaks (logogram: second strong peak (SSP), which does not contain different states for S<sub>2</sub>, such as a different charge state or states of H<sub>2</sub>O and NH<sub>3</sub> loss) to S<sub>2</sub>; *F*<sub>2</sub> identifies the difference between the calculated and experimental *m/z* of S<sub>2</sub>; *F*<sub>3</sub> identifies the intensity ratio of the second strongest peak (logogram: SSP') to S<sub>3</sub> within the range of the S<sub>3</sub> monoisotopic peak  $\pm 3$   $m/z$ . Information on different charge state ions of S<sub>3</sub> is considered, and the better result is recorded. Additionally the absolute intensities of S<sub>2</sub> and S<sub>3</sub> are required to be higher than 500 and 50, respectively. As shown, different scores correspond to different signal qualities. The confidence of the spectrum is sorted into five ranks by total score. ▲, fucose residue; ■, GlcNAc residue. 2D, two-dimensional; Endo, endoglycosidase; LCH, *L. culinaris* lectin.

(CGLVPVLAENYN\*K from apotransferrin; N\* represents the annotated glycosite) was detected in the retentate fraction. The relative intensities of all deglycosylated glycopeptides were heightened compared with the untreated sample. In the untreated sample, the failure to detect CGLVPVLAENYN\*K is ascribed to suppression by a non-glycopeptide with similar mass. In the filtrate fraction, the relative intensity of deglycosylated glycopeptides decreased to a very low level, illustrating that few glycopeptides were lost. One reported glycopeptide was not detected in the three fractions (N\*LTK from ribonuclease B). One possible reason is that its sequence is too short to detect.

**Development of Neutral Loss-dependent MS<sup>3</sup> Scan Method**—A neutral loss-dependent MS<sup>3</sup> method specifically designed for partially deglycosylated CF glycopeptides was developed. During CID, the glycosidic bond that links the two remaining sugars is prone to breakage compared with the other bonds (25). In our experiments on three partially deglycosylated CF glycopeptides, the highest peaks in the MS<sup>2</sup>

FIG. 5. Frequency histogram of intact and partial GlcNAc loss peaks in candidate MS<sup>3</sup> spectra of charge 2. The  $m/z$  values of S<sub>2</sub> were set as 0  $m/z$ . Offsets with high peak frequencies reveal potential masses of neutral losses that frequently occur on peptide-attached GlcNAc residues. The possible loss groups are shown in the table.

S <sub>2</sub> Loss	203	178	162	155	138	120	113	95	77
	C <sub>8</sub> H <sub>13</sub> NO <sub>5</sub>	C <sub>6</sub> H <sub>12</sub> NO <sub>5</sub>	C <sub>6</sub> H <sub>10</sub> O <sub>5</sub>	138+NH <sub>3</sub>	120+H <sub>2</sub> O	C <sub>4</sub> H <sub>8</sub> O <sub>4</sub>	59+3H <sub>2</sub> O	59+2H <sub>2</sub> O	59+H <sub>2</sub> O
S <sub>3</sub> Add	-	C <sub>2</sub> H	C <sub>2</sub> H <sub>3</sub> N			C <sub>4</sub> H <sub>5</sub> NO			C <sub>6</sub> H <sub>6</sub> O <sub>3</sub>
S <sub>2</sub> Loss	71	61	59	54	43	36	34	18	0
	54+NH <sub>3</sub>	43+H <sub>2</sub> O	C <sub>2</sub> H <sub>5</sub> NO	3H <sub>2</sub> O	C <sub>2</sub> H <sub>3</sub> O	2H <sub>2</sub> O	2NH <sub>3</sub>	H <sub>2</sub> O	-
S <sub>3</sub> Add		C <sub>6</sub> H <sub>8</sub> NO <sub>3</sub>	C <sub>6</sub> H <sub>8</sub> O <sub>4</sub>		C <sub>6</sub> H <sub>10</sub> NO <sub>4</sub>	C <sub>8</sub> H <sub>9</sub> NO <sub>3</sub>		C <sub>8</sub> H <sub>11</sub> NO <sub>4</sub>	C <sub>8</sub> H <sub>13</sub> NO <sub>5</sub>



