Published in final edited form as:

J Proteome Res. 2010 December 3; 9(12): 6440–6449. doi:10.1021/pr100696n.

Lung Cancer Serum Biomarker Discovery Using Glycoprotein Capture and Liquid Chromatography Mass Spectrometry

Xuemei Zeng¹, Brian L. Hood¹, Mai Sun¹, Thomas P. Conrads^{1,2,7}, Roger S. Day⁵, Joel L. Weissfeld^{3,4}, Jill M. Siegfried^{2,7}, and William L. Bigbee^{1,2,3,6}

¹Mass Spectrometry Platform, Cancer Biomarkers Facility, University of Pittsburgh Cancer Institute

²Lung and Thoracic Malignancies Program, University of Pittsburgh Cancer Institute

³Cancer Epidemiology Program, University of Pittsburgh Cancer Institute

⁴Department of Epidemiology, Graduate School of Public Health, University of Pittsburgh School of Medicine, Pittsburgh, PA

⁵Departments of Biomedical Informatics, University of Pittsburgh School of Medicine, Pittsburgh, PA

⁶Department of Pathology, University of Pittsburgh School of Medicine, Pittsburgh, PA

⁷Department of Pharmacology & Chemical Biology, University of Pittsburgh School of Medicine, Pittsburgh, PA

Abstract

Targeted glycoproteomics represents an attractive approach for conducting peripheral blood based cancer biomarker discovery due to the well-known altered pattern of protein glycosylation in cancer and the reduced complexity of the resultant glycoproteome. Here we report its application to a set of pooled non-small cell lung cancer (NSCLC) case sera (9 adenocarcinoma and 6 squamous cell carcinoma pools from 54 patients) and matched controls pools, including 8 clinical control pools with computed tomography detected nodules but being non-malignant as determined by biopsy from 54 patients, and 8 matched healthy control pools from 106 cancer-free subjects. The goal of the study is to discover biomarkers which may enable improved early detection and diagnosis of lung cancer. Immunoaffinity subtraction was used to first deplete the top most abundant serum proteins; the remaining serum proteins were then subjected to hydrazide chemistry based glycoprotein capture and enrichment. Hydrazide resin *in situ* trypsin digestion was used to release non-glycosylated peptides. Formerly N-linked glycosylated peptides were released by peptide-N-glycosidase F (PNGase F) treatment and were subsequently analyzed by liquid chromatography (LC)-tandem mass spectrometry (MS/MS). A MATLAB[®] based in-house tool was developed to facilitate retention time alignment across different LC-MS/MS runs, determination of precursor ion m/z values and elution profiles, and the integration of mass chromatograms based on determined parameters for identified peptides. A total of 38 glycopeptides from 22 different proteins were significantly differentially abundant across the case/control pools ($P < 0.01$, Student's t test) and their abundances led to a near complete separation of case and control pools based on hierarchical clustering. The differential abundances of three of

*CORRESPONDING AUTHOR FOOTNOTE. To whom correspondence should be addressed. William L Bigbee, Ph.D. Tel: 412-641-7555; Fax: 412-641-2458. bigbeewl@upmc.edu.

SUPPORTING INFORMATION AVAILABLE: Supplemental Table S1, list of glycopeptides identified from the combined 62 LC-MS/MS analyses; Supplemental Table S2, correlation between NSCLC selective proteins and the categories for biological functions/diseases from the IPA tools; Supplemental MATLAB[®] script used for batch-extraction of peptide mass chromatogram.

these candidate proteins were verified by commercially available ELISAs applied in the pools. Strong positive correlations between glycopeptide mass chromatograms and ELISA-measured protein abundance was observed for all of the selected glycoproteins.

Keywords

Lung cancer; serum biomarkers; glycoproteomics; LC-MS/MS; mass chromatogram

1. Introduction

Lung cancer is the leading cause of cancer deaths in the US partly due to its high fatality rate, with an overall 5-year survival rate of 15.6% (Surveillance Epidemiology and End Results; SEER). Poor survival from lung cancer is strongly related to the late stage at clinical presentation. The majority of lung cancer patients are diagnosed after their primary lung tumors have metastasized and have only 3.5% five-year survival rate. In contrast, early stage lung cancer has relatively favorable survival. A recent review reported that Stage 1A patients, with tumor size <10mm, who underwent complete resection, experienced an 86% overall and 100% cancer specific 5-year survival.¹ However, no current lung cancer screening is recommended by any major medical society² due to the lack of sensitive and specific screening technologies. New and more sensitive computed tomography (CT) imaging technology makes detection of early lung cancer feasible;³ however its clinical utility is limited by its limited ability of differentiating benign pulmonary nodules from lung cancer.⁴⁻⁵ Serum based biomarkers that could help guide clinical decision-making regarding the level of lung cancer risk in high-risk subjects and particularly in patients with indeterminate pulmonary nodules detected with screening CT are urgently needed to reduce lung cancer related mortality and the burden on clinical screening resources.

Peripheral blood constantly circulates through the whole body and carries important disease indicators such as shed and aberrantly secreted proteins from various tissues/organs including tumors. A number of blood-based cancer biomarkers have been identified and are being used clinically, such as prostate-specific antigen (PSA) for prostate cancer,⁶ cancer antigen CA15.3 for breast cancer,⁷ CA125 for ovarian cancer,⁸ and carcinoembryonic antigen (CEA) for colorectal cancer,⁹ supporting the enormous potential of using blood for biomarker discovery and clinical screening. Proteomics based biomarker discovery has advanced significantly recently, fueled by increasingly robust technologies, such as gel- and liquid chromatography (LC)-based protein/peptide separations and mass spectrometry (MS) instrumentation, allowing simultaneous analysis of complex protein/peptide samples. However blood-based biomarker discovery is still hampered by substantial technical and biological challenges such as the over ten orders of magnitude dynamic range of abundance and extreme complexity of the blood proteome.¹⁰

One way to circumvent these challenges is a subproteome approach that targets a specific subset of serum proteins to reduce the complexity of the serum proteome.¹¹ In particular, focus on the glycoproteome, which targets the analysis of glycosylated proteins, has received wide research interest since most known cancer biomarkers such as PSA, CA15-3, CA-125, and CEA are glycoproteins. Glycosylation is one of the most common and heterogeneous post-translational modifications in proteins. Over 50% of known eukaryotic proteins are estimated to be glycosylated.¹² In addition, cancer cells are known to express aberrant glycosylation patterns such as increased branching, sialylation, and/or fucosylation of N-linked glycans, and increased sialylation and/or truncation of O-linked glycans.¹³ Two major strategies have been developed to enrich glycoproteins or glycopeptides, chemical capture approaches based on the reaction between aldehyde groups and hydrazide,¹⁴⁻¹⁶ and

lectin affinity capture based on the specific recognition of protein glycan moieties by lectins.¹⁷

Aiming to identify non-invasive lung cancer specific biomarkers, three studies have reported application of comparative glycoproteome analysis for lung cancer serum biomarker discovery.^{18–20} All three of these published studies used lectin affinity based capture to enrich glycoproteins from serum samples and then compared the relative abundance of the enriched glycoproteins in lung adenocarcinoma samples vs. controls. Many overlapping potential candidate biomarkers were identified in these studies, supporting the important potential of glycoproteome for biomarker discovery. However all of these studies had relatively small numbers of samples, ranging from 3 to 10 individual subjects for each sample group (cancer or control). Additionally the comparison of these studies was limited to lung cancer versus healthy controls only. The clinical utility of these candidate biomarkers in predicting lung cancer risk in high risk subjects such as those with indeterminate pulmonary nodules remains to be determined.

