Published in final edited form as:

J Proteome Res. 2012 February 3; 11(2): 656–667. doi:10.1021/pr201041j.

Glycoprotein Profiles of Human Breast Cells Demonstrate a Clear Clustering of Normal/Benign versus Malignant Cell Lines and Basal versus Luminal Cell Lines

Ten-Yang Yen, Bruce A. Macher*, Claudia A. McDonald, Chris Alleyne-Chin, and Leslie C. Timpe

Department of Chemistry and Biochemistry, San Francisco State University, San Francisco, California 94132, United States

Abstract

Gene expression profiling has defined molecular subtypes of breast cancer including those identified as luminal and basal. To determine if glycoproteins distinguish various subtypes of breast cancer, we obtained glycoprotein profiles from 14 breast cell lines. Unsupervised hierarchical cluster analysis demonstrated that the glycoprotein profiles obtained can serve as molecular signatures to classify subtypes of breast cancer, as well as to distinguish normal and benign breast cells from breast cancer cells. Statistical analyses were used to identify glycoproteins that are overexpressed in normal versus cancer breast cells, and those that are overexpressed in luminal versus basal breast cancer. Among the glycoproteins distinguishing normal breast cells from cancer cells are several proteins known to be involved in cell adhesion, including proteins previously identified as being altered in breast cancer. Basal breast cancer cell lines overexpressed a number of CD antigens, including several integrin subunits, relative to luminal breast cancer cell lines, whereas luminal breast cancer cells overexpressed carbonic anhydrase 12, clusterin, and cell adhesion molecule 1. The differential expression of glycoproteins in these breast cancer cell lines readily allows the classification of the lines into normal, benign, malignant, basal, and luminal groups.

© 2011 American Chemical Society

*Corresponding Author. Telephone: 415-338-6078. Fax: 415-338-6253. macher@sfsu.edu.

ASSOCIATED CONTENT

Supporting Information

Protein and peptide identifications are included in tables. This material is available free of charge via the Internet at <http://pubs.acs.org>.

Keywords

breast cancer; glycoproteomics; protein networks; differential expression; nonmalignant vs malignant; luminal vs basal; hydrazide-modified magnetic beads; label-free quantitation; spectral counts; hierarchical clustering

INTRODUCTION

Breast cancer is the most prevalent of all cancers in American women. It is a complex disease that is heterogeneous with respect to type of tumor, chance of recurrence, and likely response to therapy. Breast cancer is also heterogeneous in the patterns of mRNA and protein expression found in tumors. An important research goal is to identify mRNA or protein biomarkers that provide clinically useful information about the diagnosis, prognosis, or response to treatment of breast cancer.

Messenger RNA and proteomics data are both being used in the search for breast cancer biomarkers. Messenger RNA expression levels have been used to create molecular taxonomies for the classification of breast cancers,^{1,2} and also to devise prognostic tests (e.g., *MammaPrint*, *Oncotype DX*, reviewed in Weigelt and Dowsett).³ More recently, proteomics data have been collected with similar goals. Some of the information available in the mRNA or protein expression patterns of tumors reflects the histological origin of the tumor cells. Normal ducts and lobules of the breast are formed by an epithelium containing luminal and basal layers. Studies of mRNA expression patterns using microarray technology have shown that breast carcinomas cluster into two groups, basal or luminal, depending on the type of cell that founded the tumor.¹ Studies of protein expression also show a division between basal and luminal tumors.^{4,5} The luminal grouping has been refined further into A and B subtypes.²

Messenger RNA and protein expression can also reveal mutations that contribute to carcinogenesis. Epithelial cells often express the receptor for epidermal growth factor (EGF). A subset of breast carcinoma cells overexpresses the HER2/neu variant of the EGF receptor, a receptor subtype that causes constitutive activation of the pathway and cell proliferation. Overexpression of HER2 indicates a poor prognosis, but these patients are also ones who benefit from treatment with trastuzumab (Herceptin).⁶ Luminal-type breast carcinoma cells frequently express receptors for estrogen, progesterone, or both. These patients have a relatively good prognosis. Furthermore, estrogen receptor (ER) positive tumors respond to treatment with tamoxifen or other agents that reduce activation of the receptor. Patients with triple negative (ER-, PR-, and HER2-) breast cancer typically have a poor prognosis and significant potential for disease recurrence.⁷ At present, immunohistochemistry of tumor sections is widely used to determine the ER, PR, and HER2 status of a patient's tumor.⁸

Biomarkers detectable in the blood have the potential either to provide information about a tumor before surgery or to be used to monitor for recurrence. Proteins found in the blood include not only secreted proteins but also proteins shed from cells or fragments of the extracellular domains of membrane proteins. Extracellular and secreted proteins are likely to be glycoproteins. Some examples of markers in current use for cancer are prostate serum antigen, carcinoembryonic antigen (colorectal cancer), and CA125 (ovarian cancer), all of which are glycoproteins. No blood test for breast cancer detection is currently in clinical use. Markers used in monitoring for breast cancer recurrence such as CA 15-3 and CA 27-29 are not specific for breast cancer and do not provide information about the type of breast

tumor.⁹ If blood tests for diagnosis, prognosis, and monitoring of breast cancer were available that were superior to existing tests, they would be widely used.

We have employed a shotgun proteomics approach to identify candidate biomarkers for breast cancer. The samples studied come from normal breast epithelial cells and from thirteen breast cancer cell lines derived from benign and malignant tumors. Glycoproteins from the cells were captured using the periodate oxidation/hydrazide magnetic bead approach¹⁰ and subjected to electrospray ionization/tandem mass spectrometry to identify their protein components. The results were evaluated to determine whether the glycoproteins produced by breast cancer cells contain sufficient information to distinguish among normal, benign, and cancer cell lines, and if the cell lines can be sub-classified as luminal versus basal based on their glycoprotein profiles. Our results show that these various breast cell types can be distinguished by cluster analysis of their glycoproteins. Furthermore, components of the extracellular matrix, or proteins that interact with it, are often expressed differentially by the different cell types and are potentially useful as biomarkers.

EXPERIMENTAL METHODS

Cell Culture

Eleven breast cancer cell lines (BT474, HCC1143, HCC1954, HCC1937, MCF7, SKBR3, SUM149PT, SUM185, SUM229, T47D, and ZR751), two benign breast tumor cell lines (MCF10A, MCF12A), and normal human mammary epithelial cells (HMEC) were grown in 10 cm culture dishes with 10 mL of culture medium (see Table 1 for details). Cells were grown at 37 °C with 5% CO₂. Each sample (12 mL lysate) for analysis was prepared from 15 culture dishes (10 cm), yielding 1.17 ± 0.58 mg of total protein.

Periodate Oxidation

Intact cells were treated with periodate to oxidize monosaccharides within the carbohydrate chains of secreted and cell surface glycoproteins as previously described.^{11,12} Normal, benign, and breast cancer cell lines were grown to ~90% confluence. The growth medium was aspirated from each plate of cells, and the cells were rinsed once with 5 mL of phosphate buffered saline (PBS). Cells were oxidized with 3 mL of 10 mM NaIO₄ in oxidation buffer (20 mM sodium acetate, 150 mM NaCl, pH 5) in the dark at 25 °C for 1 h with gentle rocking, and the periodate solution removed by aspiration. The cells were washed with 4 mL of PBS with gentle rocking for 2 min, and the PBS buffer removed.

Cell Lysate Preparation

Periodate treated cells were incubated with 0.5 mL of lysis buffer containing a 1% octyl- β -D-1-thioglucopyranoside, 1% protease inhibitor cocktail (Sigma-Aldrich, St. Louis, MO), 20 mM sodium acetate, and 150 mM NaCl at pH 5 at room temperature for 1 h with gentle rocking. Cell residue was scraped from the dishes and homogenized by multiple passes through a syringe with a series of different needle sizes ranging from 19 to 27½ gauge. The lysates were clarified by centrifugation at 14 000 rpm for 8 min at 4 °C, the supernatant was carefully collected, total protein content was determined using the Bradford assay, and the supernatant frozen at -20 °C until the samples were enriched for glycoproteins using hydrazide magnetic beads.

