# Mass Spectrometric Detection of Tissue Proteins in Plasma*[S]

Hui Zhang‡§, Alvin Y. Liu‡¶, Paul Loriaux‡, Bernd Wollscheid‖, Yong Zhou‡, Julian D. Watts‡, and Ruedi Aebersold‡‖

**It has long been thought that blood plasma could serve as a window into the state of one's organs in health and disease because tissue-derived proteins represent a significant fraction of the plasma proteome. Although substantial technical progress has been made toward the goal of comprehensively analyzing the blood plasma proteome, the basic assumption that proteins derived from a variety of tissues could indeed be detectable in plasma using current proteomics technologies has not been rigorously tested. Here we provide evidence that such tissue-derived proteins are both present and detectable in plasma via direct mass spectrometric analysis of captured glycopeptides and thus provide a conceptual basis for plasma protein biomarker discovery and analysis.** ***Molecular & Cellular Proteomics 6:64–71, 2007.***

A particular challenge in the diagnosis and treatment of human disease is the identification of molecular markers, or biomarkers, for the detection of disease at an early and treatable stage and for the molecular definition of disease progression to allow for implementation of more effective treatments (1). Although gene expression array studies on cells and tissues have shown that such markers, or marker panels, do exist and can be associated with pathological changes in the disease and its prognoses (2, 3), most tissues are not readily accessible for routine gene expression screening, and the affected tissue segments might be difficult to identify *a priori*.

In contrast, blood plasma[1] is readily accessible and thought to acquire proteins secreted, shed, or otherwise released from the tissues through which blood circulates. Thus, there has been considerable interest in blood plasma as a potentially rich source of biomarkers that, either individually or in combination, indicate the status of the different organs and tissues in our body and in technology development to identify and quantify them. To this end, recent substantial efforts in MS-based analytical methodologies have significantly increased our ability to probe the plasma proteome (4). However, the efforts to detect proteins from a variety of tissues in plasma are predicated on the assumption that such proteins are indeed both present and detectable in plasma using these methodologies, a critical assumption that has not been rigorously tested or verified.

A major hurdle for success of efforts to discover tissue- and/or cell-derived changes in the blood plasma protein profile has been the fact that blood plasma is extremely complex, consisting of minimally tens of thousands of different molecular species that span a concentration range of at least 10 orders of magnitude (5). Indeed the human plasma proteome is dominated by highly abundant proteins: 22 proteins constitute 99% of the total plasma protein mass (6), and one protein, albumin, accounts for over half the total protein. This constellation, together with the limited detection sensitivity of mass spectrometers, has made it difficult to analyze all but the most abundant protein species in plasma. Although many of these abundant plasma proteins are indicators of interesting biology and have been reported to fluctuate in response to certain types of diseases (7), they are unlikely to be useful as markers for specific disease states or as indicators of the state of specific organs. Instead many useful marker proteins are expected to be present in untreated plasma at concentrations below the limit of detection of the commonly applied MS-based proteomics methods. This expectation is based on the measured plasma concentration of the few existing known protein disease markers, the scarcity of specific biomarkers identified to date despite significant research efforts, and the large dilution effect when proteins from small lesions enter the blood stream.

In an attempt to bridge this gap and to reach into the concentration range of the desirable tissue-derived proteins in plasma, we have developed methods for high throughput and in depth analysis of cell surface proteins, secreted proteins, and plasma proteins. These approaches are based on the observation that most cell surface and secreted proteins are glycosylated and that disease-associated glycoproteins, either secreted by cells, shed from their surfaces, or otherwise released, are likely to enter into the blood stream and thus represent a rich source of potential disease markers. The validity of taking such an approach is supported by the observation that most known clinical protein markers for which

From the ‡Institute for Systems Biology, Seattle, Washington 98103, ¶Department of Urology, University of Washington, Seattle, Washington 98195, and ‖Institute of Molecular Systems Biology, Swiss Federal Institute of Technology (ETH) Zurich and Faculty of Sciences, University of Zurich, CH-8093 Zurich, Switzerland

Received, May 1, 2006, and in revised form, October 9, 2006

Published, MCP Papers in Press, October 9, 2006, DOI 10.1074/mcp.M600160-MCP200

[1] In this paper, the term plasma is used to indicate serum or plasma.

blood tests are commonly used are also known to be glycosylated (8). Furthermore the reduction in sample complexity achieved by focusing on the enriched subproteome of glycosylated plasma proteins and peptides translates into improved concentration limits of detection and thus sensitivity of detection, thus increasing the likelihood that the same polypeptide will be detectable in both tissue and plasma (9).

Here we report the identification of *N*-linked glycosylation sites (*N*-linked glycosites) and glycoproteins from cultured cells and solid tissues and the detection of many of these glycosites in plasma via glycopeptide capture and LC-MS/MS. These data confirm the assumption that numerous proteins from different tissues are indeed present and detectable in plasma. Furthermore they demonstrate the viability of a simple method for the relatively high throughput MS-based characterization of these proteins in blood plasma and thus provide a strong argument for continued and increased research efforts in this area.

## EXPERIMENTAL PROCEDURES

*Materials and Reagents*—For chromatography procedures, we used HPLC grade reagents purchased from Fisher Scientific. PNGase F[2] was purchased from New England Biolabs (Beverly, MA), and hydrazide resin was from Bio-Rad. All other chemicals used in this study were purchased from Sigma. The SK-BR-3, Ramos, and Jurkat cells were obtained from ATCC (American Type Culture Collection, Manassas, VA). Human tissue specimens were obtained from organs surgically removed because of cancer under a human subject approval for prostate and bladder cancer biomarker discovery project supported by the Early Detection Research Network from the National Cancer Institute.

