A Proteomic Analysis of Human Bile*

Troels Zakarias Kristiansen‡§, Jakob Bunkenborg‡§, Mads Gronborg‡§, Henrik Molina‡§, Paul J. Thuluvath¶, Pedram Argani‖, Michael G. Goggins‖, Anirban Maitra‖, and Akhilesh Pandey‡

We have carried out a comprehensive characterization of human bile to define the bile proteome. Our approach involved fractionation of bile by one-dimensional gel electrophoresis and lectin affinity chromatography followed by liquid chromatography tandem mass spectrometry. Overall, we identified 87 unique proteins, including several novel proteins as well as known proteins whose functions are unknown. A large majority of the identified proteins have not been previously described in bile. Using lectin affinity chromatography and enzymatically labeling of asparagine residues carrying glycan moieties by ^{18}O, we have identified a total of 33 glycosylation sites. The strategy described in this study should be generally applicable for a detailed proteomic analysis of most body fluids. In combination with "tagging" approaches for differential proteomics, our method could be used for identification of cancer biomarkers from any body fluid. ***Molecular & Cellular Proteomics 3:715–728, 2004.***

Approximately 7,500 new patients are diagnosed with biliary tract cancer each year in the United States, and nearly 4,500 patients (~65%) die from their disease (1). Once established, biliary tract cancers are notoriously challenging to diagnose and treat. At present, only those patients with a completely resectable cancer achieve a modest 5-year survival. Those with unresectable cancers have a poor prognosis. In general, the outcome for patients with biliary tract cancer at any site is disappointing, and neither radiation nor pre- or postoperative conventional chemotherapy significantly improve survival or quality of life. Therefore, identifying patients with early, potentially curable, biliary tract cancers offers the best chance for improving survival (2).

Currently, the sensitivity and specificity of laboratory tests for early detection of biliary tract cancers is less than optimal, and there is considerable difficulty in distinguishing malignant from benign causes of bile duct obstruction. For example, cytologic specimens from brush biopsies have a notorious propensity for yielding false-positives and false-negatives, with an unacceptable overall sensitivity in the range of only 33–60% (3–5). Cancer antigen (CA)[1] 19-9 is widely used for serologic detection of cholangiocarcinoma and has a sensitivity of only 50–60% and specificity of 80% (6, 7). Similarly, detection of *p53* and *K-ras* gene mutations in bile has a sensitivity of only 33% and a specificity of 87% (8). There is clearly a need to identify novel, highly sensitive, and specific biomarkers for fluid-based detection of biliary tract cancer. The development of a reliable, sensitive, and specific panel of fluid-based biomarkers will not only enable early diagnosis of cancer in at-risk individuals with a recognized risk factor for biliary tract cancer, but also provide a cost-effective alternative for noninvasive screening in populations where biliary tract cancer has a high incidence (*e.g.* American Indians and Hispanic communities).

Current proteomic technologies allow identification of disease-specific protein profiles. Changes that occur during the transformation of a healthy cell into a neoplastic cell can result in protein alterations including changes in abundance, protein modification, enzymatic activity, or subcellular localization. Identifying and understanding these changes is an underlying theme in cancer proteomics (9, 10). One would expect that biomarkers for biliary tract cancers should be more readily identifiable in the bile because of higher local concentrations of proteins derived from the biliary tract. However, no comprehensive study targeted toward defining the baseline proteome of human bile fluid has yet been performed. Here we provide the first comprehensive proteomic study of human bile fluid using a liquid chromatography and tandem mass spectrometric (LC-MS/MS) approach. We have elucidated the proteome of bile fluid using multiple fractionation techniques and affinity enrichment methods to identify several proteins that have not been previously described in bile. In addition, we provide definitive evidence for a large number of *N*-linked glycosylation sites on these proteins using lectin affinity chromatography followed by ^{18}O-labeling of the glycan attachment site by peptide *N*-glycosidase F (PNGaseF) treatment.

From the ‡McKusick-Nathans Institute of Genetic Medicine and Department of Biological Chemistry, Johns Hopkins University, Baltimore, MD 21205; §Department of Biochemistry and Molecular Biology, University of Southern Denmark, Campusvej 55, Odense M, Denmark; and Departments of ¶Gastroenterology and ‖Pathology, Johns Hopkins University, Baltimore, MD 21287

Received, January 30, 2004, and in revised form, March 30, 2004
Published, MCP Papers in Press, April 14, 2004, DOI 10.1074/mcp.M400015-MCP200

[1] The abbreviations used are: CA, cancer antigen; LC-MS/MS, liquid chromatography-tandem mass spectrometry; PNGaseF, peptide *N*-glycosidase F; ERCP, endoscopic retrograde cholangiopancreatography; Con A, concanavalin A; WGA, wheat germ agglutinin; Q-TOF, quadrupole time-of-flight; DMBT1, deleted in malignant brain tumors 1; NGAL, neutrophil gelatinase-associated lipocalin.

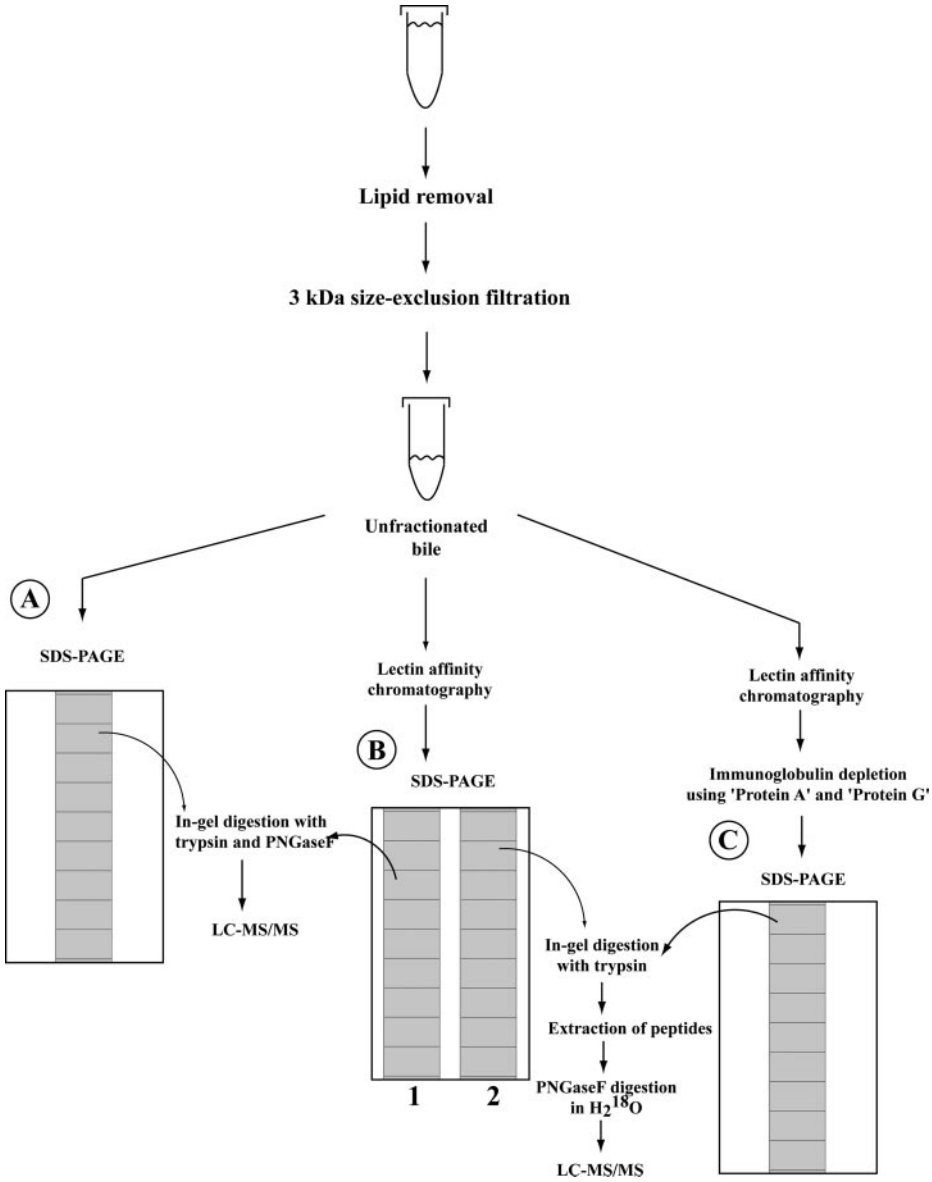

FIG. 1. **A Schematic of the purification procedure and mass spectrometric analysis of human bile.** Bile from a patient with cholangiocarcinoma was subjected to lipid removal followed by a 3-kDa size-exclusion filtration step as shown. The fraction obtained after this step was designated "unfractionated" bile and was the starting material for all subsequent experiments. The bottom part of the diagram shows an outline of the three types of experiments, indicated by *A*, *B*, and *C*, described in this study. *Scheme A* shows unfractionated bile separated by one-dimensional SDS-PAGE. The entire lane was excised into slices, followed by in-gel digestion with trypsin and PNGaseF, and finally identification of proteins by LC-MS/MS. *Scheme B* outlines two different types of experiments. Unfractionated bile was purified by lectin affinity chromatography, and the eluted proteins were separated by one-dimensional SDS-PAGE. Two identical affinity purifications are illustrated on the schematic of the gel. *Lane 1* underwent the same protocol as described for *Scheme A*, whereas *lane 2* was subjected to in-gel digestion with trypsin alone, extraction of the peptides, followed by labeling of *N*-linked glycosylation sites with PNGaseF in $H_2{}^{18}O$. *Scheme C* illustrates a two-step fractionation strategy. The bile was first subjected to lectin affinity chromatography followed by a second step, which consisted of immunoglobulin depletion, by a mixture of protein A and G. The proteins were then separated by one-dimensional SDS-PAGE and analyzed by LC-MS/MS.

