Label-Free Glycopeptide Quantification for Biomarker Discovery in Human Sera

ARTICLE *in* JOURNAL OF PROTEOME RESEARCH · JUNE 2014

Impact Factor: 4.25 · DOI: 10.1021/pr500242m · Source: PubMed

7 AUTHORS, INCLUDING:

Ehwang Song
Pacific Northwest National Laboratory
12 PUBLICATIONS **113** CITATIONS
SEE PROFILE

Abhinav Mathur
Cincinnati Children's Hospital Medical Center
8 PUBLICATIONS **32** CITATIONS
SEE PROFILE

Chuan-Yih Yu
Indiana University Bloomington
22 PUBLICATIONS **90** CITATIONS
SEE PROFILE

Yehia Mechref
Texas Tech University
188 PUBLICATIONS **5,572** CITATIONS
SEE PROFILE

Journal of proteome research

Article

pubs.acs.org/jpr

Label-Free Glycopeptide Quantification for Biomarker Discovery in Human Sera

Anoop Mayampurath,*[,†,∥] Ehwang Song,[‡] Abhinav Mathur,[†] Chuan-yih Yu,[†] Zane Hammoud,[§] Yehia Mechref,*[,‡] and Haixu Tang*[,†]

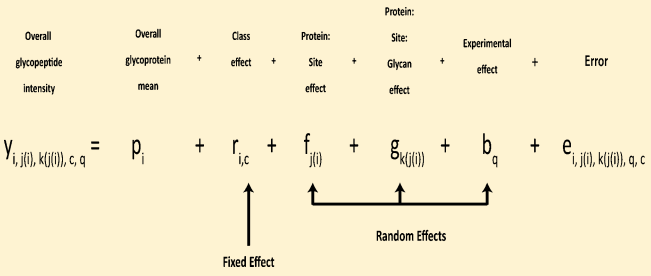
[†]School of Informatics & Computing, Indiana University, 901 East 10th Street, Bloomington, Indiana 47408, United States

[‡]Department of Chemistry & Biochemistry, Texas Tech University, Memorial Circle & Boston, Lubbock, Texas 79409, United States

[§]Cardiothoracic Surgery, Henry Ford Hospital, 2799 West Grand Boulevard, Detroit, Michigan 48202, United States

Supporting Information

ABSTRACT: Glycan moieties of glycoproteins modulate many biological processes in mammals, such as immune response, inflammation, and cell signaling. Numerous studies show that many human diseases are correlated with quantitative alteration of protein glycosylation. In some cases, these changes can occur for certain types of glycans over specific sites in a glycoprotein rather than on the global abundance of the glycoprotein. Conventional analytical techniques that analyze the abundance of glycans cleaved from glycoproteins cannot reveal these subtle effects. Here we present a novel statistical method to quantify the site-specific glycosylation of glycoproteins in complex samples using label-free mass spectrometric techniques. Abundance variations between sites of a glycoprotein as well as different glycoforms, that is, glycopeptides with different glycans attached to the same site, can be detected using these techniques. We applied our method to an esophageal cancer study based on blood serum samples from cancer patients in an attempt to detect potential biomarkers of site-specific N-linked glycosylation. A few glycoproteins, including vitronectin, showed significantly different site-specific glycosylations within cancer/control samples, indicating that our method is ready to be used for the discovery of glycosylated biomarkers.



KEYWORDS: *glycoproteomics, label-free quantification, site-specific glycosylation, biomarker discovery*

INTRODUCTION

As one of the most common post-translational modifications of proteins, glycosylation contributes to many functional attributes of proteins, from stabilizing their tertiary structures to mediating intercellular communication and immune response to pathogen infections.[1,2] Many of these functions depend on the recognition of glycan moieties within membrane or secreted-glycoproteins by various classes of glycan-binding proteins (GBPs). Thus, alteration on these glycan moieties may cause the disruption of essential cellular functions. Numerous studies show that quantitative changes in glycosylation are associated with many human diseases, including hereditary disorders, cardiovascular disease, and cancer.[1,3] In some cases, however, aberrations can occur only at a site-specific level, wherein the variation is observed in a glycan attachment to a specific glycosylation site, rather than in the global abundance of the glycoprotein.[4] Other possibilities include overall change in glycan abundance across all sites or changes that occur at only at specific sites, that is, where one site is glycosylated at a different abundance from another.