With the coupling of advanced capillary-based LC-separations online with MS/MS analyses, identification of large numbers of peptides/proteins is now becoming a routine practice for LC-MS/MS based proteomics. Label free relative quantitation, which does not require up-front isotopic labeling and permits retrospective comparison, is gaining increasing popularity for profiling relative changes in peptide/protein abundance across different samples analyzed in a LC-MS/MS workflow. Two common quantitation metrics are used in label-free LC-MS/MS experiments: integration of peptide ion mass chromatograms and spectral counting (number of MS/MS spectra assigned to a protein), each with its pros and cons.^{21–22} In particular, spectral counting is limited by the nature of data acquisition during LC-MS/MS analysis. Low abundant peptides may not be selected for MS/MS identifications, while high abundant peptides may be subjected to dynamic exclusion. On the other hand, peptide mass chromatogram integration is sensitive to the presence of other precursor ions with irresolvable mass-to-charge ratio (m/z) values and elution profiles, especially for complex samples, and the retention time variation for different LC-MS/MS runs.

In the current discovery study, we applied a comparative glycoproteomics analysis to a large set of well annotated non small cell lung cancer (NSCLC) pooled sera, including both adenocarcinoma and squamous cell carcinoma, and matched control pooled sera to identify potential lung cancer specific serum proteins. Different from previous reported studies, we included subjects who had a CT detected nodule but only with non-malignant lung disease as confirmed by biopsy in our control group. In addition, we targeted our analysis to individual glycopeptides rather than glycoproteins, thus allowing us to detect differential expression for different glycosylation sites. The general process for our experimental procedures and data analyses is illustrated in Figure 1. In brief, we used hydrazide chemistry based glycoprotein capture followed by a resin *in situ* trypsin digestion approach to enrich glycopeptides from the pooled serum samples. The enriched glycopeptides were released from the hydrazide resin through cleavage of the glycans using peptide-N-glycosidase F (PNGase F) and then analyzed using nano-LC coupled to high mass resolution ESI-MS/MS (LTQ/Orbitrap, ThermoFisher Scientific Inc., San Jose, CA). An in-house MATLAB® based tool was developed to facilitate the integration of mass chromatograms for formerly N-linked glycopeptides. We demonstrated in our studies that integration of glycopeptide mass chromatograms is highly reproducible and very useful for relative quantitation of abundance across multiple samples. Statistical analyses were applied to identify glycopeptides that discriminated lung cancers from controls and sandwich-based ELISAs were used to confirm the differential expression levels of serum proteins harboring selective glycopeptide candidates.

2. Materials and Methods

2.1. Serum samples and pooling strategies

Peripheral blood samples were obtained from NSCLC patients and control subjects recruited as part of University of Pittsburgh Cancer Institute (UPCI) Lung Nodule/Lung Cancer Proteomics/Genomics Research Registry, together with the Pittsburgh Lung Screening Study (PLuSS), supported by the UPCI Lung Cancer SPORE. A total of 54 newly diagnosed NSCLC patients (31 adenocarcinoma and 23 squamous cell carcinoma), 54 clinical controls with a CT detected nodule but only with non-malignant lung disease as confirmed by biopsy, and 106 healthy PLuSS controls were selected for current study. The clinical and demographic characteristics of NSCLC cases, clinical and PLuSS controls are summarized in Table 1. The significances for the differences in age, gender and smoking history between cases and controls were determined using Fisher's exact test (age), or the Chi-square test (gender and smoking history). None of these demographic characteristics were significantly different between cases and controls, with p values of 0.058, 0.208, and 0.782 for age, gender, and smoking history, respectively. The University of Pittsburgh Institutional Review Board (IRB) approved all aspects of the study including the opportunity for using the information and biospecimens collected from these subjects in new research studies within and outside of the Lung Cancer SPORE. Blood samples from consented Registry and PLuSS subjects were collected, processed, aliquoted, and stored using the same rigorously validated Lung Cancer SPORE protocol based on recommendations from the NIH and the NCI Early Detection Research Network (EDRN). Aliquoted serum samples were frozen at -80°C within one hour of collection and processing and were not thawed prior to the sample pooling procedures; the pooled samples were then refrozen at -80°C until use.

Given the intensive preparative and analytical workflow involved in the LC-MS/MS analysis of serum, a pooled lung cancer case/control serum sample pooling design was required. The overall goal of the serum sample pooling strategy was to construct case and matched control pools as homogenous as possible with respect to the most important clinical and demographic variables describing these samples. With this goal in mind, the individual serum samples as described above were pooled into 15 case pools (P01 – P09, adenocarcinoma; P11 – P16, squamous cell carcinoma) and 16 control pools (P17 – P24, clinical controls; P25 – P32, healthy PLuSS controls). The case pools were stratified by the major NSCLC histological subtypes, adenocarcinoma and squamous cell carcinoma; early versus late stage; gender, and smoking status (active versus ex-smoker, with never smoker cases excluded). The 16 control pools comprised 8 clinical control pools and 8 healthy PLuSS control pools matched to the cases (N=4) and to the clinical controls (N=4). Case/control matching criteria included age ($\pm 5\text{Y}$), gender, smoking status (active vs. ex-smoker), and smoking history ($<20\text{ PY}$, $20\text{--}39\text{ PY}$, $>39\text{ PY}$). PLuSS controls were also matched to cases based on calendar year of diagnosis and blood sample collection within ± 1 year.

2.2. Immunoaffinity removal of high abundant serum proteins

Agilent Human 14 Multiple Affinity Removal spin cartridges (MARS14, Agilent Technologies, Palo Alto, CA) were used to deplete the top 14 most abundant serum proteins (albumin, IgG, antitrypsin, IgA, transferrin, haptoglobin, fibrinogen, alpha-2-macroglobulin, alpha-1-acid glycoprotein, IgM, apolipoprotein AI, apolipoprotein AII, complement C3, and transthyretin). The flow-through fractions which contain the lower abundant serum proteins were desalted through buffer exchange with 50 mM NH_4HCO_3 using Amicon Ultra 5000 MWCO devices (Millipore, Bedford, MA). The micro BCATM protein assay kit (Pierce, Rockford, IL) was used to determine total protein concentrations.

2.3. Enrichment of glycopeptides

To isolate glycopeptides, we first used hydrazide chemistry based capture to enrich glycoproteins. MARS14-depleted serum proteins (100 μ g) were oxidized with 20 mM sodium periodate in the dark for 1.5 h. Sodium sulfite at 30 mM final concentration was added to inactivate excess sodium periodate. After 10 min incubation, 100 μ l 50% hydrazine resin slurry pre-washed 3 times with 100% methanol and 1 ml 1 $\times$ coupling buffer (Bio-Rad, Hercules, CA) were introduced into the solution. The solution was rotated overnight at 4°C to couple oxidized glycoproteins to the hydrazide resin. Non-glycoproteins were removed by washing the hydrazine resin 3 times with 600 μ l 8 M urea in 0.4 M NH_4HCO_3 .