MATERIALS AND METHODS

Sample Preparation—Bile fluid was obtained by endoscopic retrograde cholangiopancreatography (ERCP) from a patient with cholangiocarcinoma. One milliliter of the unfractionated bile fluid was centrifuged at 16,000 × *g* for 10 min at 4 °C. The partially cleared supernatant was then mixed with 250 μl of Cleanascite™HC (LigoChem, Inc., Fairfield, NJ) followed by rotation of the sample for 1 h at 4 °C. After incubation, the sample was centrifuged at 16,000 × *g* for 2 min to clear away the formed lipid-micelles, and the supernatant was transferred to a new tube. A pre-rinsed YM-3 centricon filter unit (molecular mass cut-off at 3 kDa) (Millipore, Bedford, MA) was loaded with the entire lipid-cleared sample and centrifuged at 6,500 × *g* until half of the sample volume had passed through the filter. MilliQ water was then added and the centrifugation repeated in order to reduce the salt concentration of the sample.

Lectin Affinity Chromatography—For the concanavalin A (Con A) affinity purification, 100 μl of concentrated bile fluid was mixed with 2× volume of Tris-buffered saline (TBS) buffer (0.05 M Tris-HCl, pH 7.1, 0.15 M NaCl) and 2× volume of a 50% slurry of Con A-agarose (Amersham Biosciences, Piscataway, NJ) followed by incubation at 4 °C overnight under rotation. After overnight incubation, the Con A beads were washed twice in TBS buffer (remove all buffer between washes in order to ensure minimal carry-over of albumin) and the bound protein eluted in 2 × 100 μl of 100 mM methyl α-D-mannopyranoside for 10 min at room temperature. The wheat germ agglutinin (WGA) (Amersham Biosciences) purification was done as described for the Con A affinity purification except that 100 mM *N*-acetyl-D-glucosamine was used for elution.

Immunoglobulin Depletion—For purification involving Con A followed by protein A and G, the initial procedure was identical to the Con A-only purification using four times the amount as starting material and all subsequent steps being scaled up four times as well (elution was done using 2 × 250 μl). The eluted proteins were then incubated with 30 μl of a 50% slurry of protein A and G (Sigma, St. Louis, MO) for 1 h at 4 °C under rotation.

Gel-electrophoresis and LC-MS/MS—All fractions were subjected to SDS-PAGE, and the gel was subsequently silver-stained as previously described (11). Gel lanes were excised for each of the samples

and divided into 28–32 sections depending on the complexity of the sample. All slices were in-gel digested with sequencing grade trypsin (Promega, Madison, WI) and PNGaseF (0.1 U/section) (Sigma) according to Küster *et al.* (12). Extracted peptides were dried down to ~10 μl and analyzed by LC-MS/MS on a Micromass Q-TOF API-US mass spectrometer (Manchester, United Kingdom).

^{18}O-labeling of N-linked Glycosylation Sites Using PNGaseF—For the $H_2{}^{18}O$-labeling of the samples by PNGaseF, the in-gel digestion was performed without PNGaseF. Extracted peptides were dried to completeness and rehydrated in 10 μl of $H_2{}^{18}O$ containing 0.1 U PNGaseF followed by overnight incubation at 37 °C. After incubation, the samples were analyzed by LC-MS/MS.

Liquid Chromatography and Mass Spectrometric Analysis—An Agilent 1100 series system (Agilent Technologies, Palo Alto, CA) was used for the chromatographic separation of the peptides. The peptides were loaded onto a pre-column packed with 10 μm C_{18} ODS-A (YMC, Ltd., Kyoto, Japan) and washed with 95% mobile phase A (100% H_2O with 0.4% acetic acid and 0.005% heptafluorobutyric acid v/v) and 5% mobile phase B (90% acetonitrile with 10% water, 0.4% acetic acid and 0.005% heptafluorobutyricacid). Subsequently, the peptides were eluted over 34 min with a flow of 300 nl/min using a linear gradient of 10–40% of mobile phase B onto an analytical column packed with 5 μm Vydac C_{18} material. The eluted peptides were analyzed by a Micromass Q-TOF API-US equipped with an ion source designed at Proxion Biosystems (Odense, Denmark). The automated data acquisition and generation of peak list files were done using MassLynx (version 4.0; Micromass). Data-dependent acquisition parameters for the scan cycle were set as follows. TOF survey scan: 0.9 s (from 350 to 1,500 *m/z*); MS/MS scans: 0.9 s (for up to three selected precursor ions) (from 50 to 2,000 *m/z*). Interscan time is 0.1 s for our instrument. For each survey scan, the three most intense ions in the spectrum were picked for MS/MS analysis, unless they appeared on the dynamic exclusion list (see below). MS/MS to MS switch criteria was set to intensity falling below 3 counts/s. Precursor ions selected for a given scan cycle were excluded for the next 180 s of the LC-MS/MS run. The collision energy was determined by charge state recognition for +2, +3, and +4 charged precursor ions.

Settings for Peak List File Generation and Database Searches—The peak list files were generated with MassLynx 4.0 using the following settings. Background subtraction: polynomial order: 0; below curve: 40%. Smoothing: smooth windows (channels): 4.00; number of smooths: 2; smooth mode: Savitzky Golay. Centroid: min. peak width at half height: 2; centroid mode: centroid top, 80%. Mascot version 1.9 installed on a Linux cluster was used to search the data using the MS/MS ion search mode for the generated peak list files. Mascot was used for database searching using the following parameters. Fixed modifications: carbamidomethyl modification of cysteines; variable modifications: oxidation of methionines, deamidation of asparagines (^{18}O modification in the relevant experiments), and pyroglutamic acid modification of amino-terminal glutamines; mass values: monoisotopic; peptide mass tolerance: ±0.4 Da; MS/MS mass tolerance: ±0.3 Da.

RESULTS AND DISCUSSION

Sample Preparation and Fractionation—Our initial strategy was to identify protein components in human bile by one-dimensional SDS-PAGE of crude bile followed by in-gel trypsin digestion and subsequent identification of the proteins by LC-MS/MS. However, the first attempts at separating crude bile on a one-dimensional gel revealed that the bile fluid, obtained by ERCP from a patient with cholangiocarcinoma, contained high amounts of lipids and bile salts, which interfered with our analysis. To circumvent this, we decided to perform a crude purification of the bile fluid for removal of these impurities (Fig. 1). The bile fluid was first subjected to a lipid removal step using Cleanascite™HC, a nonionic adsorbent used to precipitate lipids, fat droplets, cell debris, and mucinous impurities. Following lipid removal, we subjected the sample to a 3-kDa size-exclusion filtration step to remove salts and other small molecular mass components. As shown in Fig. 2, these two steps provided us with a satisfactory protein mixture that could be separated without smearing of gel bands as previously observed. The crude purified bile is referred to as "unfractionated bile" for the remainder of the article.

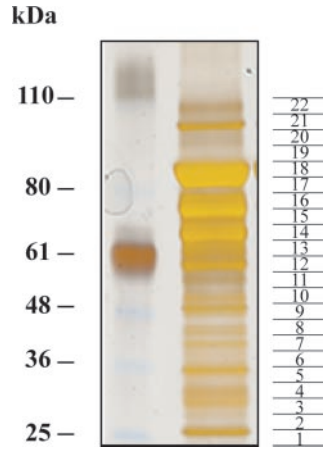

FIG. 2. **Silver-stained gel of unfractionated bile.** Twenty-five micrograms of unfractionated bile from the 3-kDa cut-off size-exclusion filtration was loaded on a gel for visualization. A larger one-dimensional gel loaded with 200 μg of material was used for the in-gel digestion in order to identify proteins. The positions of the individual gel slices are indicated to the *right* of the panel (as determined from the preparatory gel). These positions correspond to the numbers in the last column of Table I.

Identification of Protein Constituents in Bile by One-dimensional Gel Electrophoresis and LC-MS/MS—The unfractionated bile was first concentrated and purified further using a 3-kDa size-exclusion filtration step and subjected to SDS-PAGE followed by silver staining (Fig. 2). The entire lane was divided into 30 gel slices that were digested with trypsin and PNGaseF followed by LC-MS/MS. PNGaseF was added during the in-gel digestion procedure to remove *N*-linked glycans, as body fluids such as bile are expected to be rich in glycoproteins. PNGaseF is a glycopeptidase that cleaves *N*-linked high mannose, hybrid, and complex-type glycans at the linkage between the core structure and the anchoring asparagine, releasing the entire oligosaccharide and resulting in deamidation of the asparagine to an aspartic acid residue. The enzymatic cleavage of *N*-linked glycans serves two purposes. First, it results in a homogenous peptide population because *N*-linked glycans display a high degree of heterogeneity with regard to the occupancy of the site of attachment and of the sugar moieties found in individual glycan structures