Glycoprotein Enrichment using Hydrazide Magnetic Beads

Breast cell lysates were spiked with 5 μ g of periodate oxidized chicken ovalbumin (Sigma-Aldrich, St. Louis, MO), which served as an internal control for the coupling process as previously described.^{11,12} Hydrazide magnetic beads (30 mg, 1 μ m size, Bioclone, San Diego, CA) were diluted to 1 mL with coupling buffer (20 mM sodium acetate, 150 mM

NaCl, pH 5 adjusted using acetic acid) in a 2 mL centrifuge tube. The suspension was vortexed and the beads were recovered with a magnetic separator. The washing step was repeated, the beads were equally divided into eight 2 mL tubes, and each tube was mixed with 1.5 mL lysate. The suspension in the tubes was shaken for 16 h at 25 °C with a thermo mixer at 800 rpm.

Glycoprotein Reduction, Alkylation, Denaturation, and Trypsin Digestion

The magnetic beads were washed (2×, 1 mL per wash) with a series of three washing buffers (1.5 M NaCl, 60% methanol, 60% acetonitrile). Proteins bound to the magnetic beads were denatured with 8 M urea at 25 °C with constant mixing. A rinse of 1 mL of 50 mM ammonium bicarbonate was used to remove the urea. The bound proteins were reduced with 1 mL of 50 mM dithiothreitol for 40 min at 40 °C. The beads were rinsed with 1 mL of 50 mM ammonium bicarbonate buffer, and the reduced glycoproteins were alkylated with 1 mL of 50 mM iodoacetamide for 30 min at 25 °C in the dark with constant mixing. The iodoacetamide was discarded, and the immobilized glycoproteins rinsed with 1 mL of ammonium bicarbonate buffer. The glycoproteins were digested with 0.5 mL of trypsin (20 ng/μL; Promega, Madison, WI) in 50 mM ammonium bicarbonate buffer at 37 °C with constant mixing for 12 h. After digestion, the tryptic fraction was collected, and the hydrazide magnetic beads were washed with 3 mL of 50 mM ammonium bicarbonate to collect any remaining tryptic peptides. The combined eluate was processed by solid phase extraction (SPE).

N-Glycopeptide Release from the Hydrazide Resin with N-Glycosidase F

The N-linked glycopeptides bound to the hydrazide magnetic beads were released from the beads with N-glycosidase F (PNGase F, Glyco N-Glycanase 200 mU, Prozyme) (4 μL in 0.5 mL of 50 mM ammonium bicarbonate buffer) by incubating the solution overnight at 37 °C with constant mixing. The released N-linked glycopeptide fraction was collected and the beads were rinsed with 3 mL of 50 mM ammonium bicarbonate buffer. The ammonium bicarbonate rinse solution was collected and combined with the PNGase F released fraction and subsequently processed by SPE.

Solid Phase Extraction

The tryptic peptide and N-linked glycopeptide (3.5 mL each) fractions were each subjected to SPE (Strata-X reversed phase, Phenomenex). The SPE resin was activated with 1 mL of HPLC grade methanol and rinsed with 1 mL of HPLC grade water to remove the remaining methanol. The tryptic peptides, or PNGase F treated N-linked glycopeptides, were loaded onto and bound to an SPE column using a flow rate of 1.2 mL/min. The resin-bound peptides were rinsed with 2 mL of HPLC grade water to remove salts, eluted with 1 mL of 65% methanol/water, dried using a Speed-Vac apparatus, and stored at 4 °C prior to mass spectrometric analysis.

Peptide/Protein Identifications by ESI-MS/MS Analysis

Tryptic peptides and the deglycosylated N-linked peptides derived from each cell lysate were separately analyzed using an extensive set of LC/MS/MS analyses to maximize glycoprotein identification and to improve the reproducible detection of low abundance glycoproteins.^{13–15} The dried samples of tryptic peptides and the PNGase F treated N-linked glycopeptides were dissolved with 75 μL and 50 μL of 0.1% formic acid/water, respectively, and subsequently analyzed by liquid chromatography/electrospray ionization-tandem mass spectrometry (LC/ESI-MS/MS) using a Thermo LTQ ion trap mass spectrometer with dual Thermo Surveyor HPLC pump systems (Thermo Fisher, San Jose, CA) using an array of instrument settings. The use of a new NanoLC C18 column for each cell lysate sample

eliminated potential peptide carry-over. Tryptic and PNGase F released samples were subjected to a series of LC/ESI-MS/MS analyses with five different settings of the dynamic exclusion (DE) rule for the number of MS/MS spectra acquired with DE = 2 and an excluded time window of 45s, or DE = 1 and an excluded time window of 30, 45, 60, and 90. In addition, three different gas fractionation settings with mass ranges of m/z 400–900, m/z 700–1200, or m/z 1000–1800 with DE = 1 and an excluded time window of 60 s were employed to analyze tryptic and PNGase F released samples. LC/ESI-MS/MS analyses were conducted using a NanoLC reverse phase C18 column (75 $\mu\text{m} \times 130$ mm). The mobile phases for the reverse phase chromatography were (mobile phase A) 0.1% HCOOH/water and (mobile phase B) 0.1% HCOOH in acetonitrile. A four-step, linear gradient was used for nanoLC separation (5% to 35% B in the first 65 min, followed by 35% to 80% B in the next 10 min, holding at 80% B for 5 min, and return to 5% B during the final 10 min). The ESI-MS/MS data acquisition was set to collect ion signals from the eluted peptides using an automatic, data-dependent scan procedure in which a cyclical series of three different scan modes (1 full scan, 4 zoom scans, and 4 MS/MS scans of the four most abundant ions) was performed. The full scan mass range was set from m/z 400 to 1800. Online 2D-LC/ESI-MS/MS analyses were also conducted to analyze the trypsin-digested fraction. The online 2D-LC/ESI-MS/MS analysis chromatographically separated (Thermo Fisher, BioBasic SCX column, 320 $\mu\text{m} \times 100$ mm) the tryptic peptide mixtures into seven fractions using various concentrations (0, 20, 40, 60, 80, 200, 400 mM, and second wash 400 mM) of NH_4Cl . The resulting fractions were desalted in-line with a C18 trap column (300 $\mu\text{m} \times 5$ mm, Agilent), and each fraction was chromatographically resolved by reverse phase nanoLC separation using a linear gradient program similar to 1D-LC described above except that the elution time increased from 90 to 120 min. Since the use of multiple algorithms has been shown to reduce the number of false positive identifications and increase the number of protein identifications,^{16,17} two algorithms (Mascot (v2.3)¹⁸ and X!Tandem (2007.01.01.1)¹⁹) were used to identify peptides from the resulting MS/MS spectra by searching against the combined human protein database (total 22 673 proteins) extracted from SwissProt (v57.14; 2010 February) using taxonomy “homo sapiens” (22 670 proteins). BSA and fetuin provided a means for estimating the level of protein contamination resulting from fetal bovine serum proteins contained in the cell culture medium. Ovalbumin was used to estimate glycoprotein recovery. Searching parameters were set as follows: parent and fragment ion tolerances of 1.6 and 0.8 Da, respectively; carbamidomethyl (+57 Da) modification of Cys as a fixed modification; deamidation (+1 Da) of Asn or Gln, and oxidation of Met as variable modifications; trypsin as the protease with a maximum of 2 missed cleavages. Scaffold (Proteome Software) was used to merge and summarize the data obtained from the 24 runs of LC/MS/MS protein identification analyses for each cell lysate preparation (8 $\times$ 1D-LC/MS/MS acquisition for the PNGaseF released sample; 8 $\times$ 1D-LC/MS/MS acquisition; and 8 $\times$ 2D-LC/MS/MS acquisition for the tryptic digested sample). Protein identifications were based on a minimum detection of two peptides with 99% protein identification probability using the algorithm ProteinProphet.²⁰ Each peptide identified had a minimum peptide identification probability of 95% using the algorithm PeptideProphet²¹ (peptides identified from 14 cell lines and their biological replicates are listed in Supporting Information Table 1, and a summary list of N-linked glycopeptides identified in Supporting Information Table 2). The average false positive rate for the peptide identification in this study was $2.65 \pm 2.42\%$ based on results obtained with PeptideProphet. ProteinID Finder (Proteome Solutions) was used to determine whether the peptide was derived from a glycoprotein, and to extract protein information such as subcellular location, from the UniProt database for each identified protein.