After the coupling, resin *in situ* trypsin digestion was applied to remove non-glycosylated peptides. Hydrazine resin captured glycoproteins were reduced by incubating at 37°C for 1 h in 8 M urea/0.4 M NH_4HCO_3 with 10 mM DTT. After cooling to ambient temperature, 30 mM iodoacetamide was added to alkylate cysteinyl residues. After 90 min incubation in the dark, the hydrazide resin was washed twice with 50 mM NH_4HCO_3 and then 4 μ g trypsin gold (Promega, Madison, WI) was added for overnight digestion at 37°C. Non-glycosylated peptides were removed by sequential washes with 1.5 M NaCl, 100% MeOH, and 50 mM NH_4HCO_3 (3 times each).

To release the N-linked glycopeptides still covalently coupled to the hydrazide resin, 2 μ l PNGase F (New England Biolabs, Beverly, MA) was added to the hydrazide resin followed by 3 h incubation at 37°C. The supernatant was combined with a 50 mM NH_4HCO_3 wash and then desalted with a PepClean™ C-18 Spin Column (Pierce, Rockford, IL). The resulting peptide solutions contained the deglycosylated formerly N-linked glycopeptides.

The whole glycopeptide enrichment and isolation process was performed in duplicate for each pooled serum sample. The final desalted formerly N-linked glycopeptides were combined for each pooled sample and the peptide concentrations were determined using the micro BCA™ protein assay kit prior to being analyzed by LC-MS/MS.

2.4. Analysis of peptides by mass spectrometry

The deglycosylated peptides were resuspended in 0.1% TFA and were analyzed in duplicate (0.8 μ g for each injection) by LC-MS/MS using an LTQ/Orbitrap™ (ThermoFisher Scientific Inc., San Jose, CA) hybrid mass spectrometer. Peptide separations were conducted online with MS by nanoflow liquid chromatography (Dionex Ultimate 3000 Dionex Corporation, Sunnyvale, CA). Separation of the peptide digests was performed using a 75 μ m inner diameter $\times$ 360 outer diameter $\times$ 10 cm-long fused silica capillary column (Polymicro Technologies, Phoenix, AZ) packed in house with 5 μ m, 300Å pore size Jupiter C-18 stationary phase (Phenomenex, Torrance, CA). Samples were injected onto a trap column and washed with 98% mobile phase A (0.1% formic acid in water) for 3 min. Peptides were eluted by development of a linear gradient of 2% mobile phase B (0.1% formic acid in acetonitrile) to 42% mobile phase B in 80 min, then to 95% B in an additional 7 min, all at a constant flow rate of 200 nL/min. Column washing was performed at 95% B for 18 min, after which the column was re-equilibrated in mobile phase A prior to subsequent injections.

The LTQ/Orbitrap™ instrument was operated in a data-dependent MS/MS mode in which each high resolution broadband MS scan ($R = 60,000$ at mass to charge (m/z) 400, precursor ion selection scan range of m/z 350–1800) was followed by 7 MS/MS scans in the linear ion trap where the 7 most abundant peptide molecular ions dynamically determined from the MS scan were selected for tandem MS using a relative collision-induced dissociation (CID) energy of 30%. Dynamic exclusion was enabled to minimize redundant selection of peptides previously selected for CID.

2.5. MS/MS data analysis

Tandem mass spectra were searched against the UniProt human proteome database (released on March, 2010; containing 70,659 protein entries) from the European Bioinformatics Institute (<http://www.ebi.ac.uk/integr8>) using SEQUEST (ThermoFisher Scientific Inc.) with the following modifications: variable modification of cysteine (carboxyamidomethylation, +57.0214 Da), variable modification of methionine (oxidation, +15.9949 Da), and variable modification of asparagine resulting from PNGase F deglycosylation (+0.984 Da due to conversion to aspartic acid). The mass tolerance for the precursor mass was ± 20 parts per million (ppm), and the mass tolerance for the fragment ions of 1 Da. Peptides were considered legitimately identified if they achieved specific charge state and proteolytic cleavage-dependent cross-correlation (Xcorr) scores of 1.9 for $[M+H]^1+$, 2.2 for $[M+2H]^2+$, and 3.5 for $[M+3H]^3+$, and a minimum delta correlation score (ΔC_n) of 0.08. A false peptide discovery rate of approximately 0.3% was determined by searching the primary tandem MS data using the same criteria against a decoy database wherein the protein sequences are reversed.

2.6. Integration of mass chromatograms for identified peptides

An in-house MATLAB[®] based tool was used to integrate mass chromatograms for peptides identified more than once combining all 62 LC-MS/MS analyses. The extraction was carried out with the following steps: (1) convert the LC-MS/MS raw files into mzXML files using ReAdW version 4.2.0;²³ (2) align the retention times of different LC-MS/MS analyses to a common reference LC-MS/MS analysis using a combination of locally weighted scatterplot smoothing (LOWESS) regression²⁴ and linear regression: extract the retention times for MS/MS scans leading to the identification of commonly identified peptides in both the reference LC-MS/MS run and the LC-MS/MS run to be aligned; for peptides identified more than once in a particular run, use the retention time of the MS/MS scan with the highest total ion current (TIC); fit a LOWESS regression curve and a linear regression curve using the extracted retention times of common peptides; for all MS scans whose retention times fall within the retention time range of common peptides, adjust the retention times by interpolating with the fitted LOWESS regression curve; otherwise adjust the retention time by extrapolating with the fitted linear regression curve; (3) determine the range of retention times: for each peptide, extract the retention time of the MS/MS scan leading to the identification of that peptide and with the highest MS/MS signal (TIC for MS/MS scan) based on the adjusted retention time scales for each LC-MS/MS; determine the mean of retention times across all LC-MS/MS runs where the peptide was identified (meanRT); the retention time range was set as $[\text{meanRT}-0.7, \text{meanRT}+0.7]$ (min) which was determined based on manual examination of selected resultant mass chromatograms; (4) set the m/z range for the molecular ions of each peptide: determine the most commonly selected charged stage of a given peptide for CID across all of the LC-MS/MS analyses; the range of peptide molecular ion m/z was set as ± 20 ppm for A peak and A+1 peak with the determined charge stage; (5) integrate peptide mass chromatograms: for each peptide, select all full MS scans with retention times falling within the pre-determined retention time range for that peptide; for each full MS scan, sum the intensity values for peaks within the range of pre-determined precursor m/z values.