TABLE I
List of proteins identified from unfractionated bile

	Accession no.	Name of protein	Protein class	Protein assigned from peptide in gel slice no.
1.	NP_005558	Mac-2-binding protein	Adhesion molecule	21
2.	NP_005134	Haptoglobin	Carrier/transport protein	11
3.	NP_000087	Ceruloplasmin	Carrier/transport protein	4
4.	NP_000509	β globin	Carrier/transport protein	4
5.	NP_000468	Albumin	Carrier/transport protein	18
6.	NP_002635	Polymeric immunoglobulin receptor	Carrier/transport protein	22
7.	NP_000604	Hemopexin	Carrier/transport protein	15
8.	NP_001054	Transferrin	Carrier/transport protein	22
9.	NP_000033	β-2-glycoprotein I	Carrier/transport protein	11
10.	NP_002334	Lactotransferrin	Carrier/transport protein	11
11.	NP_000574	Vitamin D-binding protein	Carrier/transport protein	16
12.	NP_000362	Transthyretin	Carrier/transport protein	7
13.	NP_003935	Selenium-binding protein 1	Carrier/transport protein	16
14.	NP_000500	Fibrinogen, γ chain	Clotting protein/factor	11
15.	NP_005132	Fibrinogen, β chain	Clotting protein/factor	9
16.	NP_004029	Amylase, salivary, α-1A	Glycosidase	17
17.	NP_000690	Amylase, pancreatic, α-2A	Glycosidase	17
18.	NP_653247	Immunoglobulin J	Immune system	2
19.	NP_000599	Orosomucoid 2	Immune system	13
20.	NP_000055	Complement component 3	Immune system	4
21.	NP_003881	IgG Fc-binding protein	Immune system	21
22.	NP_001726	Complement component 5	Immune system	21
23.	NP_000583	Complement component 4B	Immune system	1
24.	NP_001701	Complement factor B	Immune system	7
25.	NP_000598	Orosomucoid-1	Immune system	14
26.	NP_000217	Keratin 9	Keratin	10
27.	NP_000412	Keratin 10	Keratin	8
28.	NP_006112	Keratin 1	Keratin	5
29.	NP_000414	Keratin 2a	Keratin	16
30.	NP_000927	Pancreatic lipase	Lipase	15
31.	NP_004181	Lipase, gastric	Lipase	1
32.	NP_004657	Vanin 1	Enzyme	22
33.	NP_002037	Glyceraldehyde-3-phosphate dehydrogenase	Enzyme	9
34.	NP_000058	Carbonic anhydrase II	Enzyme	4
35.	NP_001743	Catalase	Enzyme	17
36.	NP_001729	Carbonic anhydrase I	Enzyme	4
37.	NP_031378	Elastase 3B	Protease	5
38.	NP_002761	Trypsinogen 2	Protease	4
39.	NP_001859	Pancreatic carboxypeptidase A1	Protease	6
40.	NP_002760	Trypsinogen 1	Protease	4
41.	NP_001897	Chymotrypsinogen B1	Protease	4
42.	NP_001862	Pancreatic carboxypeptidase B1	Protease	8
43.	NP_001860	Pancreatic carboxypeptidase A2	Protease	6
44.	NP_005738	Pancreatic elastase 3 (protease E)	Protease	6
45.	NP_009203	Chymotrypsin C	Protease	7
46.	NP_000005	α_2-macroglobulin	Protease inhibitor	9
47.	NP_001076	α-1-antichymotrypsin	Protease inhibitor	18
48.	NP_000345	Thyroxine-binding globulin	Protease inhibitor	5
49.	NP_000479	Antithrombin III	Protease inhibitor	16
50.	NP_000629	Vitronectin	Protease inhibitor	4
51.	NP_000072	Beige protein homolog	Unknown	4
52.	NP_006409	hGC-1 (human G-CSF-stimulated clone-1)	Unknown	21
53.	XP_065237	Similar to FKSG30	Unknown	2
54.	NP_005555	Lipocalin 2 (oncogene 24p3)	Unknown	1
55.	NP_001176	Zn-α-2-glycoprotein 1	Unknown	13
56.	NP_001630	Serum amyloid P component	Unknown	4
57.	XP_292555	Similar to testicular metalloprotease-like	Unknown	16
58.	NP_570602	α 1B-glycoprotein	Unknown	21
59.	NP_443204	Leucine-rich α-2-glycoprotein	Unknown	14

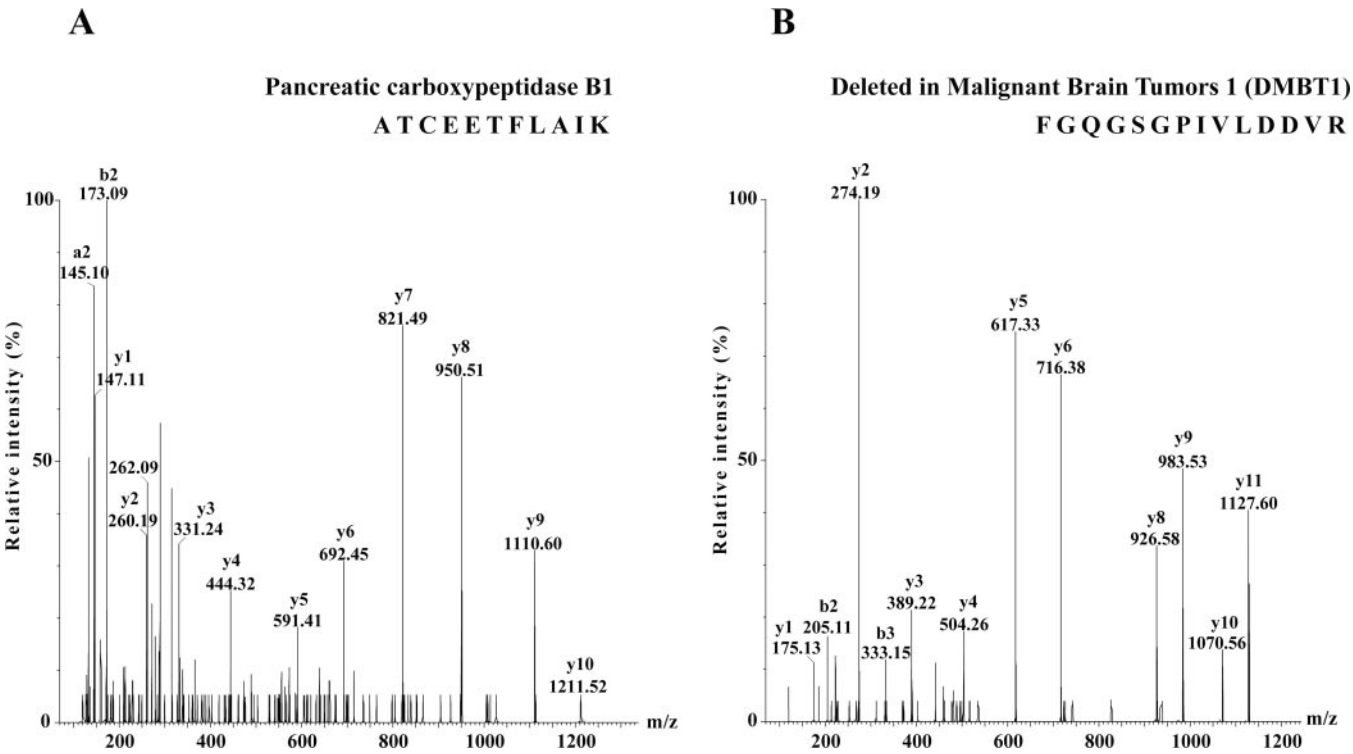

FIG. 3. **MS/MS spectra of two peptides from the unfractionated bile sample.** *A*, MS/MS spectrum of a peptide from the analysis of unfractionated bile that was derived from pancreatic carboxypeptidase B1. The spectrum shows the fragmentation pattern of a doubly charged precursor ion at *m/z* 641.82. Interpretation of the complete y-ion series provided the peptide sequence ATCEETFLAIK as shown. *B*, MS/MS analysis of a peptide found in gel band 30 from the analysis of WGA affinity-purified bile that was derived from DMBT1. The spectrum shows the fragmentation pattern of the doubly charged precursor ion at *m/z* 730.38. Interpretation of the y-ion series provided the peptide sequence FGQGSGPIVLDDVR as shown.

(13, 14). Second, electrospray experiments involving glycopeptides are difficult to perform because glycosylated peptides do not ionize as easily as their nonglycosylated counterparts. These properties of glycans and glycopeptides make analysis by electrospray mass spectrometry more tedious, unless the sugar chains are removed prior to analysis.

The result of the LC-MS/MS experiment is summarized in Table I. A total of 59 unique proteins was identified, many of which have not previously been reported in bile. Table I is divided into four columns. The first column gives the accession number in the reference sequence (RefSeq) database, the second column lists the protein name, the third column classifies the identified proteins according to their primary function, and the last column indicates the gel slice in which the peptide used to unambiguously assign the protein was found. Fig. 3*A* shows a typical MS/MS spectrum of a peptide from a protein found in Table I from unfractionated bile.

Data Processing Pipeline—The 30 peak list files generated from the LC-MS/MS runs corresponding to the 30 unfractionated bile samples were merged to a single peak list file to obtain better statistics for proteins identified in several gel bands (usually from neighboring gel bands) and to simplify the search procedure and data analysis. Database searching was done using the Mascot search engine (15). A combination of computer scoring and "human criteria" were employed in the screening of the data. First, the data was searched against the RefSeq database with tryptic constraints, and a base list of proteins was generated on which further analysis was performed. An initial list of proteins was generated by the following procedure:

1. Only proteins containing at least one unique peptide (we refer to peptide as being unique for specific protein if the sequence has not previously been used to assign to a different protein) with a peptide Mascot score greater than 20 were considered.
2. The highest scoring peptide for each of the protein entries generated in 1) was manually inspected and interpreted to confirm the identity of the peptide. If the spectrum could match a different sequence better than the assigned one, if it had poor ion statistics, or if no other good spectra pointed to the same protein, the hit was discarded. In addition, the inspected peptide match was required to have a length of at least 8 amino acids and to have a sequence tag of at least three amino acids, preferably a good y-ion series. Also, the y1 and a2-b2 pair, if present, had to be consistent with the identified peptide sequence. If the sequence tag

was not composed of y-ions and if no other peptides matched the same protein, the hit was discarded unless the spectrum was of good quality and most peaks could be explained.

3. If a protein has multiple isoforms or has multiple entries in the databases, we only specify the major form of the protein unless a specific peptide points to a region of the protein, which exists only in one of the isoforms.
4. If multiple peptides that matched the same protein are not from the same vicinity on the gel (more than two gel bands away), then extra care is taken to confirm those entries.

In order to identify proteins, which are not present in the RefSeq database, the peak list file was searched against the nonredundant (nr) database at NCBI and the results compared with the list generated by searching the RefSeq database. New entries retrieved from the nr database were tested against the same criteria as described above for identification purposes. If a spectrum that had been used to confirm an entry in the RefSeq derived dataset fitted an entry better in the search against the nr database, the original hit was removed (*e.g.* a different splice variant). Because many secreted proteins are heavily processed both within the cells (*e.g.* cleavage of signal peptide) and in the context of body fluids (*e.g.* protease processing), a high abundance of nontryptic peptide is likely to be present in the tryptic digest of our gel bands. As we are only searching a data set of tryptic peptides to minimize the amount of false-positive hits, we are potentially missing a large amount of peptides with good fragmentation spectra. The peak list was, therefore, subsequently searched against the RefSeq database with semitryptic constraints, which allows for the peptides in the database to contain only one nontryptic end. Analogous to the previous iteration, this list was compared against the combined list derived from the tryptic RefSeq and nr database searches.

Identification of Proteins in Bile After Enrichment by Lectin Affinity Chromatography—A common problem in proteomic approaches toward defining the constituents of human body fluids is the presence of high concentrations of albumin, which due to dynamic range issues prevents the identification of proteins present in low abundance (16). In keeping with this observation, we found albumin to be present in all but one fraction of unfractionated bile. To enable us to identify proteins otherwise undetectable due to the high abundance of albumin in bile, we chose lectin affinity chromatography, which allows one to remove albumin and to enrich for glycosylated proteins. We decided to use enrichment by two lectins as a complementary method for identifying proteins in bile.