spectra all resulted from subtraction of 146 Da (mass of the fucose residue) from the parent ions that had the same charge state as the corresponding parent ions (Fig. 2). Based on this trait, a neutral loss-dependent MS<sup>3</sup> scan method was utilized as an automatic event in the LTQ-FT mass spectrometer: MS<sup>3</sup> spectra were automatically collected when one of the three most intense peaks from the MS<sup>2</sup> spectrum corresponded to a neutral loss event of the fucose residue mass. MS<sup>3</sup> spectra were generated from fragmentation of the GlcNAc-attached peptides. Compared with the MS<sup>2</sup> spectra, which were generated from fragmentation of the fucosyl-GlcNAc-attached peptides, the MS<sup>3</sup> spectra have three remarkable advantages. 1) They have better spectrum quality: the peak intensity distribution of the MS<sup>3</sup> spectrum is much more homogeneous. This is beneficial because there are more fragment ion signals with good signal to noise ratios. 2) They have simpler spectrum information: the number of theoretical fragment ions in the MS<sup>3</sup> spectrum is fewer. This makes the algorithm for peak matching simpler and easier. 3) They have clearer spectrum signals: two parent ion selections (from MS to MS<sup>2</sup> and from MS<sup>2</sup> to MS<sup>3</sup>) reduce the probability of collecting interference signals adjacent to parent ions in the full scan (Fig. 3). In addition, direct assignment of CF glycosites can be deduced from the b-type and y-type ions series attached with a GlcNAc residue, providing much higher confidence levels of glycosite assignment compared with the 0.984-Da mass shift method. It should be noted

that the retained intact GlcNAc residues were found to be lost from the b and y ions (Fig. 3); therefore, these kinds of special product ions must be considered in addition to GlcNAc attached b and y ions when searching the database. This observation was taken into account for peptide scoring in the pFind 2.1 search engine (26–28). Compared with other popular software tools, pFind discovered more results (supplemental Data 1).

**Development of Candidate Spectrum-filtering and Spectrum Optimization Methods**—Due to the complexity of real samples and the massive spectra generated in these large scale glycopeptide analyses, more professional and specialized processing methods are absolutely necessary. Here a database-independent method for discovery of spectra of partially deglycosylated CF glycopeptides was developed. Two kinds of ions in MS<sup>2</sup> were scrutinized and used to judge whether the precursor was a CF glycopeptide: ions of a peptide attached to a GlcNAc residue (symbol ion 2, logogram: S<sub>2</sub>, attained from the breakage of the glycosidic bond between the remaining two monosaccharide residues) and ions of a pure peptide (symbol ion 3, logogram: S<sub>3</sub>, obtained from fragmentation between the GlcNAc and the Asn residue of the peptide). By introduction of the highly accurate parent ion mass from a full scan (recorded in FT-ICR), we can calculate the  $m/z$  of symbol ions. Next according to the quality of the symbol ions in MS<sup>2</sup>, several criteria were established to sort out the spectra. First of all the strongest peak in MS<sup>2</sup>

Precise and Large Scale CF Glycoprotein Identification

TABLE I

Bold "J" indicates the CF site. Bold "j" indicates the possible CF site. ADAM, a disintegrin and metalloprotease; ADAMTS, a disintegrin and metalloprotease with thrombospondin type 1 motifs.