Statistical Analyses

MultiExperiment Viewer (MeV, v4.6) program was used to perform hierarchical clustering analysis.^{22,23} Log₂ and Pearson correlation metric were selected for hierarchical clustering analysis. For the clustering and for the heat maps, a value of 0.5 spectral counts was added to all spectral counts, to avoid the log₂ calculation problem derived from a zero value for glycoproteins that were undetected in some samples. Multidimensional scaling was done using GGobi and GGvis.²⁴ To compare HMEC and breast cancer cell line expression for each glycoprotein a *t* statistic was calculated

$$t = (x_{\text{HMEC}} - m) / (\text{sem})$$

where x_{HMEC} = observed spectral counts for that protein in HMEC, m = the sample mean over 13 samples, and sem = the standard error of the mean. Glycoproteins with fewer than 10 total spectral counts were excluded from this comparison. The glycoproteins were then sorted on the *t* statistics. Significance analysis for microarrays (SAM)²⁵ was performed using the MeV program.²³

Flow Cytometry

Confluent cells were rinsed twice with calcium and magnesium-free phosphate buffered saline (PBS), and incubated for 15–30 min with 5 mM EDTA to release them from the plate. The cells were passed through a 40 µm strainer, pelleted at 350g for 4 min at 4 °C, and resuspended in buffer (1% bovine serum albumin, 2% fetal bovine serum, 400 µM EDTA, 0.1% NaN₃, in PBS). This wash was repeated, and cell number determined using a TC10 automated counter (Bio-Rad Laboratories).

The fluorescently conjugated mouse monoclonal antibodies used for flow cytometry were from BD Biosciences and included antibodies against CD29 (cat. #559883), CD44 (cat. #555478), and CD49c (cat. #556025). The cells (250 000–800 000) were incubated for 30 min with either one of the fluorescently conjugated mouse monoclonal antibodies or isotype controls (cat. # 555751, 555742, and 555749). Cells were washed twice and suspended in 250 µL buffer. Propidium iodide staining for cell viability was used at 270 ng/mL, and samples were analyzed on a BD FACSCalibur using CellQuest Pro v5.2.1.

Excitation of the fluorophore was obtained with either a 488 or 635 nm laser, and emissions detected with filter settings of 530/30 (FITC), 585/42 (PE), 670LP (PI), and 661/16 (APC). Propidium iodide negative cells were gated out and antigen expression measured by subtracting the median fluorescence intensity (MFI) of the isotype control from that of the samples incubated with primary antibody.

RESULTS

Enrichment of Cell Surface and Secreted Glycoproteins from Breast Cancer Cell Lines

We have developed a highly effective and efficient approach to enrich for glycoproteins that are secreted or expressed at the cell surface based on the periodate oxidation-hydrazide chemistry method originated by Zhang et al.²⁶ This approach is summarized in Figure 1 and is based on our published work (Arcinas et al.,¹¹ McDonald et al.¹²). The novel step in our protocol involves the periodate oxidation of carbohydrate residues present on glycoproteins that are exposed on the external cell surface or in the extracellular space of living cells grown in a monolayer. Our initial protocol used hydrazide-Sepharose to covalently capture oxidized glycoproteins following detergent solubilization of the cells. More recently, hydrazide-conjugated magnetic beads^{10,27} have been substituted as a matrix for the covalent

capture of glycoproteins, resulting in better yields of glycoproteins, reduced sample processing time and chemical reagent consumption, as well as improved sample-to-sample reproducibility. Glycoproteins were identified based on a minimum of two unique peptide sequences that matched proteins listed as glycoproteins within the UniProt database.

Tryptic peptides and the deglycosylated N-linked peptides derived from each cell lysate were separately analyzed using an extensive set of LC/MS/MS analyses to maximize glycoprotein identification and to improve the reproducible detection of low abundance glycoproteins.^{13–15} The use of a new NanoLC C18 column for each cell lysate sample eliminated potential peptide carry-over from sample to sample. Scaffold was used to merge together 24 LC/MS/MS data sets for each cell lysate, derived from the protein identification algorithms Mascot and X!Tandem, and to calculate total spectral counts for each identified glycoprotein based on the PeptideProphet algorithm with 95% peptide identification probability.^{18,19,21} Three hundred and fifty nine glycoproteins were identified based on the PNGase F released N-linked glycopeptides alone. Among these 359 proteins, 335 proteins were also found in tryptic peptide fraction and 24 glycoproteins were exclusively identified in PNGase F released peptides. The combination of tryptic peptides and PNGase F released N-linked glycopeptides (Supporting Information Tables 1 and 2) increased the total number of glycoproteins identified from 359 to 486 (Supporting Information Table 3). Estimates of the relative abundance of glycoproteins were made from the total number of assigned MS/MS spectra (spectral counts) for the tryptic and PNGase F released peptides associated with each identified glycoprotein. Combining the data for the tryptic and N-linked peptides increased the number of spectral counts for the proteins, which improved the statistical significance of label-free quantification (spectral counting method) as shown by Old et al.²⁸ Comparison of the spectral counts for one glycoprotein in different samples provides a good estimate of their relative abundances.^{28,29}

Overview of Data Set

Glycoprotein expression was analyzed in 14 cell lines (Table 1) including 1 from normal breast epithelial cells expanded in culture, 2 cell lines derived from benign tumors, and 11 lines derived from malignant tumors. Lysates from two biological replicates of each cell line were analyzed, with the exception of HCC1143 which was analyzed once. A total of 486 glycoproteins (excluding HLA antigen sequences) were identified among the 14 cell lines (Supporting Information Table 3). On average 162 ± 30 glycoproteins were detected per lysate. The distribution of spectral counts per glycoprotein over the entire data set was roughly exponential, with a range of 2–3104 counts and a mean of 15.1 spectral counts (Figure 2A). The distribution had a long tail to the right, corresponding to a small number of glycoproteins with very high spectral counts. Some of the glycoproteins highly expressed in these cell lines include cathepsin D, CD29, CD36, CD51, CD98, CD107a, CD107b, CD147, CD222, and CD276. Based on spectral counts, cathepsin D had the highest expression level among the glycoproteins detected, with an average spectral count of about 700 counts for each cell line.

Figure 2B shows the range of each pair of spectral count measurements plotted as a function of their mean for all glycoproteins for which measurements from two biological replicates were available. There was a tendency for the spread of spectral counts for a single glycoprotein to increase as the number of spectral counts increases. The plot shows that there is a positive association between the mean and range values of the spectral counts for the set of glycoproteins. This type of relationship has been observed in other proteomics data sets, as well as in mRNA expression data sets, and provides a justification for applying the data analysis techniques developed for mRNA to proteomics data sets.³⁰

Many of the glycoproteins in this data set are secreted proteins and plasma membrane proteins (Supporting Information Table 4). The data includes 33 cell adhesion molecules as assigned by the KEGG pathway database,³¹ 28 proteins in the extracellular matrix receptor interaction pathway, 26 in the focal adhesion pathway and others associated with cancer pathways. The largest subset of glycoproteins is one that maps to the lysosomal KEGG pathway (Supporting Information Table 4). This subset includes lysosomal proteins known to traffic to the cell surface (e.g., LAMP proteins) and those that are known to be secreted by cells in culture.^{32,33}

Relationships among Cell Lines

To examine the relationships from the biological replicates of the same cell type relative to those of other cell lines, the Euclidean distance between two samples was calculated using spectral counts from the 486 glycoproteins. To reduce the effect of glycoproteins with very high spectral counts on the distance measurements, the calculation was performed using the base two logarithm of the spectral counts. The 14 samples can be thought of as a cloud of points in a space of 486 dimensions. The technique of multidimensional scaling reduces the dimension of the space to two in a way that preserves the original distances as much as possible.²⁴ Figure 3 shows the result of a multidimensional scaling analysis on the 27 breast cancer samples. It is clear that most pairs of samples from the same cell line are closest to each other (with the exceptions of SUM149 and ZR751), indicating that the glycoprotein spectral counts are most similar for samples from the same cell line. Hence, the pairs of measurements on the same cell lines give very similar results. Furthermore, the figure shows that nonmalignant basal samples (green circles), malignant basal samples (blue squares), and luminal samples (red triangles) generally cluster together.