Lectin affinity chromatography was performed using two types of lectins with different binding specificities, Con A and WGA. Con A binds to α-D-mannopyranosyl, α-D-glucopyranosyl, and sugars with similar steric properties, implying a preference for high mannose type of *N*-linked glycans (and to a lesser extent for hybrid type of sugars), with complex types of *N*-linked glycans binding weakly to this type of lectin (17). WGA binds to oligomers (and with lesser affinity to monomers) of β (1,4)-linked *N*-acetylglucosamine and to a lesser extent sialic acid residues (18). WGA, therefore, has preference for binding various hybrid and complex sugars, and to a lesser degree, if the glycan is extensively trimmed, it also binds to the core structure of *N*-linked glycans consisting of β (1,4)-linked *N*-acetylglucosamine bound to the sugar-carrying asparagine.

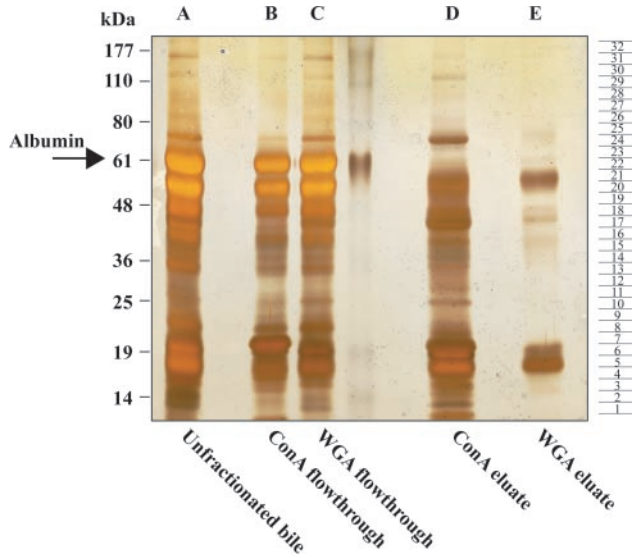

FIG. 4. **Silver-stained gel after purification by lectin affinity chromatography.** *Lane A* displays 100 μg of the unfractionated bile before affinity purification. *Lanes B* and *C* show the flowthrough from the Con A and WGA affinity purification, respectively. *Lanes D* and *E* show the eluate from the Con A and the WGA column, respectively. *Lanes D* and *E* were divided into 32 gel bands, excised, and analyzed by mass spectrometry. The positions of the individual gel slices are indicated to the *right* of the figure. These positions correspond to the numbers in the last column of Tables II and III.

The results of the affinity chromatography using these two lectins are shown in Fig. 4. The elution profile from the two lectins shows that less material was bound to WGA when compared with Con A. This is in agreement with the broader specificity and the higher affinity of Con A. Comparison with the lane containing unfractionated bile (see Fig. 4), the profiles of the two lectin-based affinity purifications are significantly different, emphasizing the ability of this method to enrich for a subset of proteins that would otherwise be difficult to identify in the presence of more-abundant proteins in bile. Notably, the lectin affinity purification practically eliminated the albumin band found in unfractionated bile, which is consistent with albumin not being *N*-glycosylated (19). In order to identify eluted proteins from the lectin-based affinity purification, both lanes were excised into smaller gel slices and subjected to in-gel digestion with PNGaseF and trypsin followed by LC-MS/MS as described above for the analysis of unfractionated bile. Also, the same data-processing pipeline used for the

TABLE II
List of proteins identified from the Con A affinity purification

	Accession no.	Name of protein	Protein class	Protein assigned from peptide in gel slice no.
1.	NP_005558	Mac-2-binding protein	Adhesion	24
2.	NP_001703	Carcinoembryonic antigen-related cell adhesion molecule 1	Adhesion	28
3.	NP_078966	CA125 ovarian cancer antigen	Adhesion	31
4.	NP_005134	Haptoglobin	Carrier/transport protein	16
5.	NP_000087	Ceruloplasmin	Carrier/transport protein	1
6.	NP_000509	β globin	Carrier/transport protein	1
7.	NP_000468	Albumin	Carrier/transport protein	24
8.	NP_002635	Polymeric immunoglobulin receptor	Carrier/transport protein	27
9.	NP_000604	Hemopexin	Carrier/transport protein	19
10.	NP_001054	Transferrin	Carrier/transport protein	27
11.	NP_000033	β-2-glycoprotein I	Carrier/transport protein	13
12.	NP_002334	Lactotransferrin	Carrier/transport protein	27
13.	NP_001124	α-albumin	Carrier/transport protein	3
14.	NP_000500	Fibrinogen, gamma chain	Clotting protein/factor	12
15.	NP_000497	Prothrombin	Clotting protein/factor	2
16.	NP_004029	Amylase, salivary, α-1A	Glycosidase	22
17.	NP_000690	Amylase, pancreatic, α-2A	Glycosidase	22
18.	NP_653247	Immunoglobulin J	Immune system	5
19.	NP_000053	Complement component 1 inhibitor	Immune system	27
20.	XP_292542	Immunoglobulin heavy chain	Immune system	22
21.	XP_036448	Immunoglobulin κ variable region	Immune system	6
22.	XP_301565	Immunoglobulin heavy chain	Immune system	22
23.	NP_000055	Complement component 3	Immune system	22
24.	NP_003881	IgG Fc-binding protein	Immune system	25
25.	NP_001726	Complement component 5	Immune system	25
26.	NP_000583	Complement component 4B	Immune system	22
27.	NP_001701	Complement factor B	Immune system	10
28.	NP_000598	Orosomucoid-1	Immune system	17
29.	NP_000217	Keratin 9	Keratin	20
30.	NP_000412	Keratin 10	Keratin	11
31.	NP_006112	Keratin 1	Keratin	2
32.	NP_000927	Pancreatic lipase	Lipase	20
33.	NP_005256	γ-glutamyltransferase 1	Enzyme	4
34.	NP_000013	Adenosine deaminase	Enzyme	16
35.	NP_001032	Sucrase-isomaltase	Enzyme	31
36.	NP_004657	Vanin 1	Enzyme	26
37.	NP_031378	Elastase 3B	Protease	8
38.	NP_002761	Trypsinogen 2	Protease	7
39.	NP_001859	Pancreatic carboxypeptidase A1	Protease	9
40.	NP_001141	Membrane alanine aminopeptidase	Protease	31
41.	NP_000893	Neprilysin	Protease	28
42.	NP_254275	Pancreatic elastase IIA	Protease	7
43.	NP_068576	Angiotensin I converting enzyme 2	Protease	28
44.	NP_000005	α_2-macroglobulin	Protease inhibitor	9
45.	NP_001076	α-1-antichymotrypsin	Protease inhibitor	21
46.	NP_000345	Thyroxine-binding globulin	Protease inhibitor	7
47.	NP_000479	Antithrombin III	Protease inhibitor	20
48.	NP_000629	Vitronectin	Protease inhibitor	5
49.	NP_000925	α-2-plasmin inhibitor	Protease inhibitor	15
50.	NP_002448	Mucin 2	Protease inhibitor	1
51.	NP_004397	DMBT1	Unknown	32
52.	NP_060145	PC-LKC gene product	Unknown	28
53.	NP_005555	Lipocalin 2 (oncogene 24p3)	Unknown	3
54.	NP_001176	α-2-glycoprotein 1, zinc	Unknown	16
55.	NP_001630	Serum amyloid P component	Unknown	6
56.	XP_292555	Similar to testicular metalloprotease-like	Unknown	22
57.	NP_570602	α 1B-glycoprotein	Unknown	26
58.	NP_443204	Leucine-rich α-2-glycoprotein	Unknown	19
59.	NP_848638	Alkaline sphingomyelinase	Enzyme	22

TABLE III
List of proteins identified from the WGA affinity purification

	Accession no.	Name of protein	Protein class	Protein assigned from peptide in gel slice no.
1.	NP_005558	Mac-2-binding protein	Adhesion	24
2.	NP_001703	Carcinoembryonic antigen-related cell adhesion molecule 1	Adhesion	25
3.	NP_005134	Haptoglobin	Carrier/transport protein	1
4.	NP_000087	Ceruloplasmin	Carrier/transport protein	18
5.	NP_000509	β globin	Carrier/transport protein	1
6.	NP_000468	Albumin	Carrier/transport protein	23
7.	NP_002635	Polymeric immunoglobulin receptor	Carrier/transport protein	26
8.	NP_000604	Hemopexin	Carrier/transport protein	18
9.	NP_000508	α 2 globin	Carrier/transport protein	1
10.	NP_000230	Lysozyme	Glycosidase	23
11.	NP_653247	Immunoglobulin J	Immune system	5
12.	NP_000053	Complement component 1 inhibitor	Immune system	26
13.	XP_292542	Immunoglobulin heavy chain	Immune system	6
14.	XP_036448	Immunoglobulin κ variable region	Immune system	7
15.	NP_000565	Decay accelerating factor for complement	Immune system	6
16.	NP_000598	Orosomucoid-1	Immune system	18
17.	NP_000217	Keratin 9	Keratin	6
18.	NP_000412	Keratin 10	Keratin	7
19.	NP_006112	Keratin 1	Keratin	7
20.	NP_000414	Keratin 2a	Keratin	7
21.	NP_002274	Keratin 5	Keratin	23
22.	NP_002268	Hard keratin 1	Keratin	23
23.	NP_002275	Keratin 6	Keratin	23
24.	NP_000927	Pancreatic lipase	Lipase	18
25.	NP_001141	Membrane alanine aminopeptidase	Protease	29
26.	NP_002248	Kallikrein 1	Protease	17
27.	NP_000005	α_2-macroglobulin	Protease inhibitor	1
28.	NP_001076	α-1-antichymotrypsin	Protease inhibitor	6
29.	NP_000072	Beige protein homolog	Unknown	7
30.	NP_004397	DMBT1	Unknown	30
31.	NP_060145	PC-LKC gene product	Unknown	30
32.	NP_001729	Carbonic anhydrase I	Enzyme	7

unfractionated bile LC-MS/MS runs was applied to the data from the lectin affinity purification.