2.7. ELISAs

Sandwich based ELISA kits for alpha-1-antichymotrypsin (ACT), insulin-like growth factor-binding protein 3 (IGFBP3), and prostaglandin D synthase (lipocalin-type) (L-PGDS) were purchased from Immunology Consultants Laboratory, Inc. (Catalog E-80CYT), R&D Systems (Catalog DGB300), and Cayman Chemical Company (Catalog No. 10007684), respectively. The assays were performed according to protocols provided by the manufacturers. Both standard samples containing recombinant ACT/IGFBP3/L-PGDS

proteins, as well as the selected pooled serum samples, were assayed in duplicate to reduce variation. For the ACT ELISA, the selected cancer pools were P01, P03, P05, P06, P07, P08, P09, P11, P14, P15, and the selected control pools were P17, P19, P22, P23, P24, P25, P27, P28, P30, and P32. For the IGFBP3 ELISA, the selected cancer pools were P01, P03, P04, P05, P06, P08, P09, P11, P12, P15, and the selected control pools were P17, P19, P21, P22, P24, P25, P26, P29, P30, and P31. For the L-PGDS ELISA, the selected cancer pools were P01, P03, P04, P05, P07, P08, P11, P14, P15, P16, and the selected control pools were P17, P18, P22, P24, P25, P26, P28, P29, P31, and P32. The selection of the sample pools was based on the availability of the remaining pooled serum samples and the expected difference between NSCLC pools and control pools based on the MS data.

2.8. Statistical analyses

Student's *t* tests were performed to determine the significance of glycopeptide mass chromatogram differences between the case and control pools. The estimation of false discovery rate was based on the linear step-up procedure originally introduced by Benjamini and Hochberg in 1995.²⁵ Unsupervised hierarchical clustering was used to evaluate the discriminative power of mass chromatogram areas of the significant glycopeptides. All statistical analyses were conducted with MATLAB® (functions *ttest2*, *mafdr*, *clustergram* for Student's *t* test, false discovery rate, and unsupervised hierarchical clustering, respectively).

3. Results

3.1. Enrichment of serum glycopeptides using hydrazide resin

As described in Materials and Methods Section 2.1, serum samples were obtained from 54 lung cancer patients (31 adenocarcinoma and 23 squamous cell carcinoma), 54 clinical controls, and 106 healthy controls. About half of the NSCLC patients had early stage disease (TMN staging of IA/IB). Table 1 summarizes the demographic and clinical characteristics of the cases and controls, including age, gender, smoking history, and tumor staging for cases. These serum samples were pooled into 9 adenocarcinoma pools, 6 squamous cell carcinoma pools, 8 clinical control pools, and 8 healthy control pools, according to the gender, smoking status, and tumor staging of the lung cancer patients. The detailed pooling strategy is described in Materials and Methods Section 2.1. The top 14 most abundant proteins in these pooled sera were depleted using Agilent Human 14 Multiple Affinity Removal Spin Cartridges (MARS14). Serum glycoproteins were enriched using hydrazide resin capture (Figure 1A). Resin *in situ* trypsin digestion was applied to the captured glycoproteins to release non-glycosylated peptides and peptide-N-glycosidase F (PNGase F) was used to release the formerly N-linked glycopeptides. The average yields of the enriched glycopeptides from 200 µg of post MARS14 depleted serum proteins were 4.4, 4.6, 4.4, and 4.0 µg for the adenocarcinoma, squamous cell carcinoma, clinical control, and healthy control pools, respectively. There was no significant difference in the amount of enriched glycopeptides obtained among these 4 groups.

3.2. Analysis of enriched glycopeptides by LC-MS/MS

The enriched formerly N-linked glycopeptides for each pooled sample (a total of 31 sample pools) were analyzed in duplicate by nano-LC coupled to ESI-MS/MS (LTQ/Orbitrap™), for a total of 62 LC-MS/MS analyses resulting in the identification of 334 different peptides with the known consensus N-linked glycosylation motif (NXS/T with deamidated N, where X is not proline) (listed in Supplemental Table S1) at an estimated false discovery rate of 0.3%. A total of 285 glycopeptides belonging to 140 proteins were identified more than once. Some of these peptides have additional deamidated asparagine residues other than the asparagine residue within the N-linked glycosylation motif, probably due to the

nonenzymatic spontaneous deamidation of asparagine *in vivo* or *in vitro*.²⁶ A total of 266 different glycosylation sites were identified from these 334 glycopeptides.

3.3. Batched integration of mass chromatograms for quantification of identified peptides

To obtain the relative abundance of the glycopeptides across the pooled case/control samples, an in-house MATLAB® based tool was utilized to batch-extract the mass chromatograms for peptides identified more than once in the 62 LC-MS/MS analyses. Prior to the extraction, peptide retention times were aligned to a common reference LC-MS/MS run using a combination of LOWESS regression and linear regression as described in Materials and Methods Section 2.6. An example of the regression curves used for retention time adjustment is shown in Figure 2A. To determine the effectiveness of the retention time alignment algorithm, the variance in retention times prior to and after alignment were determined for all peptides identified in more than 5 different LC-MS/MS analyses. As shown in Figure 2B, the retention time alignment algorithm effectively reduced the mean identification time range of 2.28 minutes to 0.45 minutes with retention time alignment.

After retention time alignment, peptide ion intensities were extracted from full MS scans according to the corresponding retention time ranges and the expected m/z (see Materials and Methods Section 2.6 for details). With proper retention time alignment, mass chromatograms for peptides not identified in a particular LC-MS/MS run can still be extracted based on the estimated retention time ranges from the other LC-MS/MS analyses. Examples of the mass chromatograms for one of the glycopeptides (R.FN*DTEVLQR.L from biotinidase precursor) from multiple LC-MS/MS runs are shown in Figure 2C (based on the original retention time scale) and Figure 2D (based on the retention time scale after alignment). Consistent with previous results, retention time alignment effectively reduced the retention time shifts of different LC-MS/MS runs.

3.4. Assessment of reproducibility of extracted peptide mass chromatogram areas

To determine whether our batch extracted peptide mass chromatograms can be useful for relative quantitation of peptide abundance across the different sample pools, we assessed the reproducibility of the mass chromatograms in the duplicate LC-MS/MS runs. Figure 3A shows the plot of all pairs of $\log_2$ - $\log_2$ plots of mass chromatogram areas for the same peptides in duplicate runs for one of the pooled samples. Good correlation was observed for mass chromatograms between duplicate LC-MS/MS analyses. The mass chromatogram areas for the majority of peptides fell within a 2-fold change. The mean Pearson's correlation between duplicate LC-MS/MS runs for all samples is 0.9867. The reproducibility for each individual peptide across different pooled samples was determined by calculating linear Pearson's correlation coefficient between duplicate LC-MS/MS analyses for each peptide. Figure 3B shows the histogram distribution of the resulting coefficients where it is evident that majority of peptides showed very strong correlation between the duplicate analyses. More than half of the peptides (51%) have correlation coefficients >0.9 . The mean correlation coefficient for all peptides identified more than once across all 62 LC-MS/MS analyses was 0.819. These results indicate that peptide mass chromatogram areas are highly reproducible and are appropriate for reliable estimation of relative peptide abundances across the sample pools in this analysis.