The relative abundances of the glycoproteins for each cell line are shown in a heat map (Figure 4A). Within the heatmap, both the samples (columns) and glycoproteins (rows) have been ordered by hierarchical clustering. Spectral counts from replicates have been averaged, reducing the number of samples to 14. The figure shows the distribution of spectral counts in the basal cell lines (left columns) and the luminal cell lines (right columns). Red bars on the side of the heatmap designate groups of glycoproteins primarily expressed in basal cell lines. The light blue bar labels a group of glycoproteins in which the expression is higher in luminal cell lines than in basal cell lines, whereas the dark blue bar at the top identifies glycoproteins expressed at similar levels in nearly all cell lines. The green bars denote glycoproteins expressed primarily in individual lines, whereas those expressed primarily in normal or nonmalignant samples are denoted by the dark fuchsia bar. The data set clearly includes glycoprotein expression profiles that are expressed differentially in normal, nonmalignant, malignant, basal, and luminal cells.

Figure 4B displays the clustering tree for the samples. Two conclusions can be drawn from this tree: (1) Two major divisions are observed, one bracketed by HMEC on the left and SUM149 on the right, and the second bracketed by SUM185 on the left and SUM229 on the right, exactly reproducing the basal/luminal groupings identified by mRNA expression studies for these cell lines.^{34,35} Thus, the information in the glycoprotein expression profiles is consistent with previous classification of the cell lines as basal or luminal. (2) The normal (HMEC) and nonmalignant (MCF10A, MCF12A) lines form a separate subtree of the basal group. Thus, nonmalignant lines are separated from the malignant cell lines within the same subtree of basal samples. These results agree with the multidimensional scaling analysis of the unaveraged data (Figure 3), which shows the identical grouping of basal versus luminal, and a similar grouping of the nonmalignant subtree within the basal samples. The similarity of the clustering result using two different methods demonstrates that the basal/luminal and nonmalignant/malignant distinctions have strong support in the glycoprotein data.

A Comparison of Normal versus Tumor Cells

The human mammary epithelial cells studied here were derived from normal breast tissue, whereas the breast cancer cell lines originate from tumors. To determine how the pattern of glycoprotein expression differed between these two types of cells, the expression level of each glycoprotein in the HMEC cells was compared to the levels in breast cancer cell lines using a one-sample *t* test (Supporting Information Table 5). Using 5% as the level of significance, one would expect to identify approximately 24 glycoproteins (of 486) as false positives. However, the observed number was 149, split 77 to 72 between large and small values of *t*, respectively. Thus, the HMEC samples included dozens of glycoproteins present at significantly higher or lower levels than that observed in the breast cell lines. *q*-values (Storey and Tibshirani (2003)³⁶) were used to evaluate the statistical significance of expression differences for the 486 glycoproteins. Figure 5 is a heat map of the 12 proteins with the lowest *q*-values (high expression in HMEC, little or no in the breast cancer cell lines) and the 6 proteins with lowest HMEC expression compared to the breast cancer cell lines (all *q*-values <0.01). Many of the proteins expressed differentially in the cell lines have previously been recognized as contributing to cancer or being altered in cancer, including Thy-1, fibronectin, laminins, thrombospondin 1, and epithelial cell adhesion molecule (see Discussion). Glycoproteins that were highly expressed in HMEC compared to the cancer cell lines also include several extracellular matrix components: fibronectin, three laminins, and a collagen. Differentially expressed glycoproteins not listed in Figure 5 (see Supporting Information Table 5) include *HER2/neu*, which was detected at significant levels in all of the cell lines that were previously reported to overexpress *HER2* (HCC1954, BT474, and SKBR3). Epithelial cell adhesion molecule (EPCAM/CD326) is notable in that it was present in all tumor cell lines, but not in HMEC (Figure 5). Overall, the proteins that were differentially expressed in HMEC cells had a distribution of functions similar to that of the entire data set, as determined by mapping to the KEGG pathways.

Differences between Malignant and Nonmalignant Samples

The breast cancer cell lines include 11 lines derived from malignant tumors and two lines derived from fibrocystic tumors, which are not malignant. To determine which glycoproteins are expressed differentially in the malignant compared to the nonmalignant samples (HMEC cell samples were grouped with the nonmalignant tumors for this analysis), a comparison between the two groups was performed using significance analysis for microarrays (SAM) method,²⁵ a method that calculates a modified *t* statistic for the difference in the means of the two groups for each glycoprotein, then finds the significance using a permutation test. SAM analysis identified only one glycoprotein, basal cell adhesion molecule, with a significant difference in expression level between nonmalignant and malignant samples (*q*-value < 0.001).

Differences between Basal and Luminal Cell Lines

A similar analysis using SAM revealed glycoproteins that were enriched either in basal or luminal cell lines as shown in Figure 6 (all *q*-values < 0.03). For example, clusterin and carbonic anhydrase 12 (a zinc metalloenzyme responsible for acidification of the microenvironment of cancer cells) were more highly expressed in luminal cell lines, whereas several CD antigens, CD142 (tissue factor), CD44, CD147 (Basigin), CD97 antigen, CD109, CD13 (Aminopeptidase N), and three integrins CD49c (integrin α -3), CD49f (integrin α -6), and CD29 (integrin β -1) were found at higher levels in basal cell lines. One glycoprotein, tissue factor (P13726), clearly distinguished all luminal cell lines from all basal cell lines.

Basal cancer cell lines have been further subdivided into subtypes A and B (this division has not been reported in tumors).³⁵ Although hierarchical clustering analysis did not result in a clear separation of basal A and basal B cell lines, two glycoproteins were found to have

substantially different expression levels between these two basal subtypes. CD13 (Aminopeptidase N) was highly expressed in basal B cell lines, whereas galectin-3-binding protein was more highly expressed in Basal A cell lines (Figure 7A and B). Aminopeptidase N had a very low expression level in luminal cells and thus, represents a basal B associated glycoprotein, whereas galectin-3-binding protein had significant expression levels in luminal cell lines.

Correlation between Antibody-Mediated Measurements and Spectral Counts

Fluorescently labeled antiglycoprotein antibodies were used to measure the cell surface expression of CD29, CD44, and CD49c in a subset of breast cell lines that had a large range of expression levels for these three glycoproteins. Figure 8 shows plots of the median fluorescence intensity measurements obtained with the antiglycoprotein antibodies evaluated by flow cytometry versus spectral counts obtained for the three glycoproteins in the cell lines identified in the figures. Cell lines used in the flow cytometry were selected to reflect the range of expression levels observed among the breast cancer cell lines based on spectral counts (spectral count values that differed by 10-fold or more at the extremes with other lines falling at intermediate levels). For example, T47D and MCF7 had very low spectral count values for CD44, whereas HCC 1143 had a spectral count value of over 300, and MCF12A fell at an intermediate level. Plotting the spectral count values versus the median fluorescence intensity value demonstrated that there is a significant correlation between these independent measures of CD44 levels in the breast cell lines. Similarly, there is a high correlation between spectral counts for the two integrins (β -1 or CD29, and α -3 or CD49c) and the median fluorescence intensity measured with the anti-integrin antibodies by flow cytometry.

DISCUSSION

Our results demonstrate that the glycoprotein patterns of normal breast cancer cells and breast cancer cell lines vary significantly. In addition, glycoproteins are differentially expressed in benign versus malignant cells, in cell lines of basal versus luminal origin, and in cell lines of basal A and B subtypes. While these distinctions have been made previously using mRNA expression arrays or immunohistochemistry on tumor sections, there are few other attempts to take a data-driven, proteomics approach to identify glycoproteins as potential serum biomarkers for breast cancer. It may be possible to exploit this variation in glycoprotein expression to infer some of the characteristics of a patient's cancer through the measurement of glycoproteins in the blood.

The shotgun proteomics analysis described here was developed to discover biomarker candidates located within the extracellular space. Our results demonstrate that this approach results in the identification of a significant number of plasma membrane and secreted proteins (Supporting Information Table 4). Similar approaches have been reported by others;³⁷ however, our method applies the periodate oxidation step to intact cells to enhance the identification of proteins in or facing the extracellular space.