In proteomics, label-free approaches have been regularly used to measure changes in protein expression, where protein abundances are derived from peptide abundances, and the ratio of a protein abundance in diseased versus healthy is used to assess if a protein is overexpressed, under-expressed, or stays the same.[5] Quantifying protein expression using statistical methods is an active area of research, where significant gains have been achieved. Hill et al.[6] have developed an ANOVA-based approach to combined normalization and differential expression characterization. The log-2 abundance of a given peptide from a particular protein in a sample is modeled as a linear combination of predictor variables or factors that account for biological and experimental variation. Oberg et al.[7] extended this approach to a stage-wise method for large data sets where the factors are divided into three categories that account for different experimental and biological effects. For general experiments, Karpievitch el al.[8] developed a probability-based quantification approach that also addresses the issue of missing values. Clough et al.[9] detailed both fixed and mixed effect ANOVA modeling of proteomic data sets for group testing and

Special Issue: Proteomics of Human Diseases: Pathogenesis, Diagnosis, Prognosis, and Treatment

Received: March 10, 2014

   dx.doi.org/10.1021/pr500242m | J. Proteome Res. XXXX, XXX, XXX−XXX

subject quantification while demonstrating the improvements in sensitivity in using an ANOVA approach.

Quantification of glycoprotein expression in complex samples is, however, still a nascent field. Conventional analytical platforms using liquid chromatography−mass spectrometry (LC−MS) can be adopted to monitor the changes of global expression levels of many glycoproteins[10−13] (i.e., similar to the quantitative proteomic profiling approaches) or observe changes in different classes of glycans from all glycoproteins in complex samples[14] (i.e., the quantitative glycomic approaches[15−17]). Both approaches involve cleavage of glycans from the attached glycoproteins prior to the analysis in LC−MS. Thus, any site-specific glycosylation characteristics are lost because although these methods can measure glycoprotein expression and alteration, they are incapable of detecting the source of variation, whether the differences are at global glycoprotein levels or at a site-specific level or on particular glycoforms (resulting from the same glycosylation site occupied by different glycans). A quantitative study focusing on profiling intact glycopeptides, that is, peptides along with attached glycans, will preserve these characteristics. Recently, site-specific glycosylation studies have been conducted on immunoglobulin glycoforms using targeted techniques[18] and multispecific proteases.[19] However, large-scale quantitative glycoproteomics for examining site-specific behavior on complex samples are yet to be undertaken. Moreover, current glycoproteomic quantification methods require extensive separation and purification of glycoproteins prior to the MS analysis, which may not achieve necessary high throughput and sensitivity in biomedical studies.

We propose a label-free approach to the quantification of site-specific glycosylations of glycoproteins in a complex sample. We use glycomaps[20] as our data sets, which are essentially a list of glycopeptides (characterized using protein, site and glycan composition) with mass, NET (normalized elution time, defined as a 0-to-1 measure indicative of glycopeptide time of elution that is determined empirically), and abundance values. We devise and apply a novel ANOVA (analysis of variance) statistical model for application to glycomaps. The abundance of a glycopeptide is broken down into a number of fixed and random effects. The fixed terms account for variation at global glycoprotein level, whereas the random effects indicate changes at either the site or glycan level. Through hypothesis testing, we show how detection of site-specific variations (i.e., differential increase or decrease in expression at a particular site and glycan or across all sites with glycans attached) between healthy and disease data sets can be made.

We first showcase efficacy of our novel statistical method by using a glycomap built from glycoproteomic data sets containing four standard glycoproteins with spiked-in fetuin protein at different concentrations. We then apply our model to the discovery of N-linked glycobiomarkers using two human sera glycomaps built from two sets of glycoproteomic samples: one (referred to as GlycoMapSera_Pooled glycomap) built from pooled human sera samples from disease and control subjects and the other (referred to as GlycoMapSera_Individual) built using glycopeptide alignment algorithms using data sets acquired from individual esophageal adenocarcinoma patients (CA, n = 15), patients suffering from Barrett's disease (BD, n = 7), high-grade dysplasia (HGD, n = 12), and disease-free control subjects (DF, n = 15), respectively. We explore these diseases due to their observed linkage; Barrett's disease is often a precursor to the precancerous high-grade dysplasia and the more severe esophageal cancer.[21,22] However, the mechanism of disease progression is not fully understood. The motivation of our work is to hypothesize that glycomic and glycoproteomic signatures might aid in delineating these three separate yet interconnected diseases. Out of our potential candidates, vitronectin, was revealed to display consistently the alteration of site-specific glycosylations in both pooled and individual glycomaps. These results demonstrate that the statistical method described here are robust for assessing quantitative alterations of protein glycosylation at site-specific levels across different classes of samples, thereby setting the stage for glycoproteomic biomarker discovery. Furthermore, our model can be easily modified to study site-specific variation in any post-translational modification such as phosphorylation and even interaction between modifications such as O-GlcNAc/phosphorylation crosstalk.[1]