Protein name	UniProt	Core fucosylated glycopeptide	Site
ADAMTS-13	Q76LX8	WVJYSCLDQAR	707 <sup>a</sup>
Afamin	P43652	JCCNTENPPGCYR YAEDKFJETTEK <sup>c</sup>	383 <sup>b</sup> 402
$\alpha_1$ -Antitrypsin	P01009	YLGJATAIFFLPDEGK <sup>c</sup>	271
$\alpha_2$ -HS-glycoprotein	P02765	VCQDCPLLAPLJDTR <sup>c</sup>	156
$\alpha_2$ -Macroglobulin	P01023	GCVLLSYLJETVTVSASLESVR VSJQTLSSLFFTVLQDVPVR <sup>c</sup> VSJQTLSSLFFTVLQDVPVRDLKPAIVK	55 1424 1424
AN1-type zinc finger and ubiquitin domain-containing protein 1	Q86XD8	MKNMJLSKK	235 <sup>b</sup>
Angiopoietin-related protein 6	Q8NI99	VLJASAEAGR	145
Apolipoprotein D	P05090	ADGTVNQIEGEATPVJLTEPAK <sup>c</sup> ADGTVNQIEGEATPVJLTEPAKLEVK	98 98
Attractin	O75882	GICJSSDVR EWLPLJR <sup>c</sup> JHSCSEGQISIFR	300 383 731
$\beta_2$ -Glycoprotein 1	P02749	PSAGJNSLYR <sup>c</sup> VYKPSAGJNSLYR <sup>c</sup>	162 162
Biotinidase	P43251	FJDTEVLQR <sup>c</sup>	130
C10orf111 protein	Q49AL1	AFCVPTAJVSVVGLNCHLEK	9 <sup>b</sup>
C1q and tumor necrosis factor-related protein 3	Q0VAN4	TGTVDJNTSTDLK <sup>c</sup>	70 <sup>b</sup>
Cadherin-5	P33151	EVYPWYJLTVEAK <sup>c</sup>	442
Calumenin	O43852	JATYGYVLDDPDDPDDGFNYK	131 <sup>a</sup>
Ceruloplasmin	P00450	EHEGAIYPDJTDFQR <sup>c</sup> AGLQAFFQVQECJK EJLTAPGSDSAVFFEQGTTR <sup>c</sup> ELHHLQEQJVSNAFLDK <sup>c</sup> ELHHLQEQJVSNAFLDKGEFYIGSK	138 358 397 762 762
Cholinesterase	P06276	EJETEIK <sup>c</sup>	284
Clusterin	P10909	EDALJETR <sup>c</sup> KKEDALJETR LAJLTQGEDQYYLR <sup>c</sup>	86 86 374
Coiled coil domain-containing protein 146	Q8IYE0	IKJATEKMMALVAELSMK	815 <sup>b</sup>
Complement C1r subcomponent-like protein	Q9NZP8	PVTPIAQJQTTLGSSR <sup>c</sup>	242
Complement C2	P06681	TMFPJLTDVR <sup>c</sup>	651
Complement C4-A	P0C0L4	GLJVTLSSTGR <sup>c</sup>	1328
Complement component C7	P10643	JYTLTGR <sup>c</sup>	754
Complement factor H	P08603	IPCSQPPQIEHGTIJSSR <sup>c</sup> ISEEJETTCYMGK <sup>c</sup> MDGASJVTCINSR	882 911 1029 <sup>a</sup>
Complement-activating component of Ra-reactive factor	P48740	FGYILHTDJR NJLTTYK <sup>c</sup>	178 <sup>a</sup> 385 <sup>a</sup>
Contactin-1	Q12860	AJSTGLVITDPTR	494
Dopamine $\beta$ -hydroxylase	P09172	SLEAIJGSLQMGQR <sup>c</sup>	184
E3 ubiquitin-protein ligase Mdm2	Q00987	LEJSTQAEEGFDVPDCKK	349 <sup>b</sup>
Exportin 5	Q5JTE9	TRSJYTKVSR	138 <sup>b</sup>
Extracellular matrix protein 1	Q16610	HIPGLIHJMTAR	444
Fibrinogen $\gamma$ chain	P02679	DLQSLEDILHQVEJK VDKDLQSLEDILHQVEJK	78 78
Fibronectin	P02751	DQCIVDDITYNVJDTFHK HEEGHMLJCTCFQGR LDAPTNLQFVJETDSTVLVR <sup>c</sup>	528 542 1007
Fibulin-1	P23142	CATPHGDJASLEATFVK <sup>c</sup>	98 <sup>a</sup>
Ficolin-3	O75636	VELEDFNGJR <sup>c</sup>	189
Galectin-3-binding protein	Q08380	ALGFEJATQALGR <sup>c</sup> DAGVVCTJETR GLJLTEDTYKPR <sup>c</sup> AAIPSALDTJSSK <sup>c</sup> TVIRPFYLTJSSGVD <sup>c</sup>	69 125 398 551 580