The 14 samples studied included cells derived from both benign and malignant tumors of the breast, and from normal breast epithelium as a normal control. The cell lines from malignant tumors varied in their phenotypes (e.g., HER2, ER, PR expression), as well as basal or luminal origin. Thus, this group of cell lines captures much of the known heterogeneity of breast cancer. To increase the number of glycoproteins identified, and the reliability of the identifications, we used protein search programs, MASCOT and X!Tandem in combination with PeptideProphet. We also used a variety of statistical classification techniques that are suited to the data. The discrimination between basal and luminal cell lines was supported by two different approaches to unsupervised classification, multidimensional scaling and

hierarchical clustering. For supervised classification, we used SAM, which employs a modified t statistic, followed by a permutation test to evaluate the significance of the result. Permutation tests and other resampling methods are potentially very useful in analyzing data from mass spectrometry given the large amount of effort and expense required to collect that data.²⁵

Our study design provides an opportunity to compare the glycoprotein-based stratification of the breast cell lines to the stratification obtained from mRNA expression profiling. Overall, cell lines previously classified as basal or luminal based on mRNA expression were grouped together by glycoprotein profiling.^{34,35} In addition, glycoprotein profiling resulted in a distinct grouping of normal and benign cells from all of the cancer cell lines, similar to the stratification obtained from mRNA expression. These similarities in stratification occurred despite the fact that there is only weak correlation between mRNA and protein expression levels. Association between spectral counts for a glycoprotein and its mRNA expression level was evaluated using our spectral count data and mRNA expression data from Neve et al.³⁵ For cathepsin D, aminopeptidase N, integrin β -1, and epithelial cell adhesion molecule the correlation coefficients were 0.41, 0.69, 0.60, and 0.67, respectively. Hence the association between mRNA and protein levels is positive but weak, as observed by several research groups.^{38–41} Nevertheless, the glycoprotein expression levels provide the same classification of the breast cancer cell lines as the mRNA expression measurements.

Supervised classification of our data set identified glycoproteins that drive the clustering of the various cell lines into basal versus luminal, and normal versus malignant subgroups. Many of the proteins expressed differentially have previously been recognized as contributing to cancer, or as being altered by cancer. Of the glycoproteins expressed in HMEC samples but not in cell lines derived from tumors, Thy-1 was expressed exclusively in HMEC cells. Thy-1 is known to have tumor suppressing activity⁴² and upregulates two other proteins listed in Figure 5, fibronectin and thrombospondin 1. An inverse relation between thrombospondin 1 expression and tumor aggressiveness has been reported.⁴³ Fibronectin is a component of the extracellular matrix that binds to integrin dimers comprising α V and β 1 subunits, which are also on the list of differentially expressed glycoproteins (Supporting Information Table 5). Fibronectin and its receptor have been observed to decline in some tumors.^{44,45} Laminins are a component of the basement membrane secreted by epithelial cells. Three laminins were easily detected in the HMEC cells, but were detected at much lower levels, if at all, in the cancer cell lines. C-type mannose receptor 2 (CD280, Endo180) is a collagen binding and recycling glycoprotein that is expressed in some basal breast tumors, and that promotes tumor growth.⁴⁶ In contrast, epithelial cell adhesion molecule (EpCAM) was highly expressed in tumor-derived samples, and undetected in HMEC. EPCAM expression has been associated with a number of cancers, including breast cancer. Silencing of EpCAM expression in breast cancer cell lines via siRNA resulted in an inhibition of proliferation, migration, and invasion.⁴⁷ A glycoprotein that distinguished malignant from nonmalignant breast cell lines is basal cell adhesion molecule. This glycoprotein is a laminin binding protein that is upregulated in some tumors and in some cell types following malignant transformation.⁴⁸ Unsupervised clustering did not result in a distinction of basal A versus B based on glycoprotein profiling, but a small group of glycoproteins could be used to clearly distinguish these basal subtypes (Figure 7). All these groups of glycoproteins represent potential biomarkers, and since many are either secreted, extracellular matrix proteins, or can be shed from the cell surface they could be found in biofluids such as plasma.

A comparison of our list of potential glycoprotein biomarkers (i.e., basal, luminal, normal, malignant) to a list of glycoproteins previously identified in human plasma^{49–51} and via the plasma data sets at the Peptide Atlas Web site (<http://www.peptideatlas.org>) demonstrates

that approximately half of our candidate biomarker glycoproteins have not been reported to occur in plasma, and another one-third have been detected with fewer than five peptides. Therefore, rising levels of these glycoproteins in blood might provide evidence for breast cancer and/or provide information on the breast cancer subtype.

One of the unexpected results from our analysis is the number of lysosomal proteins identified and their relatively high spectral count levels. More than 60 lysosomal proteins including a large number of enzymes, and other lysosomal proteins were detected. The latter include LAMP1, 2, and 3, proteins that have been shown to be associated with tumormetastasis.^{52,53} Other proteins associated with the trafficking of lysosomal proteins (cation-dependent and cation-independent mannose-6-phosphate receptors) were also identified among the glycoproteins in the breast cell lines. The protein with the highest total spectral counts in the cell lines was the lysosomal protease, cathepsin D. Expression and secretion of procathepsin D has been associated with an increase in proliferation, metastasis, and progression of breast cancer.⁵⁴ Additional analyses of a broader range of breast lines and more normal epithelial cell preparations will be necessary to further refine glycoprotein signatures that distinguish normal versus malignant breast cells.

Finally, we independently verified our mass spectrometry spectral count results for a small group of glycoproteins that had a wide range of expression levels among the breast cell lines. The results presented in Figure 8 clearly show that the spectral counts for CD29, CD44, and CD49c directly correlate with antibody binding to the breast cell lines over a broad spectral count range. Spectral counts varied linearly with the amount of protein in the sample as determined by flow cytometry. This linear relationship supports the use of spectral counts as a label-free method for relative quantification.^{28,29}

Comparison with Other Work

Whelan et al.⁵⁵ used glycoprotein capture with LC-MS/MS to identify glycoproteins in three breast cancer cell lines. Of the three lines, MCF-7 was also used in our study. Whelan et al. prepared crude membranes from the cell lines and then enriched for glycoproteins via the periodate oxidation-hydrazide resin capture method. They identified a total of 25 glycoproteins among the three cell lines. Some of the glycoproteins were only identified in one or two of the cell lines. Among the glycoproteins identified in the basal A line (MD-MBA-468) were two, CD44 and epidermal growth factor receptor, that were found in our analysis as being differentially expressed in basal cell lines. Clusterin, a glycoprotein characterized as being differentially expressed by luminal cell lines in our analysis, was identified by Whelan et al. in the luminal cell line MDA-MB-453.

Kulasingam and Diamandis⁵⁶ collected medium conditioned by three breast cancer cell lines and then identified the proteins present using mass spectrometry. Two of the three lines included in their analyses, MCF10A and BT474, were also analyzed in our experiments. The protocol of Kulasingam and Diamandis did not enrich for glycoproteins, but extracellular or membrane-bound proteins were identified by their annotations in the public databases. Kulasingam and Diamandis found 422 glycoproteins in the conditioned media from three breast cancer cell lines, similar in number to the 486 found in our study of 14 lines. Among the top 100 proteins (not all are glycoproteins) reported by Kulasingam and Diamandis in BT474 and MCF10A cell lines, 46 glycoproteins were also identified in our study. Although the approaches for protein isolation and analysis differed between the two studies, the relative expression level differences observed between BT474 and MCF10A were qualitatively similar for 19 of the glycoproteins identified in both studies.