As shown in Tables II and III, a total of 59 and 32 proteins were identified from the Con A and WGA affinity purification, respectively. Fig. 5*A* presents the overlap of proteins identified by the two different lectin-based affinity purification methods in the form of a Venn diagram, while Fig. 5*B* displays the overlap between the proteins identified from the unfractionated bile, the Con A purification, and WGA purification. In total, we identified the presence of 87 unique proteins in bile.

Identification of N-linked Glycosylation Sites by PNGaseF Treatment and ^{18}O-labeling—Glycosylation is the most common posttranslational modification found in mammalian cell systems and the site of attachment of the glycan, as well as the structure and composition of the carbohydrate moieties, has long been recognized as a biologically important characteristic. Two types of glycosylation events normally occur on proteins: *N*-linked glycosylation, where the carbohydrate chain is attached to asparagine residues, and *O*-linked glycosylation, where the carbohydrate chain is attached to serine or threonine residues. Proteins harboring *N*-linked glycosylation are commonly destined for secretion and many *N*-linked glycosylated proteins are thus often found in high abundance in extracellular environments.

As mentioned, PNGaseF treatment of the extracted peptides provides a characteristic tag on the peptide in the form of a deamidation event taking place on the glycan-linked asparagine during cleavage of the sugar by the enzyme, resulting in conversion of the asparagine residue to an aspartic acid residue with a concomitant increase of 1 Da in the mass of the amino acid residue. However, deamidation can occur spontaneously under certain conditions (20, 21), and because such an event cannot be distinguished from a deamidation event originating from the enzymatic cleavage by PNGaseF, a definitive conclusion about *N*-linked glycosylation cannot be drawn. To circumvent this problem, we decided to repeat the Con A affinity purification and to perform the PNGaseF cleavage step in the presence of $H_2{}^{18}O$ (12, 22, 23). The advantage of doing the enzymatic cleavage in $H_2{}^{18}O$ is illustrated in Fig. 6. During the deamidation process, the asparagine is converted into an aspartic acid with an ^{18}O stably incorporated, which gives rise to a mass increase of 3 Da instead of 1 Da.

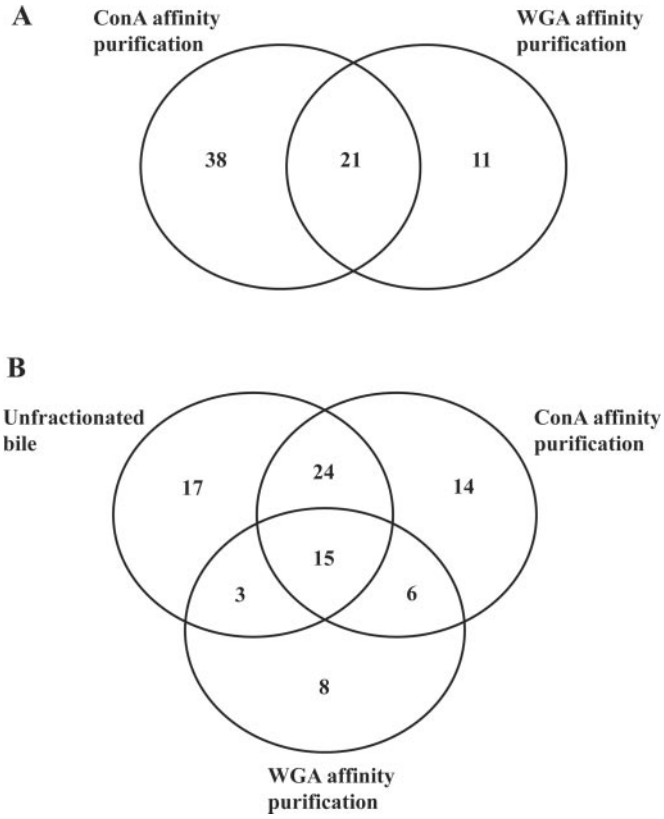

FIG. 5. **Overlap of proteins identified using different approaches.** *A*, distribution of the proteins identified using Con A and WGA affinity chromatography. A total of 68 unique proteins were identified from the two experiments. *B*, distribution of the proteins identified from unfractionated bile and from the Con A and WGA affinity chromatography. A total of 87 unique proteins were identified from the three experiments.

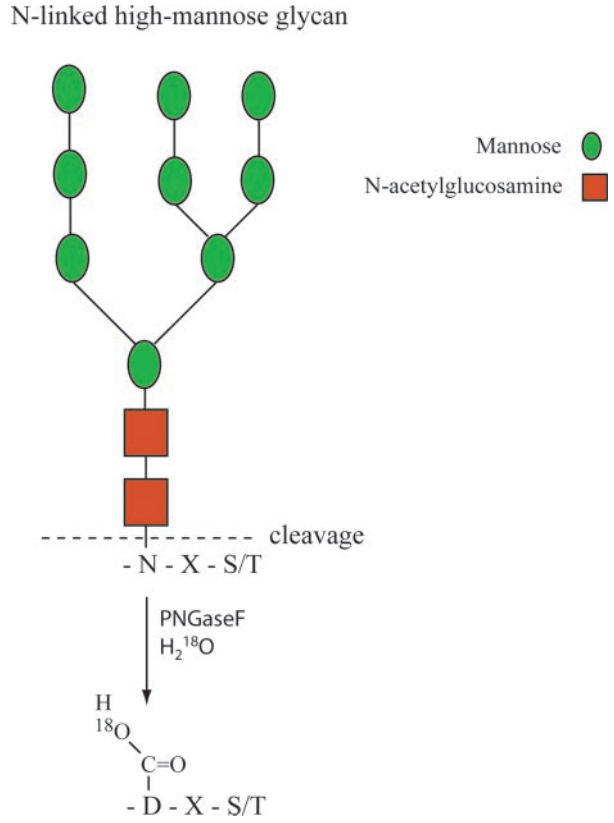

FIG. 6. **A schematic showing the strategy used for identification of glycosylation sites using PNGaseF.** Upon cleavage of the *N*-linked glycan by PNGaseF, the asparagine is converted into an aspartate along with incorporation of an ^{18}O atom, which gives rise to a 3-Da increase in mass of the residue as compared with an unmodified asparagine.

This eliminates the false-positives that arise from the spontaneously occurring deamidation events. Table IV shows a list of the identified glycosylation sites found in the Con A affinity-purified samples. Although a majority of the sites have previously been identified, some of the sites that we have identified have not been reported previously, emphasizing the continued need for detailed analysis of glycoproteins.

Immunoglobulin Depletion—During the course of analysis of our MS data, it became evident that a large proportion of the best quality MS/MS spectra originated from immunoglobulins. This was not unexpected, as most immunoglobulins found in body fluids are known to be glycosylated. Therefore, we decided to try to deplete our Con A affinity-purified proteins by using a combination of protein A and G, both of which bind heavy chains of various immunoglobulins and their subtypes. We anticipated that the automated data-dependent acquisition process, which governs the peak selection destined for sequencing by the mass spectrometer, would enable us to sequence low-abundance species that might be "squelched" in the nondepleted samples due to dynamic range issues. The immunoglobulin depletion experiment was carried out as for the "normal" lectin affinity purification, except for the one added step in which the samples were incubated with the mixture of protein A and G. Table V lists the glycosylation sites found in the depletion experiment. Again, we were able to identify a few novel glycosylation sites, although the large majority of the glycosylation sites have been previously reported.

Discussion of Proteins Identified in Human Bile Fluid—A large majority of proteins identified in our analysis include those that are synthesized by hepatocytes and thus would be somewhat expected in bile fluid. Such proteins include transport proteins (ceruloplasmin, transferrin, transthyretin (prealbumin), α_2-macroglobulin, and lactoferrin), enzymes (γ-glutamyltransferase and adenosine deaminase), proteins in the coagulation cascade (fibrinogen and antithrombin), and epithelial glycoproteins, such as the carcinoembryonic antigen-related cell adhesion molecule (CEACAM) 1. In fact, CEACAM1 is also known as biliary glycoprotein, as it was first isolated from bile fluid (24). Thus, this subset of proteins could be referred to as the "physiologic proteome" of bile fluid. As the bile sample we analyzed was obtained by ERCP, the presence of multiple pancreatic enzymes (*e.g.* pancreatic car-

TABLE IV
List of glycosylation sites identified by Con A enrichment[a]

	Peptide sequence	Whether site(s) are annotated by Swiss-Prot/Trembl	Protein name	Accession no. (Swiss-Prot/Trembl accession no.)
1.	QDQCIY**n**TTYLNVQR	Yes	α-1-acid glycoprotein 1	NP_000598 (P02763)
2.	YLG**n**ATAIFFLPDEGK	Yes	α 1-Antitrypsin	gi|231240 (P01009)
3.	VYKPSAG**n**NSLYR	Yes	β-2-glycoprotein I	NP_000033 (P02749)
4.	ILS**n**LSCNVK	No	G protein-coupled receptor 126	NP_065188 (Q8IXA4)
5.	SWPAVG**n**CSSALR	Yes	Hemopexin	NP_000604 (P02790)
6.	ALPQPQ**n**VTSLLGCTH	Yes		
7.	LSLHRPALEDLLLGSEA**n**LTCTLTGLR	Yes	Ig α-1 chain C region	gi|21619010 (P01876)
8.	TPLTA**n**ITK	Yes	Ig α-2 chain C region	gi|113585 (P01877)
9.	TKPREEQF**n**STFR* EEQF**n**STFR*	No	Ig γ-2 chain C region	gi|121043 (P01859)
10.	IIVPLNNRE**n**ISDPTSPLR	Yes	Immunoglobulin J chain	NP_653247 (P01591)
11.	GPSTPLPEDPNW**n**VTEFHTTPK (Figure 7)	No	Membrane alanine aminopeptidase	NP_001141 (P15144)
12.	A**n**LTNFPE**n**GTFVVNIAQLSQDDSGR* A**n**LTNFPE**n**GTFVVNIAQLSQDDSGRYK*	Yes Yes	Polymeric immunoglobulin receptor	NP_002635 (P01833)
13.	QQQHLFGS**n**VTDCSGNFCLFR	Yes	Transferrin	NP_001054 (P02787)
14.	LTGVAG**n**YTVCQK	No	Vanin 1	NP_004657 (O95497)

[a] * indicates peptides containing the same glycosylation site; **n** designates the glycosylated asparagine residue.

boxypepidase, pancreatic amylase, cationic trypsinogen, pancreatic lipase, and pancreatic elastase) was also not unexpected in the list of proteins identified.