## Precise and Large Scale CF Glycoprotein Identification

TABLE I—continued

Protein name	UniProt	Core fucosylated glycopeptide	Site
Haptoglobin	P00738	MVSHHJLTTGATLINEQWLLTTAK <sup>c</sup>	184
		NLFLjHSEjATAK <sup>c,d</sup>	207, 211
		VVLHPJYSQVDIGLIK <sup>c</sup>	241
Hemopexin	P02790	SWPAVGJCSSALR <sup>c</sup>	187
		ALPQPQJVTSLLGCTH <sup>c</sup>	453
Hyaluronidase-4 Ig $\alpha$ -1 chain C region	Q2M3T9	LISDMGKJVSATDIEYLAK	177 <sup>b</sup>
	P01876	PALEDLLLGSEAJLTCTLTGLR <sup>c</sup>	144
		LSLHRPALEDLLLGSEAJLTCTLTGLR	144
		PTHVJVSVMMAEVDGTCY <sup>c</sup>	340
		LAGKPTHVJVSVMMAEVDGTCY <sup>c</sup>	340
Ig $\alpha$ -2 chain C region	P01877	TPLTAJITK <sup>c</sup>	205
Ig $\gamma$ -1 chain C region	P01857	EEQYJSTYR	180
Ig $\gamma$ -2 chain C region	P01859	EEQFJSTFR	176
Ig $\gamma$ -4 chain C region	P01861	EEQFJSTYR	177 <sup>b</sup>
Ig $\mu$ chain C region	P01871	YKJNSDISSTR <sup>c</sup>	46
		GLTFQQJASSMCPDQDQTAIR <sup>c</sup>	210
Immunoglobulin J chain	P01591	EJISDPTSPLR <sup>c</sup>	49
		IIVPLNNREJISDPTSPLR	49
Insulin-like growth factor-binding protein 3	P17936	GLCVJASAVSR <sup>c</sup>	116
		AYLLPAPPAPGJASESEEDR <sup>c</sup>	136
		VDYESQSTDTQJFSSESK	199
Inter- $\alpha$ -trypsin inhibitor heavy chain H1	P19827	ICDLLVANNHFAHFFAPQJLTNMNK	285
Interleukin-6 receptor subunit $\beta$	P40189	LTVJLTNDR	390 <sup>b</sup>
Kallistatin	P29622	DFYVDEJTTVR	238
Kininogen-1	P01042	YNSQJQSNNQFVLYR	48
		ITYSIVQTJCSK	205
		LNAENJATFYFK <sup>c</sup>	294
Ligand-dependent nuclear receptor corepressor-like protein	Q8N3X6	JGTVDTGSENTEDGLDRKDSK	493 <sup>b</sup>
Metalloproteinase inhibitor 1	P01033	FVGTPEVJQTTLYQR <sup>c</sup>	53
Multimerin-1	Q13201	FNPGAESVLSJSTLK <sup>c</sup>	136
Otoancorin	Q7RTW8	JLSAVFKDLYDK	211 <sup>a</sup>
Phospholipid transfer protein	P55058	EGHFYJISEVK <sup>c</sup>	64
		VSJVSCQASVSR <sup>c</sup>	143
		JWSLPNR <sup>c</sup>	245
Plasma protease C1 inhibitor	P05155	DTFVJASR <sup>c</sup>	238
		GVTSVSQIFHSPDLAIRDTFVJASR	238
		VLSJNSDANLEINTWVAK <sup>c</sup>	253
		VGQLQLSHJLSLVILVPQNLK	352
		VPGJVTAVLGETLK <sup>c</sup>	469
Potassium voltage-gated channel subfamily H member 6	Q9H252	YJGSDPASGPSVQDK	449 <sup>b</sup>
Pro-low density lipoprotein receptor-related protein 1	Q07954	IETILLJGTDR	729
Prostaglandin-H <sub>2</sub> D-isomerase	P41222	SVVAPATDGGLJLTSTFLR <sup>c</sup>	78
Protein phosphatase 1 regulatory subunit 1C	Q8WVI7	HLKGQJESAFPEEEGTNER	89 <sup>b</sup>
Putative uncharacterized protein DKFZp686O16217	Q6N041	HYTJSSQDVTVPCR <sup>c</sup>	250 <sup>b</sup>
		MAGKPTHIJVSVMMAEADGTCY	485 <sup>b</sup>
Putative uncharacterized protein DKFZp566E164	Q9NTU4	JISKTRGWHSPGR <sup>c</sup>	58 <sup>b</sup>
Selenoprotein P	P49908	EGYSJISYIVVNHQGISSR	83
Serotransferrin	P02787	QQQHFLFGSJVTDCSGNFCLFR <sup>c</sup>	630
Signal peptide peptidase-like 2A	Q8TCT8	DMNQTLGDJITVK	155 <sup>b</sup>
Sulfhydryl oxidase 1	O00391	JGSGAVFPVAGADVQTLR <sup>c</sup>	130
Tomoregulin-2	Q9UIK5	SYDJACQIKEASCQKQEK	204 <sup>b</sup>
Uncharacterized protein ENSP00000375008	A6NH92	JTSISTAYMELSSLR	92 <sup>b</sup>
Vitronectin	P04004	JISDGFDFGIPDNVDAALALPAHSYSGR	242
von Willebrand factor	P04275	IGEADFJR <sup>c</sup>	1515
Zinc- $\alpha$ <sub>2</sub> -glycoprotein	P25311	DIVEYYNDSJGSHVLQGR <sup>c</sup>	109
		FGCEIENJR	125
4F2 cell surface antigen heavy chain	P08195	SLVTQYLJATGNR	323
ADAMTS-like protein 2	Q86TH1	DRJVTGTPLTGDK <sup>e</sup>	428 <sup>a</sup>
Afamin	P43652	DIENFJSTQK <sup>e</sup>	33
ADAM DEC1	O15204	EHAVFTSNQEEQDPAJHTCGVK <sup>e</sup>	184 <sup>a</sup>