3.5. Differential expression of identified N-linked glycopeptides based on extracted ion chromatogram area

To identify glycopeptides with differential expression between the lung cancer serum pools compared to the control serum pools, we performed 3 pair-wise comparisons using Student's *t* test to compare all case pools versus all control pools, all adenocarcinoma case pools versus all controls, and all squamous cell carcinoma case pools versus all controls. A total of

38 glycopeptides (containing 30 non-redundant sequences) from 22 different proteins yielded a P value <0.01 for at least one of the three statistical comparisons (Table 2). The false discovery rate was estimated to be about 11% using the P values resulting from Student's t test comparison of all case pools vs. all control pools. Significant glycopeptides from the same proteins identified from our studies generally showed a similar abundance trend (up or down in observed abundance) in the lung cancer serum pools, with the exception of apolipoprotein B-100 (apoB100). ApoB100 has 19 potential N-glycosylation sites.²⁷ We identified 7 apoB100 N-glycosylation sites from our LC-MS/MS analysis (Supplemental Table S1). Four of them showed significant differential abundance, with three peptides being increased, and one peptide decreased in abundance in the case pools compared to the control pools. These results may indicate the presence of differential abundance for different N-glycosylation sites of apoB100.

Several of these 22 proteins containing lung cancer selective glycopeptides have been previously associated with lung cancer, such as leucine-rich alpha-2-glycoprotein, insulin-like growth factor binding protein 3, and kininogen-1.^{20, 28} Molecular and functional analysis using Ingenuity Pathways Analysis (IPA) tools²⁹ indicated significant enrichment of proteins involved in a variety of functions and diseases including cell-to-cell signaling and interaction, cancer, hepatic system disease, molecular transport, and cell morphology (Supplemental Table S2). Consistent with the important role of inflammation in tumor development,³⁰ we found 4 positive acute phase response proteins, including isoform gamma-B of fibrinogen gamma chain, plasma protease C1 inhibitor, alpha-1-antichymotrypsin and hemopexin, being up-regulated in NSCLC pools, and one negative acute phase response protein, serotransferrin, having a diminished abundance in the NSCLC pools.

To determine the discriminatory power of these selected glycopeptides, we carried out unsupervised hierarchical clustering with these 38 glycopeptide mass chromatogram areas (Figure 3). The selected peptide abundance measures are able to separate the case pools from the majority of the control pools, the only two exceptions being the clinical control pools P18 and P20; however they are not so robust in discriminating the adenocarcinoma from the squamous cell carcinoma pools or the clinical controls from the healthy PLuSS control pools.

3.5. Verification

To verify the predicted differential serum glycoprotein abundance of candidate lung cancer biomarkers based on LC-MS/MS data, we analyzed the protein levels in selected case and control serum pools of alpha-1-antichymotrypsin (ACT), insulin-like growth factor-binding protein 3 (IGFBP3), and lipocalin-type prostaglandin D synthase (L-PGDS) using commercially available sandwich based ELISAs. The results of these ELISAs were then compared to the corresponding glycopeptide mass chromatogram areas (Figure 5). For all 3 candidates, there is a strong correlation between serum protein concentrations determined by ELISA and the derived glycopeptide mass chromatogram areas. Data shown for ACT and L-PGDS are glycopeptides K.YTGN*ASALFILPDQDK.M and K.SVVAPATDGGLN*LTSTFLR.K, respectively. Both have higher average mass chromatogram areas than the other observed glycopeptides from the same respective protein. The linear Pearson's correlation coefficients between these two endpoints were 0.918 ($P = 1.1533e-008$) for ACT, 0.576 ($P = 0.008$) for IGFBP3, and 0.6322 ($P = 0.00028$) for L-PGDS. These ELISA results confirm the legitimacy of using glycopeptide mass chromatogram area measures for the relative quantitation of peptide abundance to identify differentially abundant serum proteins as candidate lung cancer biomarkers.

4. Discussion

Despite advances in clinical diagnostics and therapeutic modalities, lung cancer patients continue to have a very dismal survival rate and will greatly benefit from tools capable of detecting lung cancer at early stages. Aiming for discovery of molecular lung cancer diagnostic biomarkers, we applied a novel targeted serum proteomic approach utilizing coupling immunoaffinity depletion of high abundant serum proteins, hydrazide chemistry based enrichment of glycoproteins, resin *in situ* trypsin digestion and PNGase F cleavage of N-linked glycans, followed by high mass resolution LC-MS/MS analysis to a large set of pooled NSCLC case and control sera. Our studies identified a total of 334 different glycopeptides containing 260 non-redundant peptide sequences from 166 different glycoproteins. A number of these identified proteins are present in very low abundance in serum/plasma, such as angiotensinogen, cathepsin D, transforming growth factor beta 1, and hepatocyte growth factor-like protein³¹ demonstrating that this discovery workflow efficiently samples the lower abundance serum glycoproteome.

We used mass chromatogram area measures for relative quantitation of glycopeptides. To reduce the effect of co-eluting peptide ions with irresolvable m/z values, we analyzed our enriched formerly N-linked glycopeptides samples with high mass resolution mass spectrometry (LTQ/OrbitrapTM). In addition, we applied retention time alignment to correct the retention time shift across different LC-MS/MS runs. We demonstrated that our alignment algorithm worked efficiently to minimize the effect of retention time shifts. The resulting aligned mass chromatograms showed very strong reproducibility between duplicative LC-MS/MS analyses. The majority of peptides have within-sample variation less than a 2-fold change. Peptides with smaller mass chromatogram areas tend to have larger variation among replicate LC-MS/MS runs. The mass chromatogram measures for these peptides will likely have larger variance among different samples and it will be more difficult to detect the true significance of these peptides if their differences between case and control pools are subtle.

Glycopeptide mass chromatograms for 285 glycopeptides were extracted using our in-house MATLAB[®] based tool. A total of 22 serum proteins containing NSCLC selective glycopeptides were identified based on P value < 0.01 for at least one of the three Student's t test comparisons (all case pools vs. control pools, adenocarcinoma case pools vs. control pools, and squamous cell carcinoma pools vs. control pools). It needs to be pointed out that a number of these NSCLC selective glycopeptides have a relatively small fold change between cases and controls (less than 2-fold), suggesting that their differences between cases and controls are subtle and will require large sample sizes of individual case and control sera for future validation. Some of these NSCLC selective proteins have been implicated in earlier published studies as potential lung cancer serum biomarkers. For example, several of these proteins such as plasma protease C1 inhibitor, leucine-rich alpha-2-glycoprotein, kininogen-1 precursor, alpha-1-antichymotrypsin, and serotransferrin were reported previously in studies for biomarker discovery for human lung adenocarcinoma.^{20, 32} In addition, we observed the same polarity of the changes between cases and controls between our data and the published studies for these potential lung cancer serum biomarkers. To verify the present use of glycopeptide mass chromatograms for relative quantitation, we measured serum levels of 3 selected glycoproteins, ACT, IGFBP3, and L-PGDS, with commercially available ELISA kits. All 3 glycoproteins contain glycopeptides that are identified in our LC-MS/MS analysis and displayed significant differences in NSCLC pooled serum versus control pooled serum based on glycopeptide mass chromatogram-derived abundance. Although whole protein based ELISAs are not directly comparable to peptide abundance, nonetheless, we observed very strong correlation between the measured

serum expression levels of all 3 candidate proteins and glycopeptide mass chromatogram areas, supporting the legitimacy of using mass chromatogram areas for relative quantitation.