In addition to hepatic and pancreatic proteins, we also identified several known "cancer-associated" proteins, perhaps reflecting the fact that the bile fluid was obtained from a patient harboring a cholangiocarcinoma. For example, we identified two epithelial apomucins, mucin 16 (also known as CA125 ovarian cancer antigen) and mucin 2 (MUC2) in the bile specimen. CA125 is a cell-surface glycoprotein that is widely used as a serum tumor marker for gastrointestinal and gynecological cancers, for diagnosis, as well as for monitoring recurrence (25, 26). CA125 levels are markedly elevated in both serum and bile in patients with cholangiocarcinomas (27–29). Along the same lines, the epithelial mucin MUC2 is normally expressed at minimal levels in the normal biliary epithelium, with MUC1 being the principal biliary mucin during development, switching to "adult-type" MUC3 expression after birth (30). However, expression of MUC2 is elevated in many pathologic conditions of the biliary tree, including chronic inflammatory states and in cancer (31, 32). Thus, our ability to detect two apomucins previously reported as differentially overexpressed in cholangiocarcinomas affirms the validity of our mass spectrometry-based approach.

We also identified additional cancer-associated proteins that have not been reported previously in the context of either cholangiocarcinomas, or even in bile fluid *per se*. These included three proteins: Mac-2-binding protein, lipocalin 2 (oncogene 24p3), and deleted in malignant brain tumors 1 (DMBT1). Mac-2-binding protein is a secreted glycoprotein that binds galectins, β_1 integrins, collagens, and fibronectin and has some relevance in cell-cell and cell-extracellular matrix adhesion (33, 34). Elevated serum levels of Mac-2-binding protein are often observed in patients with different types of solid tumors, including breast, ovarian, lung, and colorectal cancers, and are usually associated with a poor survival and metastatic spread in these malignancies (35–39). Low levels of Mac-2-binding protein are normally present in serum, semen, saliva, urine, tears, and in breast milk (33); this is the first report identifying this protein in bile fluid. Mac-2-binding protein was detectable in all three fractions (unfractionated bile, Con A, and WGA), raising the possibility that this protein may be a potential tumor marker for biliary cancer. Similarly, lipocalin 2, also known as neutrophil gelatinase-associated lipocalin (NGAL), is overexpressed in a variety of human cancers such as breast, colorectal, and pancreatic carcinomas (40–45); NGAL has recently been proposed as a tumor marker in urine for bladder cancer patients (46). Again, this is the first report of NGAL expression in bile fluid and implies that this protein could be a potential tumor marker for cholangiocarcinomas. Finally, DMBT1 is an opsonin receptor encoded by a gene located on chromosome 10q that is frequently deleted in gliomas and other malignant brain tumors (47); the DMBT1 protein is principally expressed in the lung, trachea, salivary gland, small intestine, and stomach (48). Curiously, while loss of DMBT1 protein expression has been reported in several tumor types (49, 50), a recent study suggests that this protein is overexpressed in pancreatic cancers (51). In fact, using a peptidomic approach to screening the conditioned media, the authors identified a 29-residue carboxyl-terminal fragment of DMBT1 that is secreted by pancreatic adenocarcinoma cell lines, but not by cell lines derived from normal pancreatic ductal epithelium (51). A number of keratins were detected in the different LC-MS/MS experiments. Given the nature of the sample used for the study and the way it was obtained, some of the observed keratins might

TABLE V
List of glycosylation sites identified by Con A enrichment followed by immunoglobulin depletion[a]

	Peptide sequence	Whether site(s) are annotated by Swiss-Prot/Trembl	Protein name	Accession no. (Swiss-Prot/Trembl accession no.)
1.	QIPLCANLVPVPIT**n**ATLDQITGK	Yes		
2.	QDQCIY**n**TTYLNVQR	Yes	α-1-acid glycoprotein 1	NP_000598 (P02763)
3.	QDQCIY**n**TTYLNVQRE**n**GTISR	Yes		
4.	QIPLCANLVPVPIT**n**ATLDR* PLCANLVPVPIT**n**ATLDR* PVPIT**n**ATLDR* PIT**n**ATLDR*	Yes	α-1-acid glycoprotein 2	NP_000599 (P19652)
5.	QLAHQS**n**STNIFFSPVSIATAF	Yes	α-1-antitrypsin	gi|177831 (P01009)
6.	YLG**n**ATAIFFLPDEGK	Yes		
7.	FGCEIEN**n**R	Yes	α-2-glycoprotein 1	NP_001176 (P25311)
8.	EQSTLAQMYPLQEIQ**n**LTVK	No	Angiotensin I converting enzyme 2	NP_068576 (Q9BYF1)
9.	SLTF**n**ETYQDISELVYGAK	Yes	Antithrombin III	NP_000479 (P01008)
10.	VYKPSAG**n**NSLYR* YKPSAG**n**NSLYR* KPSAG**n**NSLYR* PSAG**n**NSLYR*	Yes	β-2-glycoprotein I	NP_000033 (P02749)
11.	LG**n**WSAMPSCK	Yes		
12.	EHEGAIYPD**n**TTDFQR* PD**n**TTDFQR	Yes	Ceruloplasmin	NP_000087 (P00450)
13.	ALGFE**n**ATQALGR (Figure 7)	No	Mac-2-binding protein	NP_005558 (Q08380)
14.	P**n**VTTVER	No	γ-glutamyltransferase 1	NP_005256 (P19440)
15.	SWPAVG**n**CSSALR* PAVG**n**CSSALR*	Yes	Hemopexin	NP_000604 (P02790)
16.	ALPQPQ**n**VTSLLGCTH* PQPQ**n**VTSLLGCTH*	Yes		
17.	PALEDLLLGSEA**n**LTCTLTGLR* LSLHRPALEDLLLGSEA**n**LTCTLTGLR*	Yes	Ig α-1 chain C region	gi|21619010 (P01876)
18.	TPLTA**n**ITK* PLTA**n**ITK*	Yes	Ig α-2 chain C region	gi|9367869 (P01877)
19.	VITVQVA**n**FTLR	No	IgG Fc-binding protein	NP_003881 (Q9Y6R7)
20.	IIVPLNNRE**n**ISDPTSPLR	Yes	Immunoglobulin J chain	NP_653247 (P01591)
21.	SY**n**VTSVLFR	Yes	Lipocalin 2 (oncogene 24p3)	NP_005555 (P80188)
22.	A**n**LTNFPE**n**GTFVVNIAQLSQDDSGR	Yes		NP_002635 (P01833)
23.	LSLLEEPG**n**GTFTVILNQLTSR	Yes	Polymeric immunoglobulin receptor	
24.	VPG**n**VTAVLGETLK	Yes		
25.	KP**n**GSELMPK	No	Putative GluR6 kainate receptor	gi|15485588
26.	CGLVPVLAENY**n**K* PVLAENY**n**KSDNCEDTPEAGYFAVAVVKK* PVLAENY**n**KSDNCEDTPEAGYFAVAVVK* PVLAENY**n**K*	Yes	Transferrin	NP_001054 (P02787)
27.	QQQHLFGS**n**VTDCSGNFCLFR	Yes		
28.	QDTFIAAVYEHAAILP**n**ATLTPVSR	No	Vanin 1	NP_004657 (O95497)

[a] * indicates peptides containing the same glycosylation site; **n** designates the glycosylated asparagine residue.

be due to contaminants introduced during the sampling. However, keratin 1, 2a, 9, and 10 have all been described in the context of hepatobiliary cancers and thus likely constitute real bile components.

Notably, we have also identified a large number of proteins whose function is unknown including some proteins that were only predicted by gene prediction programs. Thus, mass spectrometry-derived data can be used for functional annotation of genomes as well as to verify the existence of predicted gene products.

Evaluation of ^{18}O-labeling in Determination of Glycosylation Site—Table IV shows a list of the identified glycosylation sites found in the Con A affinity-purified samples. The first column of the table contains the identified peptide sequence harboring the glycosylation site, the second column indicates whether the glycosylation site is annotated in Swiss-Prot (or Trembl), the third column lists the name of protein from which the peptide is derived, and the last column contains the RefSeq (or GenBank) accession number and the Swiss-Prot (or Trembl) accession number in parentheses.

Most proteins found contain one or more glycosylation sites, but in a few cases they did not contain any site. Serum albumin functions as a carrier in serum and is known to bind nonspecifically to many serum proteins; α- and β-globin are