TABLE I—continued

Protein name	UniProt	Core fucosylated glycopeptide	Site
ADAMTS-13	Q76LX8	YGEEYGJLTR <sup>e</sup>	667
ADAMTS-like protein 4	Q6UY14	LVSGJLTDR <sup>e</sup>	490
$\alpha_1$ -Antitrypsin	P01009	ADTHDEILEGLNFJLTEIPEAQIHEGFQELLR <sup>e</sup>	107
Attractin	O75882	CIJQSICEK <sup>e</sup>	1073 <sup>a</sup>
Cadherin-5	P33151	JTSLPHHVVGK <sup>e</sup>	61 <sup>a</sup>
		LDREJISEYHLTAVIVDK <sup>e</sup>	112
Cholesteryl ester transfer protein	P11597	SIDVSIQJVSVWFK <sup>e</sup>	105 <sup>a</sup>
Complement factor I	P05156	FLNJJGTCTAEGK <sup>e</sup>	103
Complement component C9	P02748	AVJITSENLIDDVSLIR <sup>e</sup>	415
Complement factor H	P08603	SPDVIJGSPISQK <sup>e</sup>	217 <sup>b</sup>
Desmoglein-2	Q14126	YVQJGTYTVK <sup>e</sup>	462
Desmocollin-2	Q02487	AJYTIK <sup>e</sup>	392
Glutaminase kidney isoform	O94925	WJNTPMDEALHFGHHDVFK <sup>e</sup>	620 <sup>b</sup>
Golgi membrane protein GP73	Q8NBJ4	AVLVNJITGER <sup>e</sup>	109
Heparin cofactor 2	P05546	JLSMPLLPADFK <sup>e</sup>	49
Histidine-rich glycoprotein	P04196	HSHNjjSSDLHPHK <sup>e</sup>	344 or 345 <sup>a</sup>
IgGfC-binding protein	Q9Y6R7	VITVQVAJFTLR <sup>e</sup>	1317 <sup>b</sup>
Ig $\mu$ chain C region	P01871	JNSDISSTR <sup>c</sup>	46
Intercellular adhesion molecule 1	P05362	AJLTVLLR <sup>e</sup>	145
FER protein	Q6PEJ9	GSTVQMNYVSVJVKSWLLMIQQTQLSRIMK <sup>e</sup>	66 <sup>b</sup>
Leucine-rich $\alpha_2$ -glycoprotein	P02750	MFSQJDR <sup>e</sup>	325
		LPPGLLAJFTLLR <sup>e</sup>	186
Macrophage mannose receptor 2	Q9UBG0	VTPACJTSLPAQR <sup>e</sup>	69 <sup>a</sup>
Membrane copper-amine oxidase	Q16853	YLYLASJHSNK <sup>e</sup>	592
Neuronal cell adhesion molecule	Q92823	VNVVJSTLAEVHWDPVPLK <sup>e</sup>	858 <sup>a</sup>
Pigment epithelium-derived factor	P36955	VTQJLTLIEESLTSEFIHDIDR <sup>e</sup>	285
Properdin	P27918	JVTFWGR <sup>e</sup>	428
Prothrombin	P00734	JFTENDLLVR <sup>e</sup>	416
Platelet glycoprotein V	P40197	LLDLSGNJLTHLPK	181
		JLSSLESVQLDHNQLETLPGDVFGALPR <sup>e</sup>	243 <sup>a</sup>
Protein HEG homolog 1	Q9ULI3	LjjSTGLQSSSVSQTKE <sup>e</sup>	313 <sup>b</sup> or 314 <sup>a</sup>
Polycystin-2	Q13563	VRJGSCSIPQDLRDEIK <sup>e</sup>	328 <sup>a</sup>
Protein PARM-1	Q6UWI2	JISIESR <sup>e</sup>	80 <sup>a</sup>
Poliovirus receptor-related protein 1	Q15223	NPJGTVTVISR <sup>e</sup>	202
Pro-low density lipoprotein receptor-related protein 1	Q07954	WTGHJVTVQR <sup>e</sup>	1511 <sup>a</sup>
Proteasome subunit $\beta$ type-4	P28070	FRJISR <sup>e</sup>	83 <sup>b</sup>
Protein phosphatase 1H	Q9ULR3	DFJMTGWAYKTIEDEDLKFPLIYGEGK <sup>e</sup>	354 <sup>b</sup>
Roundabout homolog 4	Q8WZ75	IQLEJTVLLNPDAEGPKPR <sup>e</sup>	246
Scavenger receptor cysteine-rich type 1 protein M130	Q86VB7	APGWAJSSAGSGR <sup>e</sup>	105
Tetratricopeptide repeat protein 6	P02790	jGTGHGjSTHHGPEYMR <sup>d,e</sup>	240, 246
Type-1 angiotensin II receptor	P30556	MILJSTEDGIKRIQDDCPK <sup>e</sup>	4 <sup>a</sup>
Vitronectin	P04004	NJATVHEQVGGPSLTSDLQAQSK <sup>e</sup>	86
Vasorin	Q6EMK4	LHEITJETFR <sup>e</sup>	117

<sup>a</sup> Potential glycosite is in the database.

<sup>b</sup> The glycosite is not annotated in the database.

<sup>c</sup> The CF site was identified in both the healthy and HCC samples.

<sup>d</sup> Both sites were identified as glycosylation sites, and one must be core-fucosylated.