Thoracic CT scanning, although sensitive in detecting lung cancer, has a reported range of 20–50% in detection of benign pulmonary nodules.^{33–34} While biopsy of suspicious findings on a chest CT can be performed either via transthoracic needle biopsy or via bronchoscopy, these invasive procedures are associated with significant risk due to complications, especially for patients at high risk for lung cancer, prompting the need to avoid unnecessary invasive thoracic procedures. The robustness of using significant peptides identified from our study to discriminate case pools and control pools was tested using hierarchical clustering, a statistical method for grouping subjects based on similarity of measured characteristics. The resulting analysis showed almost complete separation between NSCLC case pools and control pools with the exception of only 2 clinical control pools being clustered together with case pools. Noteworthy is the fact that many of the clinical controls, although cancer free at the time of serum sample collection, had non-malignant lung disease as demonstrated by the presence of CT detected pulmonary nodules. Two of the clinical pools were clustered together with the NSCLC case pools, probably reflecting the contributions of inflammatory proteins included in our list of significant features. Nonetheless we are able to separate the majority of clinical control pools (6 of the 8 pools) from NSCLC case pools, suggesting that our selected features may potentially be useful in increasing the specificity of thoracic CT scans for lung cancer screening, thus reducing the need of invasive thoracic procedures for patients with CT detected indeterminate nodules.

In summary, a substantial number of potential NSCLC associated glycopeptides were identified from our analysis of the enriched glycopeptides using LC-MS/MS. Further validation studies, involving both independent case/control sample sets together with complementary high-throughput quantitative assays, e.g. whole glycoprotein ELISAs or peptide based selected reaction monitoring (SRM) MS^{35–37} applied in individual patient/subject sera, are required to confirm and validate these initial discovery results. Since our analysis allowed us to also pinpoint the actual N-linked glycosylation sites that are subjected to differential regulation, targeted peptide based SRM assays are well suited for the validation of these nominated NSCLC selective glycopeptides.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by Hirtzel Foundation Postdoctoral Fellowship to XZ, P50 CA90440 NCI SPORE in Lung Cancer to JMS, and NCI Early Detection Research Network Biomarker Discovery Laboratory, U01 CA084968 to WLB.

References

1. Okada M, Nishio W, Sakamoto T, Uchino K, Yuki T, Nakagawa A, Tsubota N. Effect of tumor size on prognosis in patients with non-small cell lung cancer: the role of segmentectomy as a type of lesser resection. *J. Thorac. Cardiovasc. Surg.* 2005; 129(1):87–93. [PubMed: 15632829]
2. Smith RA, Cokkinides V, Brooks D, Saslow D, Brawley OW. Cancer screening in the United States, 2010: a review of current American Cancer Society guidelines and issues in cancer screening. *CA. Cancer J. Clin.* 2010; 60(2):99–119. [PubMed: 20228384]
3. Kaneko M, Eguchi K, Ohmatsu H, Kakinuma R, Naruke T, Suemasu K, Moriyama N. Peripheral lung cancer: screening and detection with low-dose spiral CT versus radiography. *Radiology.* 1996; 201(3):798–802. [PubMed: 8939234]

4. Welch HG, Woloshin S, Schwartz LM, Gordis L, Gøtzsche PC, Harris R, Kramer BS, Ransohoff DF. Overstating the evidence for lung cancer screening: the International Early Lung Cancer Action Program (I-ELCAP) study. *Arch. Intern Med.* 2007; 167(21):2289–2295. [PubMed: 18039986]
5. Wilson DO, Weissfeld JL, Fuhrman CR, Fisher SN, Balogh P, Landreneau RJ, Luketich JD, Siegfried JM. The Pittsburgh Lung Screening Study (PLuSS): outcomes within 3 years of a first computed tomography scan. *Am. J. Respir. Crit. Care Med.* 2008; 178(9):956–961. [PubMed: 18635890]
6. Stamey TA, Yang N, Hay AR, McNeal JE, Freiha FS, Redwine E. Prostate-specific antigen as a serum marker for adenocarcinoma of the prostate. *N. Engl. J. Med.* 1987; 317(15):909–916. [PubMed: 2442609]
7. Fujino N, Haga Y, Sakamoto K, Egami H, Kimura M, et al. Clinical evaluation of an immunoradiometric assay for CA15-3 antigen associated with human mammary carcinomas: comparison with carcinoembryonic antigen. *Jpn. J. Clin. Oncol.* 1986; 16(4):335–346. [PubMed: 3467106]
8. Gadducci A, Cosio S, Carpi A, Nicolini A, Genazzani AR. Serum tumor markers in the management of ovarian, endometrial and cervical cancer. *Biomed. Pharmacother.* 2004; 58(1):24–38. [PubMed: 14739059]
9. Tan E, Gouvas N, Nicholls RJ, Ziprin P, Xynos E, Tekkis PP. Diagnostic precision of carcinoembryonic antigen in the detection of recurrence of colorectal cancer. *Surg. Oncol.* 2009; 18(1):15–24. [PubMed: 18619834]
10. Anderson NL, Anderson NG. The human plasma proteome: history, character, and diagnostic prospects. *Mol. Cell Proteomics.* 2002; 1(11):845–867. [PubMed: 12488461]
11. Zhang H, Yan W, Aebersold R. Chemical probes and tandem mass spectrometry: a strategy for the quantitative analysis of proteomes and subproteomes. *Curr. Opin. Chem. Biol.* 2004; 8(1):66–75. [PubMed: 15036159]
12. Apweiler R, Hermjakob H, Sharon N. On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochim. Biophys. Acta.* 1999; 1473(1):4–8. [PubMed: 10580125]
13. Tong L, Baskaran G, Jones MB, Rhee JK, Yarema KJ. Glycosylation changes as markers for the diagnosis and treatment of human disease. *Biotechnol. Genet. Eng. Rev.* 2003; 20:199–244. [PubMed: 14997853]
14. 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]
15. Tian Y, Zhou Y, Elliott S, Aebersold R, Zhang H. Solid-phase extraction of N-linked glycopeptides. *Nat. Protoc.* 2007; 2(2):334–339. [PubMed: 17406594]
16. Sun B, Ranish JA, Utleg AG, White JT, Yan X, Lin B, Hood L. Shotgun glycopeptide capture approach coupled with mass spectrometry for comprehensive glycoproteomics. *Mol. Cell Proteomics.* 2007; 6(1):141–149. [PubMed: 17074749]
17. Kaji H, Saito H, Yamauchi Y, Shinkawa T, Taoka M, Hirabayashi J, Kasai K, Takahashi N, Isobe T. Lectin affinity capture, isotope-coded tagging and mass spectrometry to identify N-linked glycoproteins. *Nat. Biotechnol.* 2003; 21(6):667–672. [PubMed: 12754521]
18. Ueda K, Katagiri T, Shimada T, Irie S, Sato TA, Nakamura Y, Daigo Y. Comparative profiling of serum glycoproteome by sequential purification of glycoproteins and 2-nitrobenzenesulfonyl (NBS) stable isotope labeling: a new approach for the novel biomarker discovery for cancer. *J. Proteome Res.* 2007; 6(9):3475–3483. [PubMed: 17705522]
19. Hongsachart P, Huang-Liu R, Sinchaikul S, Pan FM, Phutrakul S, Chuang YM, Yu CJ, Chen ST. Glycoproteomic analysis of WGA-bound glycoprotein biomarkers in sera from patients with lung adenocarcinoma. *Electrophoresis.* 2009; 30(7):1206–1220. [PubMed: 19294700]
20. Heo SH, Lee SJ, Ryoo HM, Park JY, Cho JY. Identification of putative serum glycoprotein biomarkers for human lung adenocarcinoma by multilectin affinity chromatography and LC-MS/MS. *Proteomics.* 2007; 7(23):4292–4302. [PubMed: 17963278]