*Purification and Fractionation of N-Linked Glycopeptides from Plasma*—The *N*-linked glycosites identified from plasma were generated from data from four separate resources of human serum or plasma. Two of the plasma samples were from a study performed as part of the Human Proteome Organisation (HUPO) plasma proteome project (4). One of these HUPO plasma samples was an equal mixture (v/v) of plasma from one male and one postmenopausal female Caucasian-American donors. These samples were collected with sodium citrate as anticoagulant (BD Diagnostics). The second HUPO plasma sample was from the UK National Institute of Biological Standards and Control provided as a lyophilized citrated plasma standard from a pool of 25 donors (4). The third sample source for this study was generated at the Institute for Systems Biology from a pool of serum samples collected from seven healthy male donors and three healthy female donors. Following approval by Human Subject Institutional Review Board of the Institute for Systems Biology, trained phlebotomists collected blood from each donor into evacuated blood collection tubes. Blood was allowed to clot for 1 h at room temperature. Sera were collected by centrifugation at 3,000 rpm. It should be noted that, using these collection procedures for plasma and serum samples, contamination from breakage of platelet or other blood cells cannot be totally ruled out. Formerly *N*-linked glycosylated peptides were isolated using an *N*-linked glycopeptide capture procedure as described previously (10–12). For these studies, 750 $\mu$l of serum or plasma was used for *N*-linked glycopeptide isolation. The fourth set of data used for this study was generated from a previously published study of *N*-linked plasma glycopeptides from the Biological Systems Analysis and Mass Spectrometry group at Pacific Northwest National Laboratory in Richland, WA (13).

*Purification and Fractionation of N-Linked Glycopeptides from Cells and Solid Tissues*—Proteins from SK-BR-3 breast cancer cells were extracted via homogenization and fractionation of cell lysates. At confluence, SK-BR-3 cells were rinsed five times with serum-free medium followed by incubation in serum-free McCoy's 5a medium for 24 h at 37 °C in a humidified incubator at 5% $CO_2$. Cells were homogenized in 0.32 M sucrose, 100 mM sodium phosphate, pH 7.5, and separated into three fractions by sequential centrifugations (1,000 × *g* pellet, 17,000 × *g* pellet, and 17,000 × *g* supernatant) (14). Protein extraction from solid tissues was performed using cell-free supernatant after an initial digestion of the tissues with collagenase. The tissues were sliced into pieces in serum-free cell culture medium, and collagenase was added at a final concentration of 1 mg/ml. Tissues were digested overnight at room temperature with stirring, and a cell-free supernatant was obtained by centrifugation (10, 15). One-milligram aliquots of protein extracted from cultured breast cells and solid tissue samples were used for glycopeptide capture (10).

Isolation of glycopeptides from the plasma membrane of lymphocytes was by a modification of the glycopeptide capture method (10) that allows for specific labeling/isolation of just plasma membrane glycoproteins.[3] In brief, this was accomplished by the use of a biotinylated hydrazide instead of a solid-phase hydrazide to label only the cell surface glycoproteins on live B and T lymphocytes in culture. After labeling, total membrane proteins were again isolated from the cells (14) and were then proteolyzed with trypsin. Capture of plasma membrane-derived biotinylated glycopeptides was achieved via streptavidin affinity isolation (16), and the *N*-linked glycopeptides were once again recovered following cleavage with PNGase F.

*Analysis of Peptides by Mass Spectrometry*—Off-line fractionation of peptides isolated from human plasma samples by strong cation-exchange chromatography prior to analysis of each fraction via LC-MS/MS was performed as described previously (14). Peptides from other sources were analyzed by on-line reverse phase LC-MS/MS without further sample fractionation.

Fractionated peptides from plasma samples were analyzed using both an LCQ and LTQ ion trap mass spectrometer (Thermo Finnigan, San Jose, CA) as well as with an ESI-Q-TOF mass spectrometer (Waters, Milford, MA) according to standard practices and manufacturers' instructions (9).

Peptides isolated from solid tissues and breast cancer cells were identified using an LCQ or LTQ ion trap mass spectrometer. The peptides were injected in three aliquots into a homemade peptide cartridge packed with Magic $C_{18}$ (Michrom Bioresources, Auburn, CA) using a FAMOS autosampler (Dionex, Sunnyvale, CA), and then passed through a 10-cm × 75-$\mu$m-inner diameter microcapillary HPLC column packed with Magic $C_{18}$ resin. A linear gradient of acetonitrile from 5 to 32% over 100 min at a flow rate of ~300 nl/min was applied. MS/MS spectra were acquired in a data-dependent mode. Peptides isolated from B and T lymphocyte plasma membranes were analyzed on an LCQ ion trap mass spectrometer as described previously (16).

Acquired MS/MS spectra were searched against the International Protein Index (IPI) human protein database (version 2.28, containing 40,110 entries) using SEQUEST software (17). The database search parameters were set to the following modifications: carboxymethylated cysteines, oxidized methionines, and a (PNGase F-catalyzed) conversion of Asn to Asp that occurs at the original site of carbohy-

[2] The abbreviations used are: PNGase F, peptide-*N*-glycosidase F; TM, transmembrane; CD, cluster designation; IHC, immunohistochemistry; HUPO, Human Proteome Organisation.

[3] B. Wollscheid, R. Aebersold, and J. D. Watts, manuscript in preparation.

drate attachment to the peptide/protein (*i.e.* the *N*-glycosite). No other constraints were included for database searches.

Database search results were then statistically analyzed using PeptideProphet, which effectively computes a probability for the likelihood of each identification being correct (on a scale of 0 to 1) in a data-dependent fashion (18). A PeptideProphet probability score of ≥0.9 was used as a filter to remove low probability peptide identifications. This filtering step represented an estimated peptide sequence assignment error rate of 2% or less for all datasets as calculated by PeptideProphet. Although the majority of *N*-linked glycosylation occurs at a consensus N*X*(S/T) sequon (where *X* is any amino acid except proline) (19), ~20% of identified peptides did not contain such a sequon. These peptide identifications likely resulted from false positive identifications from the database search, nonspecific isolation of *N*-linked glycosites, and isolation of atypical *N*-linked glycosites (*i.e.* not containing the N*X*(S/T) motif) of which we do not have sufficient understanding to predict. Thus, to reduce the false positive rate of the identified *N*-linked glycosites and to focus on those *N*-linked glycosites we could be most confident about, the peptide sequences were additionally filtered to remove non-motif-containing peptides. Finally peptide sequences were analyzed with respect to individual unique N*X*(S/T) sequons such that overlapping sequences containing the same N*X*(S/T) sequon (*i.e.* redundant *N*-linked glycopeptides for the same *N*-linked glycosite) were resolved in favor of those peptide sequences that contained the greater number of tryptic cleavage termini.