<sup>e</sup> The CF site was identified only in the HCC sample.

must be  $S_2$  ( $\pm 0.5 m/z$  errors) with the same charge state as the parent ion. Additional information of symbol ions is then used to further evaluate their confidence into five ranks (Fig. 4). The spectra in the top two ranks are retained, and their relevant  $MS^3$  spectra are regarded as candidates. This strict spectrum-filtering method greatly improved the credibility of identification. Furthermore by statistically analyzing candidate spectra, many important neutral loss signals, which result from GlcNAc-related fragmentation, were revealed. These fragmentation patterns are always accompanied by very

strong signals and had not been reported previously (Fig. 5). In addition, diagnostic ions of GlcNAc residues were observed in  $MS^3$  spectra (Fig. 3). Based upon these observations, these interfering peaks from GlcNAc fragmentation that are very intense and would decrease the accuracy of peptide scoring were localized and subtracted from the spectra. This novel optimization method can effectively increase the identification efficacy. Both the spectrum-filtering and the spectrum-optimizing processes have been performed automatically in pFind Studio. In addition, the unidentified candidates

can be analyzed *de novo*. This can supply novel information, which is not in the database (Fig. 3).

**Identified Results and Their Illumination for Further Clinical Research**—The efficacy of our strategy was first demonstrated by implementation on healthy human plasma (IgG-extracted); 115 different CF glycopeptides (105 CF sites) from 72 glycoproteins were identified. To further demonstrate its feasibility for clinical samples, we applied this strategy to plasma from HCC patients; 108 different CF glycopeptides (106 CF sites) from 79 glycoproteins were identified. Altogether 25 novel glycosylation sites and 19 annotated potential sites were identified from these two experiments (Table I). The scale of our results shows that these innovative methods provide a breakthrough in CF glycoproteomics research and may meet the needs of clinical medicine. Although the comparison between two types of samples was not a designated outcome of this study, it still gave us illuminations in several aspects. First, the CF sites of many glycoproteins whose CF levels have been reported as altered in patients with HCC were confirmed in our research, such as  $\alpha_1$ -antitrypsin (one site),  $\alpha_2$ -HS-glycoprotein (one site),  $\alpha_2$ -macroglobulin (two sites), apolipoprotein D (one site),  $\beta_2$ -glycoprotein 1 (one site), ceruloplasmin (four sites), fibrinogen  $\gamma$  chain (one site), haptoglobin (three sites), histidine-rich glycoprotein (one site), Ig  $\alpha$ -2 chain C region (one site), Ig  $\gamma$ -1 chain C region (one site), and serotransferrin (one site) (9, 15). Direct evidence of a CF site by MS would not only help to enhance the reliability of the CF modification as a biomarker but may also lead to further clinical research at a deeper modification site level instead of the protein level. As shown previously, the CF patterns of some glycoproteins may be used as biomarkers because they are more sensitive and specific than evaluation of the respective total protein levels (19). The question of whether the specific CF site would be the more effective “marker” is interesting. This question could not be answered previously because of the limitations of the traditional techniques, but it can be tackled by application of this strategy. Second, a specific marker, CF GP-73, was reported to be more sensitive and specific for HCC diagnosis than  $\alpha$ -fetoprotein (15). This marker was specifically identified in the HCC samples in our research, whereas hemopexin (two CF sites identified), IgM (two sites), and kininogen (three sites) were identified in both of our two experiments. These glycoproteins have not previously been reported in healthy plasma (9). These results remind us that although CF glycoproteomics research has significantly advanced during recent years and impressive results have been obtained in clinical research more extensive research is needed. This further research inevitably depends on the acquisition of massive qualitative and quantitative data on CF glycoproteins and CF sites. Recently fucosylated haptoglobin was reported as a novel marker for pancreatic cancer, and site-specific increases in fucosylation were observed (29). However, the specificity of this marker is still not ideal for diagnosis; evaluation of the CF levels of a combination of

glycoproteins would permit more reliable discrimination among different disease stages. In our research, all three tryptic CF glycopeptides of haptoglobin were identified. Moreover our strategy possesses the merit that stable isotope labeling techniques can be embedded for quantitative research. The relative abundance of CF glycoproteins in some diseases, such as pancreatic cancer, could be quantified with the strategy. It should be mentioned that because lectin enrichment strategy was used in the early step the quantitation information obtained would only represent the relative difference in CF glycoprotein abundance, whereas the ratios between glycans with and without core fucose could not be reached as reported in other researches (8, 9).

In conclusion, this study holds promise for the large scale identification of CF glycoproteins, which can serve as a tool for the discovery of novel biomarker panels from clinical samples, such as body fluids or tissue biopsies. In addition, it is our hope that both identified and unidentified candidate spectra (over 10,000) will be a useful resource for the improvement of database searching methods for glycopeptides. Spectra data sets of this sort are rare and should arouse the interest of scientists in both glycoproteomics and bioinformatics research fields.

**Acknowledgments**—We thank Ji-Yang Zhang, You Li, Chao Liu, Wen-Ping Wang, Li-yun Xiu, Xue-qun Zhang, and Lin-Juan Tian for contributions. We also thank the Digestive Department of the First Affiliated Hospital, College of Medicine, Zhejiang University for the offering of HCC plasma.