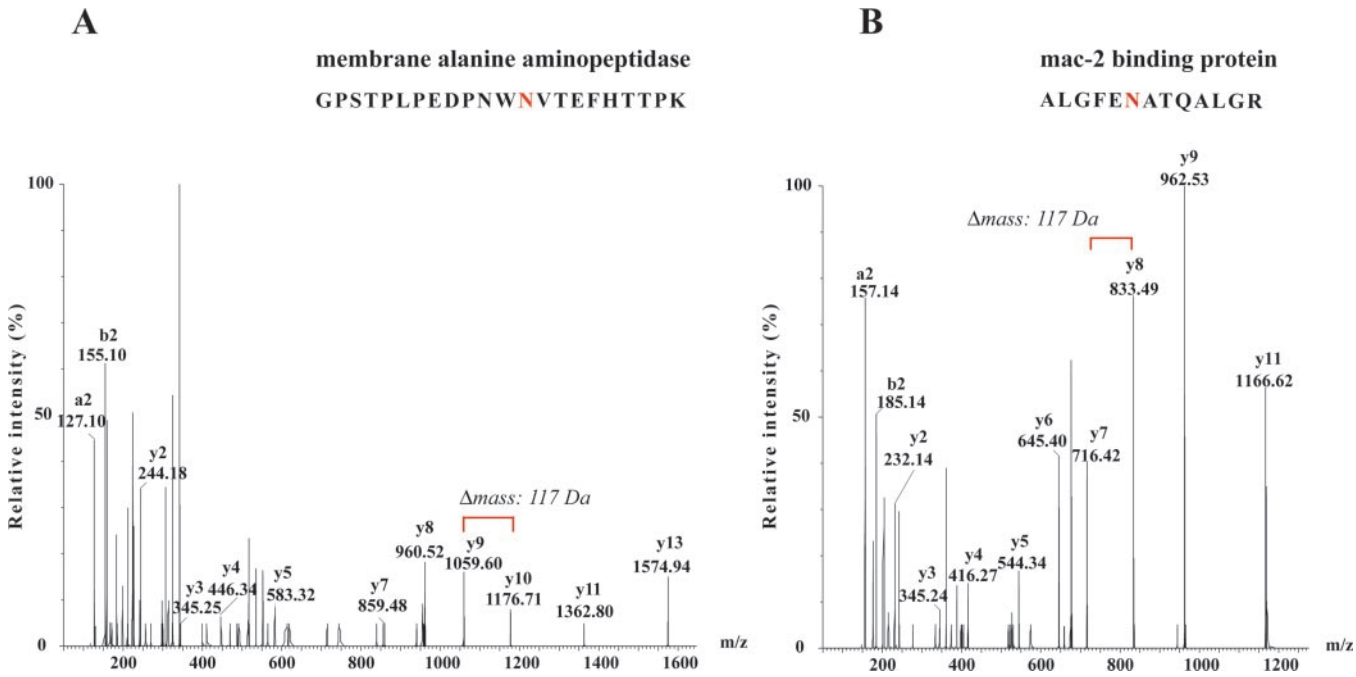

FIG. 7. **MS/MS spectra showing localization of two novel *N*-linked glycosylation sites.** *A* shows localization of a glycosylation site from the protein membrane alanine aminopeptidase that was identified from the Con A affinity purification followed by PNGaseF cleavage of the *N*-linked glycans in the presence of $H_2{}^{18}O$. The spectrum shows the fragmentation pattern of a triply charged precursor ion at *m/z* 823.10. The peptide sequence of the precursor ion is shown and the deamidated asparagine is indicated with *red coloring*. The mass difference between the y9 and y10 ion corresponds to 117 Da, indicating an aspartic acid with one ^{18}O atom incorporated (see schematic in Fig. 6). *B* shows the localization of a glycosylation site in the protein Mac-2-binding protein identified from the Con A purification followed by immunoglobulin depletion using protein A and G. The deamidated asparagine residue is indicated with *red coloring* as in *A*.

not glycosylated but are noncovalently but tightly bound to the plasma glycoprotein haptoglobin. Thus identifying a protein in the lectin affinity-purified gel does not necessarily mean that it is glycosylated and that is why our ^{18}O-labeling approach is necessary for definitive identification of a protein as a glycoprotein.

Two instances where we were able to localize the *N*-glycosylation site (membrane alanine aminopeptidase and Mac-2-binding protein) are shown in Fig. 7. Due to the incorporation of ^{18}O, the mass of deglycosylated asparagine residue is 117 Da, leading to an unambiguous assignment of the glycosylation site. While most of the proteins identified by lectin affinity chromatography were glycoproteins, only 15 glycosylation sites were identified by this method. This could indicate that the method is limited by the complexity of the sample and that it is necessary to further decrease the sample complexity prior to deglycosylation. Because immunoglobulins are glycosylated and present in bile in fairly high amounts, we tried depleting the immunoglobulins by protein A and G chromatography, and, using this strategy, we were able to identify a total of 28 glycosylation sites (Table V). Reassuringly, many of these were also found in the Con A-purified samples listed in Table IV. In all, we definitively identified 33 glycosylation sites. In some cases, several forms of the same glycopeptide was found with different N termini. This could result from in-source fragmentation during analysis or from proteolytic processing by aminopeptidases present in bile. Because proline residues are prone to fragmentation, it is possible that these shorter forms are the result of in-source fragmentation. Proline residues are also known to slow the trimming of the peptide end by exopeptidases and could also explain why several of the shorter versions of peptides begin with a proline residue.

CONCLUSIONS

A limited number of proteomic studies to analyze bile have been performed thus far. Using two-dimensional electrophoresis, He *et al.* studied the composition of vesicular and micellar proteins of human gall bladder (52). Upon comparison with reference two-dimensional electrophoresis maps of human plasma, red blood cells, and liver cells, the authors identified eight serum proteins in the bile samples. In a different study aimed at isolating and identifying hydrophobic polypeptides in human bile, Stark *et al.* (53) managed, through chloroform/methanol extraction, specialized reversed-phase chromatography and gel-filtration, and MALDI-TOF mass spectrometry, to identify a small subset of five proteins, of which three had not been described in bile previously. Using one-dimensional gel electrophoresis and LC-MS/MS, Jones *et al.* (54) analyzed bile in rats before and after treatment with 1,1-dichloroethylene or diclofenac. The rat bile samples obtained prior to exposure with 1,1-dichloroethylene or diclofenac allowed the authors to identify a total of 23 proteins that

included several immunoglobulins, as well as hemoglobin α-1 and β chains.

Whereas the above-mentioned studies have targeted specific fractions of bile or at comparing specific states, our article aims to produce a catalog of protein components that exist in bile. The 87 unique proteins we have identified is the largest catalog of human bile protein components to date. As mentioned earlier, there is a need for better biomarkers to diagnose biliary tract cancers, and we believe that having a reliable catalog of proteins present in this body fluid could ease the difficult task of identifying potential biomarker candidates. We have used multiple fractionation and purification methods to obtain our catalog, and the Venn diagrams presented in Fig. 5 clearly shows the need for combining multiple techniques. Failure to use one of the three methods would have resulted in missing 8–17 proteins corresponding to 9–20% of our catalog. We believe that the catalog of proteins published in this article is only a starting point. Given the complexity of human serum (55), we will hopefully be able to expand on defining the bile proteome further using additional fractionation techniques to move closer to identification of biomarkers for hepatobiliary cancers using differential proteomics.

* This work is supported by the family of Margaret Lee. A. P. is supported by National Institutes of Health Grant CA62924 and the Alexander and Margaret Stewart Trust and holds Sidney Kimmel Scholar and Beckman Young Investigator Awards. A. M. is supported by a grant from the Cancer Research Foundation of America and a Johns Hopkins Clinical Scientist Award. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

** To whom correspondence should be addressed: Department of Biological Chemistry, Johns Hopkins University, Baltimore, MD 21205. Tel.: 410-502-6662; Fax: 410-502-7544; E-mail: pandey@jhmi.edu.

REFERENCES

1. de Groen, P. C., Gores, G. J., LaRusso, N. F., Gunderson, L. L., and Nagorney, D. M. (1999) Biliary tract cancers. *N. Engl. J. Med.* **341,** 1368–1378
2. Gores, G. J. (2000) Early detection and treatment of cholangiocarcinoma. *Liver Transpl.* **6,** S30–S34
3. Mansfield, J. C., Griffin, S. M., Wadehra, V., and Matthewson, K. (1997) A prospective evaluation of cytology from biliary strictures. *Gut* **40,** 671–677
4. Ponsioen, C. Y., Vrouenraets, S. M., van Milligen de Wit, A. W., Sturm, P., Tascilar, M., Offerhaus, G. J., Prins, M., Huibregtse, K., and Tytgat, G. N. (1999) Value of brush cytology for dominant strictures in primary sclerosing cholangitis. *Endoscopy* **31,** 305–309
5. Sturm, P. D., Rauws, E. A., Hruban, R. H., Caspers, E., Ramsoekh, T. B., Huibregtse, K., Noorduyn, L. A., and Offerhaus, G. J. (1999) Clinical value of K-ras codon 12 analysis and endobiliary brush cytology for the diagnosis of malignant extrahepatic bile duct stenosis. *Clin. Cancer Res.* **5,** 629–635
6. Bjornsson, E., Kilander, A., and Olsson, R. (1999) CA 19–9 and CEA are unreliable markers for cholangiocarcinoma in patients with primary sclerosing cholangitis. *Liver* **19,** 501–508
7. Patel, A. H., Harnois, D. M., Klee, G. G., LaRusso, N. F., and Gores, G. J. (2000) The utility of CA 19-9 in the diagnoses of cholangiocarcinoma in patients without primary sclerosing cholangitis. *Am. J. Gastroenterol.* **95,** 204–207
8. Muller, P., Ostwald, C., Puschel, K., Brinkmann, B., Plath, F., Kroger, J., Barten, M., Nizze, H., Schareck, W. D., Hauenstein, K., Liebe, S., and Lohr, J. M. (2001) Low frequency of p53 and ras mutations in bile of patients with hepato-biliary disease: A prospective study in more than 100 patients. *Eur J. Clin. Invest.* **31,** 240–247
9. Srinivas, P. R., Srivastava, S., Hanash, S., and Wright, G. L., Jr. (2001) Proteomics in early detection of cancer. *Clin. Chem.* **47,** 1901–1911
10. Verma, M., Wright, G. L., Jr., Hanash, S. M., Gopal-Srivastava, R., and Srivastava, S. (2001) Proteomic approaches within the NCI early detection research network for the discovery and identification of cancer biomarkers. *Ann. N. Y. Acad. Sci.* **945,** 103–115
11. Shevchenko, A., Wilm, M., Vorm, O., and Mann, M. (1996) Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. *Anal. Chem.* **68,** 850–858
12. Kuster, B., and Mann, M. (1999) ^{18}O-labeling of *N*-glycosylation sites to improve the identification of gel-separated glycoproteins using peptide mass mapping and database searching. *Anal. Chem.* **71,** 1431–1440
13. Robertson, E. R., and Kennedy, J. F. (1996) Glycoproteins: A consideration of the potential problems and their solutions with respect to purification and characterisation. *Bioseparation* **6,** 1–15
14. Charlwood, J., Bryant, D., Skehel, J. M., and Camilleri, P. (2001) Analysis of *N*-linked oligosaccharides: Progress towards the characterisation of glycoprotein-linked carbohydrates. *Biomol. Eng.* **18,** 229–240
15. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20,** 3551–3567
16. Anderson, N. L., and Anderson, N. G. (2002) The human plasma proteome: history, character, and diagnostic prospects. *Mol. Cell. Proteomics* **1,** 845–867
17. Bhattacharyya, L., and Brewer, C. F. (1989) Interactions of concanavalin A with asparagine-linked glycopeptides. Structure/activity relationships of the binding and precipitation of oligomannose and bisected hybrid-type glycopeptides with concanavalin A. *Eur. J. Biochem.* **178,** 721–726
18. Kawashima, H., Sueyoshi, S., Li, H., Yamamoto, K., and Osawa, T. (1990) Carbohydrate binding specificities of several poly-*N*-acetyllactosamine-binding lectins. *Glycoconj. J.* **7,** 323–334
19. Durand, G., and Seta, N. (2000) Protein glycosylation and diseases: Blood and urinary oligosaccharides as markers for diagnosis and therapeutic monitoring. *Clin. Chem.* **46,** 795–805
20. Liu, D. T. (1992) Deamidation: A source of microheterogeneity in pharmaceutical proteins. *Trends Biotechnol.* **10,** 364–369
21. Yuksel, K. U., and Gracy, R. W. (1986) *In vitro* deamidation of human triosephosphate isomerase. *Arch. Biochem. Biophys* **248,** 452–459
22. Gonzalez, J., Takao, T., Hori, H., Besada, V., Rodriguez, R., Padron, G., and Shimonishi, Y. (1992) A method for determination of *N*-glycosylation sites in glycoproteins by collision-induced dissociation analysis in fast-atom-bombardment mass-spectrometry—Identification of the positions of carbohydrate-linked asparagine in recombinant α-amylase by treatment with peptide-*N*-glycosidase-F in ^{18}O-labeled water. *Anal. Biochem.* **205,** 151–158
23. 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
24. Hinoda, Y., Neumaier, M., Hefta, S. A., Drzeniek, Z., Wagener, C., Shively, L., Hefta, L. J., Shively, J. E., and Paxton, R. J. (1988) Molecular cloning of a cDNA coding biliary glycoprotein I: Primary structure of a glycoprotein immunologically crossreactive with carcinoembryonic antigen. *Proc. Natl. Acad. Sci. U. S. A.* **85,** 6959–6963
25. Modugno, F. (2003) Ovarian cancer and high-risk women—Implications for prevention, screening, and early detection. *Gynecol. Oncol.* **91,** 15–31
26. Haga, Y., Sakamoto, K., Egami, H., Yoshimura, R., Mori, K., and Akagi, M. (1986) Clinical significance of serum CA125 values in patients with cancers of the digestive system. *Am. J. Med. Sci.* **292,** 30–34
27. Chen, C. Y., Shiesh, S. C., Tsao, H. C., and Lin, X. Z. (2002) The assessment of biliary CA 125, CA 19–9 and CEA in diagnosing cholangiocarcinoma—The influence of sampling time and hepatolithiasis. *Hepatogastroenterology* **49,** 616–620
28. Ker, C. G., Chen, J. S., Lee, K. T., Sheen, P. C., and Wu, C. C. (1991) Assessment of serum and bile levels of CA19-9 and CA125 in cholangitis