*Subcellular Localization of Identified Proteins*—To predict the likely subcellular localization of identified peptides/proteins, we utilized freely available prediction software for determination of (secretion) signal peptides and likely cell membrane-spanning sequences. Signal peptides were predicted using SignalP 2.0 (20) and transmembrane (TM) regions were predicted using TMHMM (version 2.0) (21) for protein topology and the number of TM helices. Information from both SignalP and TMHMM were combined to allow for sorting of the identified *N*-glycosylated proteins into the following categories: (i) *cell surface*, proteins that contained predicted non-cleavable signal peptides and no predicted TM segments; (ii) *secreted*, proteins that contained predicted cleavable signal peptides and no predicted TM segments; (iii) *transmembrane*, proteins that contained predicted TM segments and extracellular loops and intracellular loops; and (iv) *intracellular*, proteins that contained neither predicted signal peptides nor predicted TM segments.

## RESULTS AND DISCUSSION

The goal of this study was to test whether *bona fide* peptides derived from a variety of cell or tissue types were also detectable in blood plasma. Because cell surface and secreted proteins are both likely to be deposited into the blood and most of them are also glycosylated, we set out to target the glycoprotein subproteome that could be readily identified from both selected cultured cell lines and solid tumor samples. We then went on to determine whether a significant subset of these cell- and tissue-derived glycoproteins were indeed similarly detectable and thus present in blood plasma.

The general approach we used for these analyses is summarized in Fig. 1 and consists of four basic steps. 1) *Protein extraction*. Proteins were extracted from cells via homogenization and differential centrifugations (14). For protein extraction from solid tissues, tissues were digested with collagenase to obtain a cell-free supernatant (15). 2) *Glycopeptide capture*. Proteins from tissues/cells and plasma were processed by the recently described solid phase-based method for the isolation of *N*-linked glycopeptides (10). The end product for this procedure is the isolation of deglycosylated peptides that originally contain *N*-linked carbohydrates in the native protein (10). This also results in the conversion of the formerly glycosylated Asn to an Asp side chain. 3) *Peptide identification*. Isolated peptides were analyzed by automated LC-MS/MS. SEQUEST database search was performed for peptide sequence identification (17) followed by implementation of PeptideProphet (18) for statistical determination of the peptide identifications most likely to be correct. 4) *Peptide comparison*. Peptides identified from the different samples were compared against each other to determine the peptides in common between different cell and tissue types as well as with peptides identified from plasma to determine which cell/tissue-derived proteins/peptides were also detectable in plasma. Because our general isolation procedures specifically targeted *N*-linked glycosylation and because there is a known consensus sequence for this modification (N*X*(S/T) where X can be any amino acid except Pro), we limited our comparisons solely to the identified peptide sequences that contained at least one such *N*-linked glycosylation motif to simplify and to further reduce false positive rates.

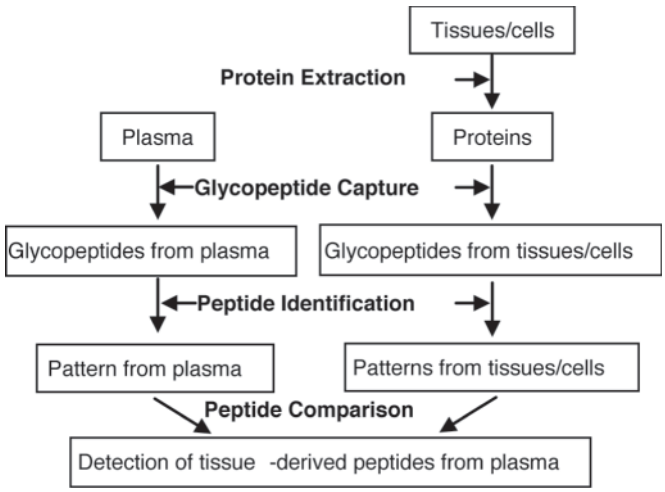



FIG. 1. **Schematic diagram of detection of *N*-linked glycopeptides from tissues/cells in plasma.** 1) *Protein extraction*. Proteins were extracted from cells using homogenization and differential centrifugation (14) or from solid tissues using collagenase digestion of tissues (15). 2) *Glycopeptide capture*. Proteins from tissues/cells and plasma were processed by recently described solid-phase extraction of glycopeptides (10). Peptides that contained *N*-linked carbohydrates in the native protein are isolated in their deglycosylated form (10). 3) *Peptide identification*. Isolated peptides were analyzed to generate identified peptide patterns from LC-MS/MS analysis and SEQUEST search (17). 4) *Peptide comparison*. Peptides obtained from different samples were compared, and peptides identified from both tissues/cells and plasma were determined.

We first set out to identify proteins from a cellular source that would have a high likelihood that polypeptides secreted, shed, or otherwise released would end up in plasma and would be as free from plasma contamination as possible. To