21. Old WM, Meyer-Arendt K, Aveline-Wolf L, Pierce KG, Mendoza A, Sevinisky 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]
22. Wong JW, Cagney G. An overview of label-free quantitation methods in proteomics by mass spectrometry. *Methods Mol. Biol.* 2010; 604:273–283. [PubMed: 20013377]
23. Thermo Xcalibur raw file to mzXML converter, ReAdW version 4.2.0 was downloaded. from <http://tools.proteomecenter.org/wiki/index.php?title=Software:ReAdW>.
24. Cleveland WS. Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* 1979; 74(368):829–836.
25. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc.* 1995; 57:289–300.
26. Lindner H, Helliger W. Age-dependent deamidation of asparagine residues in proteins. *Exp. Gerontol.* 2001; 36(9):1551–1563. [PubMed: 11525877]
27. Harazono A, Kawasaki N, Kawanishi T, Hayakawa T. Site-specific glycosylation analysis of human apolipoprotein B100 using LC/ESI MS/MS. *Glycobiology*. 2005; 15(5):447–462. [PubMed: 15616123]
28. Chen B, Liu S, Xu W, Wang X, Zhao W, Wu J. IGF-I and IGFBP-3 and the risk of lung cancer: a meta-analysis based on nested case-control studies. *J. Exp. Clin. Cancer Res.* 2009; 28:89–94. [PubMed: 19549343]
29. The networks and functional analyses were generated through the use of Ingenuity Pathways Analysis (Ingenuity® Systems, www.ingenuity.com)."
30. Grivennikov SI, Greten FR, Karin M. Immunity, inflammation, and cancer. *Cell*. 2010; 140(6): 883–899. [PubMed: 20303878]
31. 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]
32. Okano T, Kondo T, Kakisaka T, Fujii K, Yamada M, Kato H, Nishimura T, Gemma A, Kudoh S, Hirohashi S. Plasma proteomics of lung cancer by a linkage of multi-dimensional liquid chromatography and two-dimensional difference gel electrophoresis. *Proteomics*. 2006; 6(13): 3938–3948. [PubMed: 16767791]
33. Welch HG, Woloshin S, Schwartz LM, Gordis L, Gøtzsche PC, Harris R, Kramer BS, Ransohoff DF. Overstating the evidence for lung cancer screening: the International Early Lung Cancer Action Program (I-ELCAP) study. *Arch. Intern. Med.* 2007; 167(21):2289–2295. [PubMed: 18039986]
34. Wilson DO, Weissfeld JL, Fuhrman CR, Fisher SN, Balogh P, Landreneau RJ, Luketich JD, Siegfried JM. The Pittsburgh Lung Screening Study (PLuSS): outcomes within 3 years of a first computed tomography scan. *Am. J. Respir. Crit. Care Med.* 2008; 178(9):956–961. [PubMed: 18635890]
35. Anderson L, Hunter CL. Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. *Mol. Cell Proteomics*. 2006; 5(4):573–588. [PubMed: 16332733]
36. Kuzyk MA, Smith D, Yang J, Cross TJ, Jackson AM, Hardie DB, Anderson NL, Borchers CH. Multiple reaction monitoring-based, multiplexed, absolute quantitation of 45 proteins in human plasma. *Mol. Cell Proteomics*. 2009; 8(8):1860–1877. [PubMed: 19411661]
37. Abbatiello SE, Pan YX, Zhou M, Wayne AS, Veenstra TD, Hunger SP, Kilberg MS, Eyler JR, Richards NG, Conrads TP. Mass spectrometric quantification of asparagine synthetase in circulating leukemia cells from acute lymphoblastic leukemia patients. *J. Proteomics*. 2008; 71(1): 61–70. [PubMed: 18541474]

Figure 1.

(A) Schematic illustration of pooled serum sample processing. Immunoaffinity depletion was first applied to the crude serum pools to remove high abundance proteins. Glycoproteins were then captured with hydrazide resin, and resin *in situ* trypsin digestion was used to release non-glycosylated peptides. PNGase F was applied to release captured glycopeptides which are then analyzed by high mass resolution LC-MS/MS. (B) Schematic illustration for data analysis. LC-MS/MS results were submitted to the SEQUEST cluster for peptide identification. To extract the peptide ion mass chromatograms, the raw files generated by LC-MS/MS were converted into mzXML files using ReAdw tool and then an in-house MATLAB® based tool was used to extract mass chromatogram for every peptide identified more than once combining all LC-MS/MS runs across the case/control sample pools. Statistical tests were then used to identify significant features.

Figure 2.

Batch-extraction of peptide mass chromatograms. (A) Scatterplot of the retention times (RT) of common peptides and the fitted calibration curve used to adjust retention time shifts. The red line represents the LOWESS regression curve and the blue lines represent the linear regression curve. (B) Scatterplot of peptide retention time ranges before and after retention time adjustment. The center red line corresponds to no change before and after calibration. Data points below the red line are peptides with a retention time range reduced by the alignment and vice-versa. Retention time alignment significantly reduced the retention time differences across the multiple LC-MS/MS runs for the majority of peptides. (C and D) Extracted mass chromatograms for a formerly N-linked biotinidase glycopeptide (R.FN*DTEVLQR.L) from 7 different LC-MS/MS runs using retention time scales before (C) and after adjustment (D). Each color represents one LC-MS/MS run.

Figure 3.

(A) Log₂-log₂ scatter plot of peptide mass chromatogram areas between duplicate LC-MS/MS runs of one of the sample pools (P01). The center red line indicates a ratio of 1 (meaning no change between duplicate runs), and the 2 blue lines delineate 2-fold change. The vast majority of peptides have a mass chromatogram area ratio within a 2-fold change in duplicative LC-MS/MS runs. (B) Histogram distribution of the linear Pearson's correlation coefficients between duplicative LC-MS/MS runs across all sample pools for each peptide. The majority of peptides showed very strong correlation between duplicative LC-MS/MS runs, with more than 50% of peptides with Pearson's correlation coefficients greater than 0.9.

Figure 4.

Hierarchical clustering based on mass chromatogram areas of the significant glycopeptides ($P < 0.01$). Both sample distance (column distance) and peptide feature distance (row distance) were based on Pearson's correlation and the linkage of both clusters was based on average linkage. The values for mass chromatogram areas were standardized for each peptide. The selected features are able to separate case pools (pools P01–P16) from the majority of controls (pools P17–P32) (with only 2 exceptions, P18 and P20); however they are not so robust in discriminating adenocarcinoma (pools P01–P09; red) from squamous cell carcinoma (pools P11–P16; purple) or clinical controls (pools P17–P24; blue) from healthy PLuSS controls (pools P25–P32; green).