Bateman et al.⁵⁷ conducted a global proteomic analysis of three breast cancer cell lines, as well as the benign breast cell line MCF10A, which they considered to be a normal-like

control. Fold change relative to MCF10A was the statistic used for identifying proteins that were differentially expressed. Bateman et al. found 82 proteins, including many extracellular matrix proteins, that were detected in all four samples and expressed differentially. Network analysis revealed that many of these proteins cluster with focal adhesion kinase. Focal adhesion kinase is a nonglycoprotein tyrosine kinase localized intracellularly to focal adhesions, and is part of a signaling pathway that is activated by extracellular matrix components including integrins and laminins. Focal adhesion kinase and extracellular matrix proteins that interact with it were identified as among the most important network clusters in the study of Bateman et al., in which there was no experimental bias favoring glycoproteins, secreted proteins, or membrane-associated proteins. Our results (Figure 5) clearly demonstrate that several proteins (laminins $\alpha 3$, $\beta 3$, and $\gamma 2$, fibronectin, and thrombospondin-1) which are differentially expressed in normal breast epithelial cells versus breast cancer cell lines form a network of interactions that have a direct association with focal adhesion kinase (Figure 9). These proteins, as well as several integrins ($\alpha 2$, 3, 6, V, and $\beta 1$) which we also found to be differentially expressed between normal versus cancer lines, have been found to be altered in normal versus breast cancer and are critical for cancer invasiveness.

If any of the differentially expressed glycoproteins identified here are to become useful as a serum biomarker, several questions must be answered.^{58–61} How would the glycoprotein be detected?⁶² A method using antibodies is attractive, but requires suitable antibodies. While there is clear evidence for the differential expression of glycoproteins in these cultured cells, it is not known whether these glycoproteins change expression level during carcinogenesis in vivo. It was noted above that over half of the differentially expressed glycoproteins found in our study have either not been detected at all, or were detected at very low levels, in plasma. It remains to be determined whether these glycoproteins rise to a high enough level to be detectable in the blood of cancer patients, and do so early enough to provide useful information. Finally, it is likely that more information can be obtained from a set of proteins than from individual ones. It will require effective use of statistical classification to find panels of glycoproteins for tests of sufficiently high sensitivity, selectivity, and positive predictive value that they are clinically useful.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Support for this work was provided by grants from the National Institutes of Health, Grant P20 MD000544, and the National Science Foundation, Grant CHEM-0619163. The authors thank Ms. Judi Wong, Ms. Christina Litsakos-Cheung, and Mr. Roger Yen for their help with some of the technical aspects of this work. The authors also thank Dr. Susan Fisher for providing some of the breast cancer cell lines.

ABBREVIATIONS

EGF	epidermal growth factor
ER	estrogen receptor
PR	progesterone receptor
<i>HER2/neu</i>	c-erbB2 variant of the EGF receptor
LC/ESI-MS/MS	liquid chromatography/electrospray ionization-tandem mass spectrometry

DE	dynamic exclusion
PNGase F	peptide- <i>N</i> -glycosidase F
SPE	solid phase extraction
SAM	statistical analysis for microarrays
HMEC	human mammary epithelial cells
CD	cluster of differentiation

REFERENCES

1. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D. Molecular portraits of human breast tumours. *Nature*. 2000; 406(6797):747–752. [PubMed: 10963602]
2. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein Lonning P, Borresen-Dale AL. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. U.S.A.* 2001; 98(19):10869–10874. [PubMed: 11553815]
3. Weigel MT, Dowsett M. Current and emerging biomarkers in breast cancer: prognosis and prediction. *Endocr.-Relat. Cancer*. 2010; 17(4):R245–R262. [PubMed: 20647302]
4. Abd El-Rehim DM, Ball G, Pinder SE, Rakha E, Paish C, Robertson JF, Macmillan D, Blamey RW, Ellis IO. High-throughput protein expression analysis using tissue microarray technology of a large well-characterised series identifies biologically distinct classes of breast cancer confirming recent cDNA expression analyses. *Int. J. Cancer*. 2005; 116(3):340–350. [PubMed: 15818618]
5. Jacquemier J, Ginestier C, Rougemont J, Bardou VJ, Charafe-Jauffret E, Geneix J, Adelaide J, Koki A, Houvenaeghel G, Hassoun J, Maraninchi D, Viens P, Birnbaum D, Bertucci F. Protein expression profiling identifies subclasses of breast cancer and predicts prognosis. *Cancer Res.* 2005; 65(3):767–779. [PubMed: 15705873]
6. Goldhirsch A, Ingle JN, Gelber RD, Coates AS, Thürlimann B, Senn H-J. members, P. Thresholds for therapies: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2009. *Ann. Oncol.* 2009; 20(8):1319–1329. [PubMed: 19535820]
7. Gluz O, Liedtke C, Gottschalk N, Pusztai L, Nitz U, Harbeck N. Triple-negative breast cancer--current status and future directions. *Ann. Oncol.* 2009; 20(12):1913–1927. [PubMed: 19901010]
8. Cianfrocca M, Gradishar W. New molecular classifications of breast cancer. *Ca-Cancer. J. Clin.* 2009; 59(5):303–313. [PubMed: 19729680]
9. Sturgeon CM, Duffy MJ, Stenman U-H, Lilja H, Brunner N, Chan DW, Babaian R, Bast RC Jr, Dowell B, Esteva FJ, Haglund C, Harbeck N, Hayes DF, Holten-Andersen M, Klee GG, Lamerz R, Looijenga LH, Molina R, Nielsen HJ, Rittenhouse H, Semjonow A, Shih I-M, Sibley P, Soletormos G, Stephan C, Sokoll L, Hoffman BR, Diamandis EP. National Academy of Clinical Biochemistry Laboratory Medicine Practice Guidelines for Use of Tumor Markers in Testicular, Prostate, Colorectal, Breast, and Ovarian Cancers. *Clin. Chem.* 2008; 54(12):e11–e79. [PubMed: 19042984]
10. Zou Z, Ibisate M, Zhou Y, Aebbersold R, Xia Y, Zhang H. Synthesis and evaluation of superparamagnetic silica particles for extraction of glycopeptides in the microtiter plate format. *Anal. Chem.* 2008; 80(4):1228–1234. [PubMed: 18197692]
11. Arcinas A, Yen TY, Kebebew E, Macher BA. Cell surface and secreted protein profiles of human thyroid cancer cell lines reveal distinct glycoprotein patterns. *J. Proteome Res.* 2009; 8(8):3958–3968. [PubMed: 19530676]
12. McDonald CA, Yang JY, Marathe V, Yen TY, Macher BA. Combining results from lectin affinity chromatography and glyco-capture approaches substantially improves the coverage of the glycoproteome. *Mol. Cell. Proteomics.* 2009; 8(2):287–301. [PubMed: 18923192]
13. Spahr CS, Davis MT, McGinley MD, Robinson JH, Bures EJ, Beierle J, Mort J, Courchesne PL, Chen K, Wahl RC, Yu W, Luethy R, Patterson SD. Towards defining the urinary proteome using