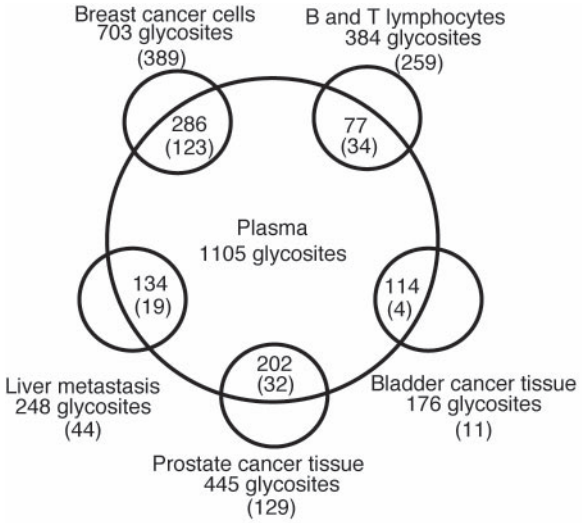



FIG. 2. **Comparison of *N*-linked glycosites identified from cell/tissue and plasma.** The total number of *N*-linked glycosites and tissue-specific *N*-linked glycosites are compared with the *N*-linked glycosites identified from plasma. Peptide identification was defined as scoring ≥0.9 with PeptideProphet (18). An identified *N*-linked glycosite was defined as cell/tissue-specific if it was only detected in one cell/tissue type in this study. The *numbers under* each cell/tissue type (703, 384, 176, 445, and 248) indicate the total number of *N*-linked glycosites identified from the specific cell/tissue type. The number of *N*-linked glycosites unique to each cell/tissue type (389, 259, 11, 129, and 44) is given in *parentheses*. The identifications that are common to a given cell/tissue and plasma are listed in *small circles* representing the cell/tissue (286, 77, 114, 202, and 134). The numbers of identified *N*-linked glycosites between each cell/tissue and plasma that were unique to each cell/tissue are indicated in *parentheses* (123, 34, 4, 32, and 19).

this end, we chose to characterize glycoproteins expressed on the surface of two human lymphocyte cell lines, one of B cell and one of T cell lineage (Ramos and Jurkat, respectively). Because lymphocytes naturally circulate in the blood, they come in contact with the blood plasma as much or more than any other cell type, thus maximizing the likelihood of their proteins being deposited into the plasma.

To achieve this aim, we isolated and identified *N*-linked glycopeptides from the plasma membranes of both Jurkat and Ramos cells for comparison with a previously compiled list of identified *N*-linked glycosites derived from plasma glycoproteins (11–13). In this way we were able to identify a total 384 *N*-linked glycosites from B and T cell surface glycoproteins with PeptideProphet score of ≥0.9. When compared with previously compiled data on 1,105 identified *N*-linked glycosites from plasma proteins (similarly scoring ≥0.9 with PeptideProphet), we found that 77 of the *N*-linked glycosites were in common with those already identified from plasma (Fig. 2 and Supplemental Table 1). This represented a significant portion (20%) of the total identifications from the B and T lymphocyte cell plasma membranes, thus confirming that lymphocyte-derived glycoproteins are both present and readily detectable in plasma when using this fairly simple glycoprotein/glycopeptide enrichment protocol upstream of identification by LC-MS/MS.

Because these identifications were achieved using cells grown in culture media supplemented with bovine serum, there was no potential for human blood contamination for these samples. However, some identifications could be attributed to bovine proteins should there be sufficient sequence homologies with human. To investigate this possibility, we submitted the sequences of the 77 *N*-linked glycosites representing this lymphocyte/plasma overlap to a search of the bovine protein database (bovine.nci.20051213). These results indicated that only 10 of the 77 *N*-linked glycosites were conserved between human and bovine. For these 10 *N*-linked glycosites, the source of origin could not be reliably assigned. However, for the remaining 67 *N*-linked glycosites that were not conserved, we could conclude that they could only have originated from the human cells under study, thus indicating that most or all of the plasma membrane glycoproteins identified from the human lymphocytes originated from the cells themselves rather than the culture medium. Thus, these data combined clearly indicated that glycoproteins expressed on the surface of lymphocytes were indeed detectable in the blood via solid phase-based isolation and LC-MS analysis of *N*-linked glycopeptides.

Because blood cells such as B and T lymphocytes and platelets naturally circulate in the blood, it was also possible that proteins could have been artificially introduced from such cells into the plasma during the blood/plasma collection rather than by natural release into the blood *in vivo*. Although this eventuality was difficult to experimentally exclude completely during the serum/plasma collection process, a clue as to whether this was generally a problem might be inferable from microarray data. To this end, we compared proteins identified in both prostate and plasma in this study with the transcriptional profiling data of these proteins in whole blood from available published microarray analyses (22, 23). We found transcription data for 162 of 202 *N*-linked glycosites that were identified in both prostate tissue and plasma (Fig. 2 and Supplemental Table 1) of which 78 were not detected in blood cells (an average difference value of 200 was used as threshold to make present/absent calls (23)). For 84 *N*-linked glycosites that were shown to be present in blood cells, genes for 20 *N*-linked glycosites were highly expressed in blood cells (expression in blood cells was 5-fold of the median value for 64 tissues or cells used). Therefore, the tissue origin of these 20 *N*-linked glycosites cannot be determined. On the other hand, a number of *N*-linked glycosites identified in both prostate tissue and plasma were preferentially expressed in prostate tissue but not in blood cells as shown by microarray analyses (Supplemental Table 1). These included CD26, lumican, MAC-2-binding protein, basement membrane-specific heparan sulfate proteoglycan core protein, and desmoglein (Supplemental Table 1). These observations suggest that the majority of proteins that were detected in both tissues and

plasma were likely deposited into the plasma from tissues *in vivo*.

We next sought to test whether the observation of such an overlap between *N*-linked glycosites identified from both lymphocytes and blood plasma could be extended to other cell types and tissues whose cells do not circulate in the blood stream. For this, we selected four different but representative cell/tissue types pertinent to cancer biomarker discovery to determine whether the *N*-linked glycosites identifiable from these sources are also present in the larger plasma dataset. Specifically we chose SK-BR-3 breast cancer cells, primary bladder and prostate cancer tissue, and a liver metastasis of prostate cancer.

*N*-Linked glycopeptides from the cultured SK-BR-3 breast cancer cells were isolated from a whole-cell lysate via conventional solid-phase glycoprotein/glycopeptide enrichment method. Similarly hydrazide-based isolation of *N*-linked glycopeptides from tissues was carried out with cell-free supernatants of collagenase-digested prostate, bladder, and liver metastasis tissue specimens (Fig. 1) (10, 15). The identification of isolated *N*-linked glycopeptides was via LC-MS/MS, and the results were similarly compared with the plasma dataset (24). When combined with the lymphocyte data, these data showed that of the total 1,257 *N*-linked glycosites identified in the two cell and three tissue types, 832 of these were identified in only one of the sample types (Supplemental Table 1). Fig. 2 summarizes the total number of *N*-linked glycosites identified in each cell/tissue type and the number of these that were unique to each specific cell or tissue type as well as the subsets of these that additionally overlapped with the plasma-derived *N*-linked glycosite dataset.