Figure 5. Correlation between glycopeptide mass chromatogram areas (MCA) and serum ACT (A), or IGFBP3 (B), or L-PGDS (C) glycoprotein levels determined by ELISAs. Serum ACT, IGFBP3, and L-PGDS levels for selected case and control pooled sera were determined using commercially available ELISA kits.

Table 1

Demographic and clinical characteristics of the individual subjects contributing to the pooled NSCLC case and control samples

	Cases		Controls	
	N	%	N	%
Age (years)				
<50	0	0.00	2	1.25
50–54	3	5.56	22	13.75
55–59	5	9.26	36	22.50
60–64	10	18.52	25	15.63
65–69	12	22.22	31	19.38
70–74	13	24.07	18	11.25
75+	11	20.37	26	16.25
Gender				
Male	23	42.59	84	52.50
Female	31	57.41	76	47.50
Smoking history				
Previous	35	64.81	107	66.88
Current	19	35.19	53	33.13
Histology of lung cancers				
Adenocarcinoma	31	57.41		
Squamous cell carcinoma	23	42.59		
Stage of lung cancers				
IA/IB	27	50.00		
IIA/IIB	8	14.81		
IIIA/IIIB	14	25.93		
IV	1	1.85		
Unknown	4	7.41		

Table 2

Identified glycopeptides with differential mass chromatogram areas with a minimum P value^d less than 0.01.

Accession	Protein	Identified Peptide	Area ^a	ADC ^b	SCC ^c	P Value ^d
P01011-1	Alpha-1-antichymotrypsin	K.YTGN*ASALFILPDQDK.M	5.84E+08	1.56	1.15	1.96E-04
		R.TLN*QSSDELQLSM@GNAM@FVK.E	1.87E+08	1.80	1.33	1.57E-04
		K.FN*LTETSEAEIHQSFQHLRL.T	4.80E+06	2.52	1.38	4.58E-03
P04217	Alpha-1B-glycoprotein	R.EGDHEFLEVEPEAQEDVEATFP... VHQPCN*YSC#SYR.T	2.66E+07	1.66	1.41	6.93E-03
P04114	Apolipoprotein B-100	K.YDFN*SSM@LYSTAK.G	8.68E+07	-1.23	-1.39	4.14E-03
		R.FN*SSYLQGTN^QITGR.Y	7.66E+07	-1.17	-1.35	3.84E-03
		K.FVEGSHN*STVSLTTK.N	6.72E+06	-1.06	-1.44	2.86E-03
		R.FEVDSPVYN*ATWSASLK.N	3.60E+06	1.55	1.10	1.47E-03
		K.EAFSN*FSSSTLTEVQAISTHGGVGD.K.I	9.88E+05	1.59	1.55	3.35E-03
P07307-1	Asialoglycoprotein receptor 2	K.N^PVGLIGAEN*ATGETDPSHSK.F	1.17E+06	-2.41	-1.54	6.47E-03
P43251	Biotinidase	K.GSLSYLN*VTR.K	1.46E+06	-1.42	-1.12	3.63E-03
P07339	Cathepsin D	R.SIPAC^hVPWSPYLFQPN*DTG^hVSGWG.R.E	1.88E+05	3.15	1.56	5.70E-03
P05156	Complement factor I	K.LDAPTN^LQFVN*ETDSTVLVR.W	3.62E+06	1.64	1.23	5.53E-03
P02751-1	Fibronectin	K.AVLVNN*ITTGER.L	9.28E+05	1.37	1.51	5.72E-03
Q8NBJ4-1	Golgi membrane protein 1	K.ALQPQN*VTSLGCG^TH.-	3.26E+09	1.30	1.11	3.64E-03
P02790	Hemopexin	R.C^hSDGWSFDAITLDDN*GTM@LFFK.G	1.10E+06	2.62	1.53	7.82E-03
P17936	Insulin-like growth factor-binding protein 3	R.YKVDYESQSTDTQN*FSSESK.R	9.03E+05	-1.67	-1.55	4.93E-04
Q14624-1	Inter-alpha-trypsin inhibitor heavy chain H4	K.LPTQN*ITFQTESSVAEQAEFQSPK.Y	1.26E+08	1.27	1.09	3.96E-03
P15151-1	Isoform Alpha of Poliovirus receptor	R.VEDEGN*YTC^hLFVTFPQSGR.S	3.86E+05	1.45	1.22	7.37E-03
P02679-1	Isoform Gamma-B of Fibrinogen gamma chain	K.VDKDQLQSLEDILHQVEN*K.T	2.04E+05	9.42	2.68	9.61E-03
P01042-1	Isoform HMW of Kininogen-1	K.YNSQN*QSNNQFVLYR.I	1.37E+07	-1.31	-1.23	2.60E-04
P29622	Kallistatin	K.FLN*DTM@AVYEAK.L	7.58E+07	-1.38	-1.43	8.05E-04

Accession	Protein	Identified Peptide	Area ^a	ADC ^b	SCC ^c	P Value ^d
P02750	Leucine-rich alpha-2-glycoprotein	K.LPPGLLAN [*] FTLLR.T	5.14E+06	2.71	2.34	1.42E-03
Q13201-1	Multimerin-1	K.FNPGAESVVLSN [*] STLK.F	5.31E+06	1.21	1.31	9.62E-03
P80108-1	Phosphatidylinositol-glycan-specific phospholipase D	K.LGTSLSGGHVLN [@] N [*] GTLK.Q	1.63E+06	-1.30	-1.40	4.59E-03
P05155	Plasma protease C1 inhibitor	K.VGQLQLSHN [*] LSLVILVPQNLIK.H	2.26E+05	3.60	1.12	2.70E-03
P41222	Prostaglandin-H2 D-isomerase	K.SVVAPATDGGGLN [*] LTSTFLR.K	5.94E+06	1.59	1.37	3.71E-03
P02787	Serotransferrin	R.WFSAGLASN [*] SSWLR.E	2.19E+05	1.71	1.18	7.12E-04
		R.QQHLFGSN [*] VTDC [#] SGNFC [#] LFR.S	3.70E+07	-1.44	-1.89	8.11E-03
		K.C [#] GLVPVLAEN [^] YN [*] K.S	2.85E+07	-3.16	-2.23	8.03E-04

For redundant sequences (sharing the same glycosylation sites), only peptides with the highest average mass chromatogram areas are shown.

^a average mass chromatogram area of all samples;
^b average fold change for adenocarcinoma pools;
^c average fold change for squamous cell carcinoma pools;
^d the minimal P value of 3 Student's *t* tests comparing all cases vs. controls, all adenocarcinoma vs. controls, and all squamous cell carcinoma vs. controls. Modifications of amino acid:
* conversion of asparagine to aspartic acid due to PNGase F treatment;
^ spontaneous conversion of asparagine to aspartic acid;
@ oxidation of methionine;
carboxyamidomethylation of cysteine.