- liquid chromatography-tandem mass spectrometry. I. Profiling an unfractionated tryptic digest. *Proteomics*. 2001; 1(1):93–107. [PubMed: 11680902]
14. Wolters DA, Washburn MP, Yates JR 3rd. An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.* 2001; 73(23):5683–5690. [PubMed: 11774908]
 15. Zhang Y, Wen Z, Washburn MP, Florens L. Effect of dynamic exclusion duration on spectral count based quantitative proteomics. *Anal. Chem.* 2009; 81(15):6317–6326. [PubMed: 19586016]
 16. Nesvizhskii AI. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Proteomics*. 2010; 73(11):2092–2123. [PubMed: 20816881]
 17. Kapp EA, Schutz F, Connolly LM, Chakel JA, Meza JE, Miller CA, Fenyo D, Eng JK, Adkins JN, Omenn GS, Simpson RJ. An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis. *Proteomics*. 2005; 5(13):3475–3490. [PubMed: 16047398]
 18. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*. 1999; 20(18):3551–3567. [PubMed: 10612281]
 19. Craig R, Beavis RC. A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Commun. Mass Spectrom.* 2003; 17(20):2310–2316. [PubMed: 14558131]
 20. Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* 2003; 75(17):4646–4658. [PubMed: 14632076]
 21. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* 2002; 74(20):5383–5392. [PubMed: 12403597]
 22. Eisen M, Spellman P, Brown P, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.* 1998; 95(25):14863–14868. [PubMed: 9843981]
 23. Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M, Sturn A, Snuffin M, Rezantsev A, Popov D, Ryltsov A, Kostukovich E, Borisovsky I, Liu Z, Vinsavich A, Trush V, Quackenbush J. TM4: a free, open-source system for microarray data management and analysis. *Biotechniques*. 2003; 34(2):374–378. [PubMed: 12613259]
 24. Buja A, Swayne DF, Littman ML, Dean N, Hofmann H, Chen L. Data visualization with multidimensional scaling. *J. Comput. Gr. Stat.* 2008; 17(2):444–472.
 25. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U.S.A.* 2001; 98(9):5116–5121. [PubMed: 11309499]
 26. Zhang H, Li XJ, Martin DB, Aebersold R. Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry. *Nat. Biotechnol.* 2003; 21(6):660–666. [PubMed: 12754519]
 27. Berven FS, Ahmad R, Clauser KR, Carr SA. Optimizing performance of glycopeptide capture for plasma proteomics. *J. Proteome Res.* 2010; 9(4):1706–1715. [PubMed: 20235580]
 28. Old WM, Meyer-Arendt K, Aveline-Wolf L, Pierce KG, Mendoza A, Sevinsky JR, Resing KA, Ahn NG. Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol. Cell. Proteomics*. 2005; 4(10):1487–1502. [PubMed: 15979981]
 29. Liu H, Sadygov RG, Yates JR 3rd. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* 2004; 76(14):4193–4201. [PubMed: 15253663]
 30. Pavelka N, Fournier ML, Swanson SK, Pelizzola M, Ricciardi-Castagnoli P, Florens L, Washburn MP. Statistical similarities between transcriptomics and quantitative shotgun proteomics data. *Mol. Cell. Proteomics*. 2008; 7(4):631–644. [PubMed: 18029349]
 31. Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 2009; 37(1):1–13. [PubMed: 19033363]
 32. Kornfeld S, Mellman I. The biogenesis of lysosomes. *Annu. Rev. Cell Biol.* 1989; 5:483–525. [PubMed: 2557062]

33. Sleat DE, Zheng H, Qian M, Lobel P. Identification of sites of mannose 6-phosphorylation on lysosomal proteins. *Mol. Cell. Proteomics*. 2006; 5(4):686–701. [PubMed: 16399764]
34. Charafe-Jauffret E, Ginestier C, Monville F, Finetti P, Adelaide J, Cervera N, Fekairi S, Xerri L, Jacquemier J, Birnbaum D, Bertucci F. Gene expression profiling of breast cell lines identifies potential new basal markers. *Oncogene*. 2006; 25(15):2273–2284. [PubMed: 16288205]
35. Neve RM, Chin K, Fridlyand J, Yeh J, Baehner FL, Fevr T, Clark L, Bayani N, Coppe JP, Tong F, Speed T, Spellman PT, DeVries S, Lapuk A, Wang NJ, Kuo WL, Stilwell JL, Pinkel D, Albertson DG, Waldman FM, McCormick F, Dickson RB, Johnson MD, Lippman M, Ethier S, Gazdar A, Gray JW. A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell*. 2006; 10(6):515–527. [PubMed: 17157791]
36. Storey JD, Tibshirani R. Statistical significance for genome-wide studies. *Proc. Natl. Acad. Sci. U.S.A.* 2003; 100(16):9440–9445. [PubMed: 12883005]
37. Tian Y, Kelly-Spratt KS, Kemp CJ, Zhang H. Mapping tissue-specific expression of extracellular proteins using systematic glycoproteomic analysis of different mouse tissues. *J. Proteome Res.* 2010; 9(11):5837–5847. [PubMed: 20828161]
38. Anderson L, Seilhamer J. A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis*. 1997; 18(3–4):533–537. [PubMed: 9150937]
39. Gygi SP, Rochon Y, Franz BR, Aebersold R. Correlation between Protein and mRNA Abundance in Yeast. *Mol. Cell. Biol.* 1999; 19(3):1720–1730. [PubMed: 10022859]
40. Ginestier C, Charafe-Jauffret E, Bertucci F, Eisinger F, Geneix J, Bechlian D, Conte N, Adelaide J, Toiron Y, Nguyen C, Viens P, Mozziconacci MJ, Houlgatte R, Birnbaum D, Jacquemier J. Distinct and complementary information provided by use of tissue and DNA microarrays in the study of breast tumor markers. *Am. J. Pathol.* 2002; 161(4):1223–1233. [PubMed: 12368196]
41. Chen G, Gharib TG, Wang H, Huang CC, Kuick R, Thomas DG, Shedden KA, Misek DE, Taylor JM, Giordano TJ, Kardias SL, Iannettoni MD, Yee J, Hogg PJ, Orringer MB, Hanash SM, Beer DG. Protein profiles associated with survival in lung adenocarcinoma. *Proc. Natl. Acad. Sci. U.S.A.* 2003; 100(23):13537–13542. [PubMed: 14573703]
42. Abeysinghe HR, Li LQ, Guckert NL, Reeder J, Wang N. THY-1 induction is associated with up-regulation of fibronectin and thrombospondin-1 in human ovarian cancer. *Cancer Genet. Cytogenet.* 2005; 161(2):151–158. [PubMed: 16102586]
43. Lawler J. Thrombospondin-1 as an endogenous inhibitor of angiogenesis and tumor growth. *J. Cell. Mol. Med.* 2002; 6(1):1–12. [PubMed: 12003665]
44. Engbring JA, Kleinman HK. The basement membrane matrix in malignancy. *J. Pathol.* 2003; 200(4):465–470. [PubMed: 12845613]
45. Williams CM, Engler AJ, Slone RD, Galante LL, Schwarzbauer JE. Fibronectin expression modulates mammary epithelial cell proliferation during acinar differentiation. *Cancer Res.* 2008; 68(9):3185–3192. [PubMed: 18451144]
46. Wienke D, Davies GC, Johnson DA, Sturge J, Lambros MB, Savage K, Elsheikh SE, Green AR, Ellis IO, Robertson D, Reis-Filho JS, Isacke CM. The collagen receptor Endo180 (CD280) is expressed on basal-like breast tumor cells and promotes tumor growth in vivo. *Cancer Res.* 2007; 67(21):10230–10240. [PubMed: 17974964]
47. Osta WA, Chen Y, Mikhitarian K, Mitas M, Salem M, Hannun YA, Cole DJ, Gillanders WE. EpCAM is overexpressed in breast cancer and is a potential target for breast cancer gene therapy. *Cancer Res.* 2004; 64(16):5818–5824. [PubMed: 15313925]
48. Rettig WJ, Dracopoli NC, Goetzger TA, Spengler BA, Biedler JL, Oettgen HF, Old LJ. Somatic cell genetic analysis of human cell surface antigens: chromosomal assignments and regulation of expression in rodent-human hybrid cells. *Proc. Natl. Acad. Sci. U.S.A.* 1984; 81(20):6437–6441. [PubMed: 6593709]
49. Liu T, Qian WJ, Gritsenko MA, Camp DG 2nd, Monroe ME, Moore RJ, Smith RD. Human plasma N-glycoproteome analysis by immunoaffinity subtraction, hydrazide chemistry, and mass spectrometry. *J. Proteome Res.* 2005; 4(6):2070–2080. [PubMed: 16335952]
50. Omenn GS, States DJ, Adamski M, Blackwell TW, Menon R, Hermjakob H, Apweiler R, Haab BB, Simpson RJ, Eddes JS, Kapp EA, Moritz RL, Chan DW, Rai AJ, Admon A, Aebersold R, Eng J, Hancock WS, Hefta SA, Meyer H, Paik YK, Yoo JS, Ping P, Pounds J, Adkins J, Qian X, Wang