Similar to the comparison between lymphocytes and plasma, all four of these additional datasets showed a significant overlap with the plasma dataset. As can be seen from Fig. 2, some of the *N*-linked glycosites identified in both a particular cell/tissue and plasma were unique to that cell/tissue type. For example, of the 286 *N*-linked glycosites in common between plasma and breast cancer cells, 123 were not identified in any of the other cell/tissue samples evaluated. These results again support the contention that glycoproteins originating from cells or tissues are detectable in plasma using the relatively simple methodological approach of LC-MS analysis of enriched *N*-linked glycoproteins. Furthermore they indicate that glycoproteins from all or most cell and tissue types are likely to be found in the blood and be present at detectable levels for such an analytic approach.

In the above studies, proteins were identified by LC-MS/MS. In this method, not all proteins from cells, tissues, or plasma are identified due to the random sampling of peptide precursor ions during the analytical process. Therefore, we focused this study on the proteins commonly detected in both cell/tissue and plasma and put less value on the proteins only detected in specific tissues (tissue specificity). In addition, tumor cells and tissues were used to isolate the cell/tissue

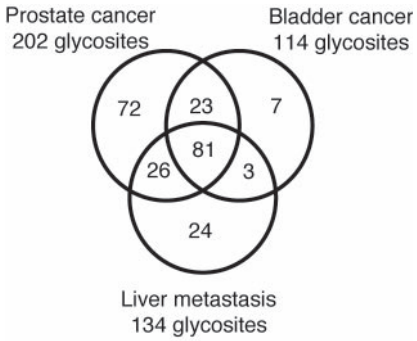
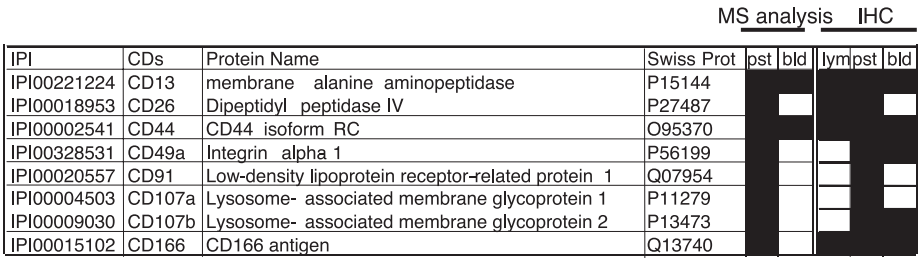



FIG. 3. **Tissue-derived *N*-linked glycosite identifications are also common to multiple tissue types.** Shown in this overlap are only the *N*-linked glycosites identified in prostate, bladder, or liver metastasis of prostate cancer that were also identified in plasma (from Fig. 2).

*N*-linked glycopeptides, whereas the dataset for plasma proteins was derived from samples obtained from non-cancer patient donors. Therefore, without quantitative comparison of protein concentration in normal and cancer plasma, we cannot confirm that the *N*-linked glycosites identified in common between tissues/cells and plasma shown here are associated with cancer. Conversely *N*-linked glycosites identified from cancer cells/tissues but not detected in the current plasma dataset could be potential cancer biomarkers for detection in plasma of cancer patients. For example, two prostate cancer tissue proteins, prostatic acid phosphatase and prostate-specific antigen, were not found in the plasma dataset. The levels of these proteins have been shown to be elevated in the plasma of prostate cancer patients and are unlikely to be detected in plasma of normal donors (8).

Unlike cultured cells, tissues are vascularized. One would thus expect that some contamination of the tissue glycoproteins by common circulating blood glycoproteins would inevitably occur. To investigate this possibility, we next examined the cell/tissue-derived data to see whether the overlap of *N*-linked glycosites detected in both plasma and the respective tissue sources could be explained by simple contamination from blood proteins. If this were the case, then we would expect that such contaminating plasma-derived glycoproteins would be a general effect and thus be detected in multiple tissues.

When we made this comparison, we found that a significant number of identified *N*-linked glycosites were indeed common to multiple tissues (Fig. 3 and Supplemental Table 1). For example, we identified 202 unique *N*-linked glycosites in both prostate tissue and plasma. By referencing available database annotations for these proteins, we were able to determine that 94 of these *N*-linked glycosites likely originated from proteins made by prostatic cells with another 96 likely originating from blood. The remaining 12 *N*-linked glycosites were annotated as hypothetical proteins whose origin could not be determined. Furthermore when we compared the *N*-linked glycosites identified from both prostate cancer tissue and plasma



| | | | | MS analysis | | IHC | | |
|---|---|---|---|---|---|---|---|---|
| IPI | CDs | Protein Name | Swiss Prot | pst | bld | lym | pst | bld |
| IPI00221224 | CD13 | membrane alanine aminopeptidase | P15144 | ■ | ■ | ■ | ■ | ■ |
| IPI00018953 | CD26 | Dipeptidyl peptidase IV | P27487 | ■ | □ | ■ | ■ | □ |
| IPI00002541 | CD44 | CD44 isoform RC | O95370 | ■ | ■ | ■ | ■ | ■ |
| IPI00328531 | CD49a | Integrin alpha 1 | P56199 | ■ | □ | □ | ■ | ■ |
| IPI00020557 | CD91 | Low-density lipoprotein receptor-related protein 1 | Q07954 | ■ | □ | □ | ■ | □ |
| IPI00004503 | CD107a | Lysosome- associated membrane glycoprotein 1 | P11279 | ■ | □ | □ | ■ | □ |
| IPI00009030 | CD107b | Lysosome- associated membrane glycoprotein 2 | P13473 | ■ | □ | □ | ■ | ■ |
| IPI00015102 | CD166 | CD166 antigen | Q13740 | ■ | □ | ■ | ■ | ■ |