\* This study was supported by National Natural Science Foundation of China Grants 30621063 and 20735005; National Key Program for Basic Research Grants 2006CB910801, 2002CB713807, 2004CB518707, and 2007CB914104; Hi-Tech Research and Development Program of China Grants 2006AA02A308, 2007AA02Z315, and 2008AA02Z309; and Chinese Academy of Sciences Knowledge Innovation Program Grant KGGX1-YW-13.

☒ The on-line version of this article (available at <http://www.mcponline.org>) contains supplemental material.

¶¶ These authors contributed equally to this work.

‡‡ To whom correspondence may be addressed. E-mail: yingwtll@yahoo.com.cn.

§§ To whom correspondence may be addressed. E-mail: smhe@ict.ac.cn.

¶¶¶ To whom correspondence may be addressed. E-mail: qianxh1@yahoo.com.cn.

## REFERENCES

1. Parodi, A. J. (2000) Protein glycosylation and its role in protein folding. *Annu. Rev. Biochem.* **69**, 69–93
2. Walsh, G., and Jefferis, R. (2006) Post-translational modifications in the context of therapeutic proteins. *Nat. Biotechnol.* **24**, 1241–1252
3. Dwek, R. A., Butters, T. D., Platt, F. M., and Zitzmann, N. (2002) Targeting glycosylation as a therapeutic approach. *Nat. Rev. Drug Discov.* **1**, 65–75
4. Kondo, A., Li, W., Nakagawa, T., Nakano, M., Koyama, N., Wang, X., Gu, J., Miyoshi, E., and Taniguchi, N. (2006) From glycomics to functional glycomics of sugar chains: identification of target proteins with functional changes using gene targeting mice and knock down cells of FUT8 as examples. *Biochim. Biophys. Acta* **1764**, 1881–1889
5. Ma, B., Simala-Grant, J. L., and Taylor, D. E. (2006) Fucosylation in prokaryotes and eukaryotes. *Glycobiology* **16**, 158–184