- R, Wasinger V, Wu CY, Zhao X, Zeng R, Archakov A, Tsugita A, Beer I, Pandey A, Pisano M, Andrews P, Tammen H, Speicher DW, Hanash SM. 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*. 2005; 5(13):3226–3245. [PubMed: 16104056]
51. Qian W-J, Kaleta DT, Petritis BO, Jiang H, Liu T, Zhang X, Mottaz HM, Varnum SM, Camp DG, Huang L, Fang X, Zhang W-W, Smith RD. Enhanced Detection of Low Abundance Human Plasma Proteins Using a Tandem IgY12-SuperMix Immunoaffinity Separation Strategy. *Mol. Cell. Proteomics*. 2008; 7(10):1963–1973. [PubMed: 18632595]
52. Saitoh O, Wang WC, Lotan R, Fukuda M. Differential glycosylation and cell surface expression of lysosomal membrane glycoproteins in sublines of a human colon cancer exhibiting distinct metastatic potentials. *J. Biol. Chem.* 1992; 267(8):5700–5711. [PubMed: 1544942]
53. Kanao H, Enomoto T, Kimura T, Fujita M, Nakashima R, Ueda Y, Ueno Y, Miyatake T, Yoshizaki T, Buzard GS, Tanigami A, Yoshino K, Murata Y. Overexpression of LAMP3/TSC403/DC-LAMP promotes metastasis in uterine cervical cancer. *Cancer Res.* 2005; 65(19):8640–8645. [PubMed: 16204031]
54. Ohri SS, Vashishta A, Proctor M, Fusek M, Vetvicka V. The propeptide of cathepsin D increases proliferation, invasion and metastasis of breast cancer cells. *Int. J. Oncol.* 2008; 32(2):491–498. [PubMed: 18202773]
55. Whelan SA, Lu M, He J, Yan W, Saxton RE, Faull KF, Whitelegge JP, Chang HR. Mass spectrometry (LC-MS/MS) site-mapping of N-glycosylated membrane proteins for breast cancer biomarkers. *J. Proteome Res.* 2009; 8(8):4151–4160. [PubMed: 19522481]
56. Kulasingam V, Diamandis EP. Proteomics analysis of conditioned media from three breast cancer cell lines: a mine for biomarkers and therapeutic targets. *Mol. Cell. Proteomics*. 2007; 6(11):1997–2011. [PubMed: 17656355]
57. Bateman NW, Sun M, Hood BL, Flint MS, Conrads TP. Defining central themes in breast cancer biology by differential proteomics: conserved regulation of cell spreading and focal adhesion kinase. *J. Proteome Res.* 2010; 9(10):5311–5324. [PubMed: 20681588]
58. Drake PM, Schilling B, Niles RK, Braten M, Johansen E, Liu H, Lerch M, Sorensen DJ, Li B, Allen S, Hall SC, Witkowska HE, Regnier FE, Gibson BW, Fisher SJ. A lectin affinity workflow targeting glycosite-specific, cancer-related carbohydrate structures in trypsin-digested human plasma. *Anal. Biochem.* 2011; 408(1):71–85. [PubMed: 20705048]
59. Rifai N, Gillette MA, Carr SA. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat. Biotechnol.* 2006; 24(8):971–983. [PubMed: 16900146]
60. Anderson NL, Anderson NG, Pearson TW, Borchers CH, Paulovich AG, Patterson SD, Gillette M, Aebersold R, Carr SA. A human proteome detection and quantitation project. *Mol. Cell. Proteomics*. 2009; 8(5):883–886. [PubMed: 19131327]
61. Makawita S, Diamandis EP. The bottleneck in the cancer biomarker pipeline and protein quantification through mass spectrometry-based approaches: current strategies for candidate verification. *Clin. Chem.* 2010; 56(2):212–222. [PubMed: 20007861]
62. Pan S, Chen R, Aebersold R, Brentnall TA. Mass Spectrometry Based Glycoproteomics—From a Proteomics Perspective. *Mol. Cell. Proteomics*. 2011; 10(1):1–14.

Figure 1.

Workflow diagram of the glyco-capture (periodate oxidation/hydrazide magnetic bead) method for profiling the cell surface membrane, and secreted glycoproteins, database searching and statistical classification.

Figure 2.

(A) Distribution of spectral counts per glycoprotein over the entire data set. Spectral counts ranged from 2 to 3104 counts with a mean of 15.1 spectral counts. (B) Plot of the range of each pair of spectral count measurements as a function of their mean for all glycoproteins. There is a positive association between mean and range values.

Figure 3.

Multidimensional scaling analysis on the 27 breast cancer samples. The distance between points indicates their similarity using the Euclidean norm. One half count was added to each spectral count, and the base 2 log taken for the analysis. Biological replicates on the same cell line are generally close to each other, showing that pairs of measurements on the same cell lines give very similar results. Blue boxes are nonmalignant lines. Green circles are basal lines. Red triangles are luminal lines.

Figure 4.
 (A) Heat map of relative abundances of the glycoproteins (253 proteins) for each cell line. Cell line (columns) and glycoproteins (rows) have been ordered by hierarchical clustering. Bars on the left side identify glycoproteins that are expressed at similar levels in nearly all cell lines (dark blue), primarily expressed in basal cell lines (red), luminal cell lines (light blue), in individual lines (green), and in normal or nonmalignant lines (dark fuchsia). (B) Clustering tree resulting from hierarchical clustering shown in (A).

Figure 5.

Glycoproteins differentially expressed between HMEC and breast cancer cell lines. Expression level of glycoprotein in the HMEC cells was compared to the levels in breast cancer cell lines using a one-sample *t* test; the expression profiles of glycoproteins with extreme values of *t* are displayed in a heat map. One half count was added to each spectral count, and the base 2 log taken for constructing the heat map. Expression levels are represented by red for high expression level and green for low expression level. Glycoprotein UniProt IDs and names are given on the right.

Figure 6.

Glycoproteins differentially expressed between basal and luminal cell lines. Basal cell lines are represented in the first seven columns. Expression levels are represented by red for high level expression levels and green for low expression levels. Glycoprotein UniProt IDs and names are given on the right.

Figure 7. Glycoproteins, (A) aminopeptidase (CD13) and (B) galectin-3-binding protein, differentially expressed between basal A and basal B cell lines. Basal A lines are the three HCC cell lines. Spectral counts are shown on the Y-axis (from the average of two biological replicates for each cell type except for HCC1143, which represents spectral counts from a single analysis).

Figure 8.

(A–C) Spectral counts for CD29, CD44, and CD49c directly correlate with antibody binding (median fluorescence intensity) to the breast cell lines. Spectral counts obtained from mass spectrometry analysis for the three CD antigens were plotted against the median fluorescence intensity values obtained from flow cytometry. Measurements for the various breast cancer cell lines are indicated for CD44 (A), CD49c (B), and CD29 (C), and R^2 for the fit of the data to a line are shown. (●) Basal cell lines; (◆) luminal cell lines.

Figure 9. STRING (version 9.0) network showing interactions between glycoproteins that are differentially expressed in normal breast epithelial cells versus breast cancer cell lines and focal adhesion kinase. Gene symbols are given in the figure for glycoproteins that had significantly different expression levels in normal breast cells versus the breast cancer cell lines analyzed and focal adhesion kinase (PTK2).

Table 1Cell Lines Studied in These Experiments^a

cell lines	ER status	PR status	HER2 status	origin	gene cluster
HMEC ^b	NA	NA	NA	human mammary epithelial cells	NA
MCF10A ^c	–	–	–	fibrocystic disease	basal
MCF12Ac	–	–	–	fibrocystic disease	basal
HCC1143 ^d	–	–	–	ductal carcinoma	basal
HCC1937 ^d	–	–	–	ductal carcinoma	basal
HCC1954 ^c	–	–	+	ductal carcinoma	basal
SUM149 ^d	–	–	–	inflammatory ductal carcinoma	basal
BT474 ^c	+	+	+	invasive ductal carcinoma	luminal
MCF7 ^c	+	+	–	invasive ductal carcinoma	luminal
SKBR3 ^c	–	–	+	adenocarcinoma	luminal
SUM185 ^d	–	–	–	ductal carcinoma	luminal
T47D ^c	+	+	–	invasive ductal carcinoma	luminal
ZR751 ^c	+	–	–	invasive ductal carcinoma	luminal
SUM229 ^d	NA	NA	NA	NA	NA

^aThe assignment to gene cluster, ER, PR, and HER2 status are from Neve et al.³⁵ (NA: not available or applicable.)

^bCell line was from Cell Applications, Inc. (San Diego, CA) and was grown according to protocols provided by the vendor.

^cCell lines were from Dr. Susan Fisher at UCSF and were grown according to protocols described by Neve et al.³⁵

^dCell lines were from American Type Culture Collection (ATCC; Manassas, VA) and were grown according to protocols provided by ATCC.