FIG. 4. **Immunological confirmation of selected CD antigens identified from solid tissues.** CD antigens that were identified by MS in both plasma and tissue (prostate (*pst*) or bladder (*bld*)) had their tissue-specific expression confirmed by IHC. Immunohistochemistry studies were performed using fixed prostate or bladder tissues generated by the Urologic Epithelial Stem Cell Genome Analysis Project (Institute for Systems Biology). The presence of these same proteins in lymphocytes (*lym*) was also tested to assess the possibility of their detection by MS having partially or wholly resulted via lymphocyte infiltration of the tissue.

with the *N*-linked glycosites identified from the other two tissues (bladder cancer and liver metastasis) and plasma, we found that 81 of the *N*-linked glycosites identified were shared among all three tissues. Of these, 57 (70%) were annotated as classical plasma proteins (Fig. 3 and Supplemental Table 1). In contrast, we would expect that the peptides identified from only one of these tissues would be far more likely to represent *bona fide* tissue-derived proteins. Indeed for the 129 *N*-linked glycosites that were uniquely identified in prostate cancer tissue, we found that only seven *N*-linked glycosites (5%) were annotated as classical plasma proteins. These observations again suggested that we were able to identify significant numbers of genuine tissue-derived glycoproteins in both tissue and plasma samples without being overwhelmed by high abundance plasma proteins.

The initial premise for specifically targeting *N*-linked glycosites in our study was 2-fold. First, the reduction in sample complexity achieved by selectively focusing on the subproteome of *N*-linked glycopeptides was expected to improve the detection sensitivity in mass spectrometric analysis of the resulting sample mixtures. Second, the vast majority of intracellular proteins are non-glycosylated, whereas a significant proportion of plasma membrane-bound, extracellular, and secreted proteins, including plasma proteins, are glycosylated. Thus glycoproteins should represent an ideal class of proteins to target for the discovery of new markers of disease that are detectable and quantifiable in the blood.

To test whether we were indeed sampling these expected categories of proteins in our analyses, we applied an informatics approach for the prediction of likely subcellular localization for the glycoproteins identified in the various tissues and cells studied, classifying them into four general groups: 1) cell surface proteins, 2) secreted proteins, 3) transmembrane proteins, and 4) intracellular proteins. We would expect glycoproteins to fall into one of the first three of these groups, and not surprisingly, our analyses confirmed that 1,168 of a total of 1,257 (93%) *N*-linked glycosites identified from tissues, cells, or plasma were classified as such (see Supplemental Table 1). Indeed the true percentage of such proteins in our dataset was likely even higher than 93% because some of the *N*-linked glycosites predicted as intracellular proteins were in fact immunoglobulin isoforms, proteins known to be secreted in actuality. In contrast, applying the same informatics methodology to all 40,110 entries in the human protein sequence database we used for searching our MS/MS data showed that about a third of proteins in the database could be similarly classified (data not shown). These observations thus confirmed our initial premise that the targeted isolation and identification of *N*-linked glycoproteins and glycopeptides significantly enriched for the desired secreted, extracellular, and cell membrane proteins, *i.e.* proteins that likely represent good candidates for both markers of disease and their quantification in the blood. To further reduce the false positive identification of *N*-linked glycosites, the protein subcellular location for the identified *N*-linked glycosites can be further used as a filter to remove the *N*-linked glycosites from intracellular proteins.

Another largely unanswered question relating to blood biomarker discovery was whether the simple, robust, and affordable methodologies required for the necessary high throughput screens were able to access the lower abundance proteins that are generally assumed to be of greater significance for predictive or diagnostic purposes. The data presented here also indicated that by targeting the identification of *N*-linked glycosites we were indeed able to access such lower abundance plasma proteins that also might have originated from specific tissues. A representative list of such proteins is shown in Fig. 4 (see also Supplemental Table 1), including 217 *N*-linked glycosites from cluster designation (CD) cell surface antigens. Of these, 56 *N*-linked glycosites from CD antigens were also identified from plasma samples, and 140 of the *N*-linked glycosites from CD antigens were identified from lymphocyte membranes (Supplemental Table 1). This high proportion of detection in lymphocytes was to be expected because CD antigens were originally characterized as white blood cell surface proteins (25), many of which are now used routinely for typing lymphocytes. However, the expression of many CD antigens is not restricted only to lymphocytes or cells of the hematopoietic system. In this study, we also identified 77 *N*-linked glycosites from CD an-

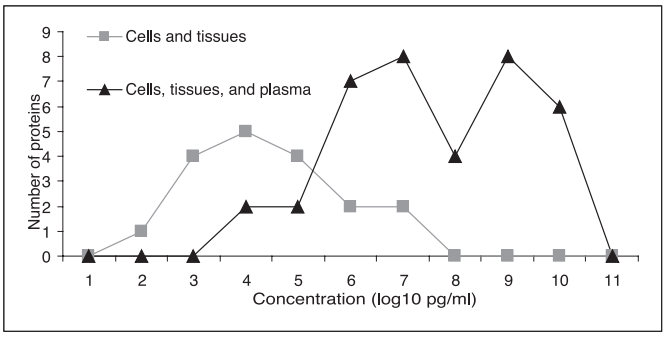



FIG. 5. **The known normal plasma concentration distribution for cell/tissue- and plasma-derived *N*-linked glycoproteins.** The histograms for those proteins identified from both cell/tissue and plasma or from cell/tissue only and that had also recently been shown to be candidate disease markers with known concentrations in normal plasma (27, 28) (also see Supplemental Table 1) are displayed. For convenience, published protein concentrations were binned across sequential plasma concentration ranges each spanning 1 order of magnitude and were plotted on a log scale.

tigens in tissues or cells other than lymphocytes (Supplemental Table 1). Because the expression of some CD antigens on cancer cells has been shown to differ from their normal counterparts, cancer-specific CD antigens found in plasma might also serve as markers for the detection of cancer of specific tissues (26). To confirm that these *N*-linked glycosites from CD antigens identified from tissues were in fact derived from the tissues themselves rather than via contamination from infiltrating lymphocyte proteins present in the tissues, we examined available immunohistochemistry (IHC) data for some of these CD molecules and found that in cases where MS identification had been made from a tissue sample the IHC data were supportive of those findings (Fig. 4).