and bile duct carcinoma. *J. Gastroenterol. Hepatol.* **6,** 505–508
29. Brockmann, J., Emparan, C., Hernandez, C. A., Sulkowski, U., Dietl, K. H., Menzel, J., Wolters, H., Glodny, B., and Senninger, N. (2000) Gallbladder bile tumor marker quantification for detection of pancreato-biliary malignancies. *Anticancer Res.* **20,** 4941–4947
30. Sasaki, M., Nakanuma, Y., Terada, T., and Kim, Y. S. (1995) Biliary epithelial expression of MUC1, MUC2, MUC3 and MUC5/6 apomucins during intrahepatic bile duct development and maturation. An immunohistochemical study. *Am. J. Pathol.* **147,** 574–579
31. Sasaki, M., Nakanuma, Y., and Kim, Y. S. (1996) Characterization of apomucin expression in intrahepatic cholangiocarcinomas and their precursor lesions: An immunohistochemical study. *Hepatology* **24,** 1074–1078
32. Higashi, M., Yonezawa, S., Ho, J. J., Tanaka, S., Irimura, T., Kim, Y. S., and Sato, E. (1999) Expression of MUC1 and MUC2 mucin antigens in intrahepatic bile duct tumors: Its relationship with a new morphological classification of cholangiocarcinoma. *Hepatology* **30,** 1347–1355
33. Koths, K., Taylor, E., Halenbeck, R., Casipit, C., and Wang, A. (1993) Cloning and characterization of a human Mac-2-binding protein, a new member of the superfamily defined by the macrophage scavenger receptor cysteine-rich domain. *J. Biol. Chem.* **268,** 14245–14249
34. Inohara, H., Akahani, S., Koths, K., and Raz, A. (1996) Interactions between galectin-3 and Mac-2-binding protein mediate cell-cell adhesion. *Cancer Res.* **56,** 4530–4534
35. Iacobelli, S., Arno, E., D'Orazio, A., and Coletti, G. (1986) Detection of antigens recognized by a novel monoclonal antibody in tissue and serum from patients with breast cancer. *Cancer Res.* **46,** 3005–3010
36. Marchetti, A., Tinari, N., Buttitta, F., Chella, A., Angeletti, C. A., Sacco, R., Mucilli, F., Ullrich, A., and Iacobelli, S. (2002) Expression of 90K (Mac-2 BP) correlates with distant metastasis and predicts survival in stage I non-small cell lung cancer patients. *Cancer Res.* **62,** 2535–2539
37. Fusco, O., Querzoli, P., Nenci, I., Natoli, C., Brakebush, C., Ullrich, A., and Iacobelli, S. (1998) 90K (MAC-2 BP) gene expression in breast cancer and evidence for the production of 90K by peripheral-blood mononuclear cells. *Int. J. Cancer* **79,** 23–26
38. Scambia, G., Panici, P. B., Baiocchi, G., Perrone, L., Iacobelli, S., and Mancuso, S. (1988) Measurement of a monoclonal-antibody-defined antigen (90K) in the sera of patients with ovarian cancer. *Anticancer Res.* **8,** 761–764
39. D'Ostilio, N., Natoli, C., Grassadonia, A., Rossi, N., Di Stefano, P., Amatetti, C., Tinari, N., and Iacobelli, S. (1996) Prognostic value of a novel interferon-inducible 90K tumor antigen. *Ann. N. Y. Acad. Sci.* **784,** 288–293
40. Bratt, T. (2000) Lipocalins and cancer. *Biochim. Biophys. Acta* **1482,** 318–326
41. Stoesz, S. P., and Gould, M. N. (1995) Overexpression of neu-related lipocalin (NRL) in neu-initiated but not ras or chemically initiated rat mammary carcinomas. *Oncogene* **11,** 2233–2241
42. Nielsen, B. S., Borregaard, N., Bundgaard, J. R., Timshel, S., Sehested, M., and Kjeldsen, L. (1996) Induction of NGAL synthesis in epithelial cells of human colorectal neoplasia and inflammatory bowel diseases. *Gut* **38,** 414–420
43. Stoesz, S. P., Friedl, A., Haag, J. D., Lindstrom, M. J., Clark, G. M., and Gould, M. N. (1998) Heterogeneous expression of the lipocalin NGAL in primary breast cancers. *Int. J. Cancer* **79,** 565–572
44. Terris, B., Blaveri, E., Crnogorac-Jurcevic, T., Jones, M., Missiaglia, E., Ruszniewski, P., Sauvanet, A., and Lemoine, N. R. (2002) Characterization of gene expression profiles in intraductal papillary-mucinous tumors of the pancreas. *Am. J. Pathol.* **160,** 1745–1754
45. Furutani, M., Arii, S., Mizumoto, M., Kato, M., and Imamura, M. (1998) Identification of a neutrophil gelatinase-associated lipocalin mRNA in human pancreatic cancers using a modified signal sequence trap method. *Cancer Lett.* **122,** 209–214
46. Monier, F., Surla, A., Guillot, M., and Morel, F. (2000) Gelatinase isoforms in urine from bladder cancer patients. *Clin. Chim. Acta* **299,** 11–23
47. Mollenhauer, J., Wiemann, S., Scheurlen, W., Korn, B., Hayashi, Y., Wilgenbus, K. K., von Deimling, A., and Poustka, A. (1997) DMBT1, a new member of the SRCR superfamily, on chromosome 10q25.3–26.1 is deleted in malignant brain tumours. *Nat. Genet.* **17,** 32–39
48. Holmskov, U., Mollenhauer, J., Madsen, J., Vitved, L., Gronlund, J., Tornoe, I., Kliem, A., Reid, K. B., Poustka, A., and Skjodt, K. (1999) Cloning of gp-340, a putative opsonin receptor for lung surfactant protein D. *Proc. Natl. Acad. Sci. U. S. A.* **96,** 10794–10799
49. Wu, W., Kemp, B. L., Proctor, M. L., Gazdar, A. F., Minna, J. D., Hong, W. K., and Mao, L. (1999) Expression of DMBT1, a candidate tumor suppressor gene, is frequently lost in lung cancer. *Cancer Res.* **59,** 1846–1851
50. Mori, M., Shiraishi, T., Tanaka, S., Yamagata, M., Mafune, K., Tanaka, Y., Ueo, H., Barnard, G. F., and Sugimachi, K. (1999) Lack of DMBT1 expression in oesophageal, gastric and colon cancers. *Br J Cancer* **79,** 211–213
51. Sasaki, K., Sato, K., Akiyama, Y., Yanagihara, K., Oka, M., and Yamaguchi, K. (2002) Peptidomics-based approach reveals the secretion of the 29-residue COOH-terminal fragment of the putative tumor suppressor protein DMBT1 from pancreatic adenocarcinoma cell lines. *Cancer Res.* **62,** 4894–4898
52. He, C., Fischer, S., Meyer, G., Muller, I., and Jungst, D. (1997) Two-dimensional electrophoretic analysis of vesicular and micellar proteins of gallbladder bile. *J. Chromatogr. A* **776,** 109–115
53. Stark, M., Jornvall, H., and Johansson, J. (1999) Isolation and characterization of hydrophobic polypeptides in human bile. *Eur. J. Biochem.* **266,** 209–214
54. Jones, J. A., Kaphalia, L., Treinen-Moslen, M., and Liebler, D. C. (2003) Proteomic characterization of metabolites, protein adducts, and biliary proteins in rats exposed to 1,1-dichloroethylene or diclofenac. *Chem. Res. Toxicol.* **16,** 1306–1317
55. Anderson, N. L., Polanski, M., Pieper, R., Gatlin, T., Tirumalai, R. S., Conrads, T. P., Veenstra, T. D., Adkins, J. N., Pounds, J. G., Fagan, R., and Lobley, A. (2004) The human plasma proteome: A non-redundant list developed by combination of four separate sources. *Mol. Cell. Proteomics* **3,** 311–326