6. Wang, X., Inoue, S., Gu, J., Miyoshi, E., Noda, K., Li, W., Mizuno-Horikawa, Y., Nakano, M., Asahi, M., Takahashi, M., Uozumi, N., Ihara, S., Lee, S. H., Ikeda, Y., Yamaguchi, Y., Aze, Y., Tomiyama, Y., Fujii, J., Suzuki, K., Kondo, A., Shapiro, S. D., Lopez-Otin, C., Kuwaki, T., Okabe, M., Honke, K., and Taniguchi, N. (2005) Dysregulation of TGF- $\beta$ 1 receptor activation leads to abnormal lung development and emphysema-like phenotype in core fucose deficient mice. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15791–15796
7. Wang, X., Gu, J., Ihara, H., Miyoshi, E., Honke, K., and Taniguchi, N. (2006) Core fucosylation regulates epidermal growth factor receptor-mediated intracellular signaling. *J. Biol. Chem.* **281**, 2572–2577
8. Block, T. M., Comunale, M. A., Lowman, M., Steel, L. F., Romano, P. R., Fimmel, C., Tennant, B. C., London, W. T., Evans, A. A., Blumberg, B. S., Dwek, R. A., Mattu, T. S., and Mehta, A. S. (2005) Use of targeted glycoproteomics to identify serum glycoproteins that correlate with liver cancer in woodchucks and humans. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 779–784
9. Comunale, M. A., Lowman, M., Long, R. E., Krakover, J., Philip, R., Seeholzer, S., Evans, A. A., Hann, H. W., Block, T. M., and Mehta, A. S. (2006) Proteomic analysis of serum associated fucosylated glycoproteins in the development of primary hepatocellular carcinoma. *J. Proteome Res.* **5**, 308–315
10. Okuyama, N., Ide, Y., Nakano, M., Nakagawa, T., Yamanaka, K., Moriwaki, K., Murata, K., Ohigashi, H., Yokoyama, S., Eguchi, H., Ishikawa, O., Ito, T., Kato, M., Kasahara, A., Kawano, S., Gu, J., Taniguchi, N., and Miyoshi, E. (2006) Fucosylated haptoglobin is a novel marker for pancreatic cancer: a detailed analysis of the oligosaccharide structure and a possible mechanism for fucosylation. *Int. J. Cancer* **118**, 2803–2808
11. Barrabés, S., Pagès-Pons, L., Radcliffe, C. M., Tabarés, G., Fort, E., Royle, L., Harvey, D. J., Moenner, M., Dwek, R. A., Rudd, P. M., De Llorens, R., and Peracaula, R. (2007) Glycosylation of serum ribonuclease 1 indicates a major endothelial origin and reveals an increase in core fucosylation in pancreatic cancer. *Glycobiology* **17**, 388–400
12. Geng, F., Shi, B. Z., Yuan, Y. F., and Wu, X. Z. (2004) The expression of core fucosylated E-cadherin in cancer cells and lung cancer patients: prognostic implications. *Cell Res.* **14**, 423–433
13. Saldova, R., Royle, L., Radcliffe, C. M., Abd Hamid, U. M., Evans, R., Arnold, J. N., Banks, R. E., Hutson, R., Harvey, D. J., Antrobus, R., Petrescu, S. M., Dwek, R. A., and Rudd, P. M. (2007) Ovarian cancer is associated with changes in glycosylation in both acute-phase proteins and IgG. *Glycobiology* **17**, 1344–1356
14. Tabarés, G., Radcliffe, C. M., Barrabés, S., Ramírez, M., Alexandre, R. N., Hoesel, W., Dwek, R. A., Rudd, P. M., Peracaula, R., and de Llorens, R. (2006) Different glycan structures in prostate-specific antigen from prostate cancer sera in relation to seminal plasma PSA. *Glycobiology* **16**, 132–145
15. Drake, R. R., Schwegler, E. E., Malik, G., Diaz, J., Block, T., Mehta, A., and Semmes, O. J. (2006) Lectin capture strategies combined with mass spectrometry for the discovery of serum glycoprotein biomarkers. *Mol. Cell. Proteomics* **5**, 1957–1967
16. Wright, L. M., Kreikemeier, J. T., and Fimmel, C. J. (2007) A concise review of serum markers for hepatocellular cancer. *Cancer Detect. Prev.* **31**, 35–44
17. Zhang, H., Li, X. J., Martin, D. B., and Aebersold, R. (2003) Identification and quantification of N-linked glycoproteins using hydrazide chemistry stable isotope labeling and mass spectrometry. *Nat. Biotechnol.* **21**, 660–666
18. Kaji, H., Saito, H., Yamauchi, Y., Shinkawa, T., Taoka, M., Hirabayashi, J., Kasai, K., Takahashi, N., and Isobe, T. (2003) Lectin affinity capture, isotope-coded tagging and mass spectrometry to identify N-linked glycoproteins. *Nat. Biotechnol.* **21**, 667–672
19. Zhao, J., Simeone, D. M., Heidt, D., Anderson, M. A., and Lubman, D. M. (2006) Comparative serum glycoproteomics using lectin selected sialic acid glycoproteins with mass spectrometric analysis: application to pancreatic cancer serum. *J. Proteome Res.* **5**, 1792–1802
20. Häggglund, P., Bunkenborg, J., Elortza, F., Jensen, O. N., and Roepstorff, P. (2004) A new strategy for identification of N-glycosylated proteins and unambiguous assignment of their glycosylation sites using HILIC enrichment and partial deglycosylation. *J. Proteome Res.* **3**, 556–566
21. Häggglund, P., Matthiesen, R., Elortza, F., Højrup, P., Roepstorff, P., Jensen, O. N., and Bunkenborg, J. (2007) An enzymatic deglycosylation scheme enabling identification of core fucosylated N-glycans and O-glycosylation site mapping of human plasma proteins. *J. Proteome Res.* **6**, 3021–3031
22. Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J., and Gygi, S. P. (2003) Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J. Proteome Res.* **2**, 43–50
23. Alvarez-Manilla, G., Atwood, J., III, Guo, Y., Warren, N. L., Orlando, R., and Pierce, M. (2006) Tools for glycoproteomic analysis: size exclusion chromatography facilitates identification of tryptic glycopeptides with N-linked glycosylation sites. *J. Proteome Res.* **5**, 701–708
24. Tarentino, A. L., Quinones, G., Changchien, L. M., and Plummer, T. H. (1993) Multiple endoglycosidase F activities expressed by *Flavobacterium meningosepticum* endoglycosidases F2 and F3. *J. Biol. Chem.* **268**, 9702–9708
25. Wuhrer, M., Catalina, M. I., Deelder, A. M., and Hokke, C. H. (2007) Glycoproteomics based on tandem mass spectrometry of glycopeptides. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.* **849**, 115–128
26. Fu, Y., Yang, Q., Sun, R., Li, D., Zeng, R., Ling, C. X., and Gao, W. (2004) Exploiting the kernel trick to correlate fragment ions for peptide identification via tandem mass spectrometry. *Bioinformatics* **20**, 1948–1954
27. Li, D., Fu, Y., Sun, R., Ling, C. X., Wei, Y., Zhou, H., Zeng, R., Yang, Q., He, S., and Gao, W. (2005) pFind: a novel database-searching software system for automated peptide and protein identification via tandem mass spectrometry. *Bioinformatics* **21**, 3049–3050
28. Wang, L. H., Li, D. Q., Fu, Y., Wang, H. P., Zhang, J. F., Yuan, Z. F., Sun, R. X., Zeng, R., He, S. M., and Gao, W. (2007) pFind 2.0: a software package for peptide and protein identification via tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **21**, 2985–2991
29. Miyoshi, E., and Nakano, M. (2008) Fucosylated haptoglobin is a novel marker for pancreatic cancer: detailed analyses of oligosaccharide structures. *Proteomics* **8**, 3257–3262