As an additional test of the sensitivity of our approach toward the identification of lower abundance proteins from cells, tissues, and plasma, we compared our *N*-linked glycosite dataset with recently published literature-derived lists of proteins that have been linked to both cardiac disease and cancer and could thus also represent candidate biomarkers; datasets that also included reported blood concentrations for some of the proteins were also published (27, 28). When we compared these two published datasets with our *N*-linked glycosite dataset presented here, we found 314 *N*-linked glycosites were from 141 candidate biomarkers (see Supplemental Table 1). Of these, normal plasma concentrations were also reported for 56 of these proteins. Several of these proteins detected in both cell/tissue and plasma in our study were known to be present in normal plasma at concentrations in the ng/ml to low $\mu$g/ml range. Such proteins included prothrombin, tissue inhibitor of metalloproteinase 1, von Willebrand factor, tenascin, L-selectin, CD54, and others (Supplemental Table 1). Fig. 5 shows a histogram for these known protein concentrations in normal plasma for the proteins we had also detected in both cells/tissues and plasma or cells/tissues alone. As expected, the proteins identified for which normal blood concentrations were also reported were indeed biased toward the more abundant proteins present in the blood. However, these data also showed that despite this we were nevertheless still able to sample *N*-glycosylated plasma proteins spanning a wide concentration range spanning at least the top 8 orders of magnitude of the full plasma protein concentration range. From these results, we concluded that through targeting *N*-linked glycopeptide enrichment identification via LC-MS/MS we were able to access the lower abundance tissue- and cell-derived proteins that many believe constitute the richest source of potentially new disease markers.

Thus, through the application of solid-phase glycopeptide enrichment and LC-MS, we are clearly able to detect in plasma cell surface CD antigens as well as other molecules known to reflect important physiological information about the state of a particular tissue or cell type. In fact, expression patterns of some CD molecules have already been correlated to disease states of certain tissues, including cancer of the colon, thyroid, and prostate (29–31). Two other proteins identified in this study, the MAC-2-binding protein and metalloproteinase inhibitor 1, have also been identified as potential cancer markers from multiple tissue types with their quantification in blood being of use in monitoring cancer progression (15, 32).

In conclusion, in the present study, we have isolated *N*-linked glycopeptides from tissues, cells, and plasma and have identified the peptide sequences and proteins that they represent via MS-based proteomics. Glycoproteins identified from the individual tissue and cell types were compared with those identified from plasma. In each case, a significant overlap was observed between the tissue/cell glycoproteins and those observed in plasma. Taken together, these data demonstrate that extracellular glycoproteins originating from tissues and cells are released into the blood at levels that are detectable by MS. They also demonstrate that the use of a single, simple solid phase-based enrichment of glycoproteins/glycopeptides from blood plasma, upstream of LC-MS analysis, is sufficient to allow for measurement and profiling of such tissue-derived and cellular proteins in plasma. Thus we have both demonstrated that the largely untested assumption that MS-based proteomics screens are able to detect tissue/cell-derived proteins in the blood is indeed correct and described a methodology capable of accessing such proteins and the potential biological and physiological insights they promise.

*Acknowledgment*—We thank Laurel Feltz for proofreading.

* This work was supported in part with federal funds from the NHLBI, National Institutes of Health, under Contract N01-HV-28179 (to R. A.); with federal funds from NCI, National Institutes of Health, by Grants R21-CA-114852 (to H. Z.) and U01-CA-111244 (to A. L.), and under Contract N01-CO-12400 (to J. W.); by National Institutes of Health Grant R01-AI-41109-01 (to J. W.); and by the Entertainment

Industry Foundation and its Women's Cancer Research Fund (to H. Z.). 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.

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

§ To whom correspondence should be addressed. Current address: Dept. of Pathology, Johns Hopkins Medical Institutions, Baltimore, MD 21287. Tel.: 410-502-8142; E-mail: hzhang32@jhmi.edu.

REFERENCES

1. Etzioni, R., Urban, N., Ramsey, S., McIntosh, M., Schwartz, S., Reid, B., Radich, J., Anderson, G., and Hartwell, L. (2003) The case for early detection. *Nat. Rev. Cancer* **3,** 243–252
2. Schadt, E. E., Lamb, J., Yang, X., Zhu, J., Edwards, S., Guhathakurta, D., Sieberts, S. K., Monks, S., Reitman, M., Zhang, C., Lum, P. Y., Leonardson, A., Thieringer, R., Metzger, J. M., Yang, L., Castle, J., Zhu, H., Kash, S. F., Drake, T. A., Sachs, A., and Lusis, A. J. (2005) An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.* **37,** 710–717
3. Dai, H., van't Veer, L., Lamb, J., He, Y. D., Mao, M., Fine, B. M., Bernards, R., van de Vijver, M., Deutsch, P., Sachs, A., Stoughton, R., and Friend, S. (2005) A cell proliferation signature is a marker of extremely poor outcome in a subpopulation of breast cancer patients. *Cancer Res.* **65,** 4059–4066
4. Omenn, G. S., States, D. J., Adamski, M., Blackwell, T. W., Menon, R., Hermjakob, H., Apweiler, R., Haab, B. B., Simpson, R. J., Eddes, J. S., Kapp, E. A., Moritz, R. L., Chan, D. W., Rai, A. J., Admon, A., Aebersold, R., Eng, J., Hancock, W. S., Hefta, S. A., Meyer, H., Paik, Y. K., Yoo, J. S., Ping, P., Pounds, J., Adkins, J., Qian, X., Wang, R., Wasinger, V., Wu, C. Y., Zhao, X., Zeng, R., Archakov, A., Tsugita, A., Beer, I., Pandey, A., Pisano, M., Andrews, P., Tammen, H., Speicher, D. W., and Hanash, S. M. (2005) Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. *Proteomics* **5,** 3226–3245
5. Anderson, N. L., and Anderson, N. G. (2002) The human plasma proteome: history, character, and diagnostic prospects. *Mol. Cell. Proteomics* **1,** 845–867
6. Tirumalai, R. S., Chan, K. C., Prieto, D. A., Issaq, H. J., Conrads, T. P., and Veenstra, T. D. (2003) Characterization of the low molecular weight human serum proteome. *Mol. Cell. Proteomics* **2,** 1096–1103
7. Nedelkov, D., Kiernan, U. A., Niederkofler, E. E., Tubbs, K. A., and Nelson, R. W. (2005) Investigating diversity in human plasma proteins. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 10852–10857
8. Ludwig, J. A., and Weinstein, J. N. (2005) Biomarkers in cancer staging, prognosis and treatment selection. *Nat. Rev. Cancer* **5,** 845–856
9. Zhang, H., Yi, E. C., Li, X. J., Mallick, P., Kelly-Spratt, K. S., Masselon, C. D., Camp, D. G., II, Smith, R. D., Kemp, C. J., and Aebersold, R. (2005) High throughput quantitative analysis of serum proteins using glycopeptide capture and liquid chromatography mass spectrometry. *Mol. Cell. Proteomics* **4,** 144–155
10. 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
11. Desiere, F., Deutsch, E. W., Nesvizhskii, A. I., Mallick, P., King, N. L., Eng, J. K., Aderem, A., Boyle, R., Brunner, E., Donohoe, S., Fausto, N., Hafen, E., Hood, L., Katze, M. G., Kennedy, K. A., Kregenow, F., Lee, H., Lin, B., Martin, D., Ranish, J. A., Rawlings, D. J., Samelson, L. E., Shiio, Y., Watts, J. D., Wollscheid, B., Wright, M. E., Yan, W., Yang, L., Yi, E. C., Zhang, H., and Aebersold, R. (2005) Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol.* **6,** R9
12. Deutsch, E. W, Eng, J. K., Zhang, H., King, N. L., Nesvizhskii, A. I., Lin, B., Lee, H., Yi, E. C., Ossola, R., and Aebersold, R. (2005) Human Plasma PeptideAtlas. *Proteomics* **5,** 3497–3500
13. Liu, T., Qian, W. J., Gritsenko, M. A., Camp, D. G., II, Monroe, M. E., Moore, R. J., and Smith, R. D. (2005) Human plasma N-glycoproteome analysis by immunoaffinity subtraction, hydrazide chemistry, and mass spectrometry. *J. Proteome Res.* **4,** 2070–2080
14. Han, D. K., Eng, J., Zhou, H., and Aebersold, R. (2001) Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nat. Biotechnol.* **19,** 946–951
15. Liu, A. Y., Zhang, H., Sorensen, C. M., and Diamond, D. L. (2005) Analysis of prostate cancer by proteomics using tissue specimens. *J. Urol.* **173,** 73–78
16. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17,** 994–999
17. Eng, J., McCormack, A. L., and Yates, J. R., III (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5,** 976–989
18. Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74,** 5383–5392
19. Bause, E. (1983) Structural requirements of N-glycosylation of proteins. Studies with proline peptides as conformational probes. *Biochem. J.* **209,** 331–336
20. Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. (1997) A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Int. J. Neural Syst.* **8,** 581–599
21. Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305,** 567–580
22. Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M. P., Walker, J. R., and Hogenesch, J. B. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 6062–6067
23. Su, A. I., Cooke, M. P., Ching, K. A., Hakak, Y., Walker, J. R., Wiltshire, T., Orth, A. P., Vega, R. G., Sapinoso, L. M., Moqrich, A., Patapoutian, A., Hampton, G. M., Schultz, P. G., and Hogenesch, J. B. (2002) Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 4465–4470
24. Zhang, H., Loriaux, P., Eng, J., Campbell, D., Keller, A., Moss, P., Bonneau, R., Zhang, N., Zhou, Y., Wollscheid, B., Cooke, K., Yi, E. C., Lee, H., Peskind, E. R., Zhang, J., Smith, R. D., and Aebersold, R. (2006) UniPep, a database for human N-linked glycosites: a resource for biomarker discovery. *Genome Biol.* **7,** R73
25. True, L. D., and Liu, A. Y. (2003) A challenge for the diagnostic immunohistopathologist. Adding the CD phenotypes to our diagnostic toolbox. *Am. J. Clin. Pathol.* **120,** 13–15
26. Liu, A. Y., Roudier, M. P., and True, L. D. (2004) Heterogeneity in primary and metastatic prostate cancer as defined by cell surface CD profile. *Am. J. Pathol.* **165,** 1543–1556
27. Anderson, L. (2005) Candidate-based proteomics in the search for biomarkers of cardiovascular disease. *J. Physiol.* **563,** 23–60
28. Polanski, M., and Anderson, N. A. (2006) A list of candidate cancer biomarkers for targeted proteomics. *Biomarker Insights* **2,** 1–48
29. Weichert, W., Knosel, T., Bellach, J., Dietel, M., and Kristiansen, G. (2004) ALCAM/CD166 is overexpressed in colorectal carcinoma and correlates with shortened patient survival. *J. Clin. Pathol.* **57,** 1160–1164
30. Kholova, I., Ryska, A., Ludvikova, M., Pecen, L., and Cap, J. (2003) Dipeptidyl peptidase IV (DPP IV, CD 26): a tumor marker in cytologic and histopathologic diagnosis of lesions of the thyroid gland. *Cas. Lek. Cesk.* **142,** 167–171
31. Kristiansen, G., Pilarsky, C., Wissmann, C., Stephan, C., Weissbach, L., Loy, V., Loening, S., Dietel, M., and Rosenthal, A. (2003) ALCAM/CD166 is up-regulated in low-grade prostate cancer and progressively lost in high-grade lesions. *Prostate* **54,** 34–43
32. 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