MATERIALS AND METHODS

Materials

HPLC-grade solvents, including methanol, isopropanol, and water, were procured from Macron Fine Chemicals Avantor Performance Materials (Center Valley, PA). HPLC-grade acetonitrile was purchased from Fisher Scientific (Pittsburgh, PA). Sodium chloride (NaCl) and disodium phosphate (Na_2HPO_4) were acquired from Mallinckrodt Chemicals (Phillipsburg, NJ). Acetic acid and MS-grade formic acid were acquired from Sigma-Aldrich (St. Louis, MO). D,L-Dithiothreitol (DTT) and iodoacetamide (IAA) were also purchased from Sigma-Aldrich. Mass-spectrometry-grade trypsin was obtained from Promega (Madison, WI). Agarose-bound *Sambucus nigra* agglutinin and *Aleuria aurantia* lectin were purchased from Vector Laboratories (Burlingame, CA).

Depletion of the Seven Most Abundant Blood Serum Proteins in Patient Samples

Blood serum samples were obtained from patients diagnosed with esophageal adenocarcinoma (CA, N = 15), high-grade dysplasia (HGD, N = 12), Barrett's disease (BD, N = 7), as well as age and sex-matched disease-free (DF, N = 15) subjects. Agilent Plasma 7 Multiple Affinity Removal Spin Cartridge from Agilent Technologies (Santa Clara, CA) was used to deplete the seven most abundant human serum proteins: albumin, IgG, antitrypsin, IgA, transferin, haptoglobin, and fibrinogen. A 15 μL aliquot of human blood serum was depleted, as stated in the protocol recommended by the manufacturer. The buffer of the depleted sample was exchanged into 50 mM PBS, a phosphate-buffered saline containing 50 mM of Na_2HPO_4 and 150 mM of NaCl, using 5 kDa MWCO spin concentrators from Agilent Technologies. This buffer is needed for efficient lectin enrichment.

Protein Assay

Prior to lectin enrichment, the protein concentration of depleted blood serum (BS) was determined through micro BCA protein assay (Thermo Scientific/Pierce, Rockford, IL). A bovine serum albumin (BSA) standard stock solution of 2.0 mg/mL concentration provided in the micro BCA assay kit was used to prepare a set of diluted BSA standard samples with concentrations of 200, 40, 20, 10, 5, 2.5 and 1 μg/mL. 50 mM PBS was used to prepare the BSA standard samples. The micro BCA working reagent required for the assay was prepared by mixing reagents A, B, and C (provided by the vendor) at a ratio

dx.doi.org/10.1021/pr500242m | J. Proteome Res. XXXX, XXX, XXX−XXX

of 50:48:3. Next, a 10 μL aliquot of depleted BS samples was diluted in 140 μL of 50 mM PBS. BSA standard samples and depleted BS samples were then mixed with 150 μL aliquots of the working buffer and transferred to a 96-well plate prior to incubation at 37 °C for 2 h. The concentration was then measured at 620 nm wavelength on Multiskan plate-reader (Thermo Scientific, Rockford, IL). The calculated concentration was multiplied by the dilution factor to determine the original concentration of the depleted BS samples. These numbers suggest a 90.8% depletion of high abundant proteins.

Lectin Affinity Enrichment of N-Linked Glycoproteins using SNA and AAL Mixture

Two agarose-bound lectins were used to enrich glycoproteins from 49 depleted blood serums associated with three esophagus diseases and disease-free subjects. *Sambucus nigra* agglutinin (SNA, 3 mg/mL of settled gel) is binding to sialic acid, while *Aleuria aurantia* lectin (AAL, 2 mg/mL of settled gel) is specific to bind fucose-related structures. 60 μL aliquots of SNA and 90 μL aliquots of AAL were mixed and washed five times with 200 μL of 50 mM PBS. 49 lectin mixtures were prepared for each depleted BS subject. The depleted BS samples were mixed with lectin mixtures and incubated overnight with agitation at 4 °C. The unbound fractions were then discarded after the centrifugation at a speed of 2000*g* for 5 min. The bound fractions were washed five times with 50 mM PBS to remove nonspecific binding proteins. The enriched glycoproteins were released by incubation with 400 μL of 100 mM acetic acid at 4 °C with agitation for 4 h. The eluted fractions were collected and subjected to buffer exchange, as previously described. The exchanged buffer was 50 mM ammonium bicarbonate. BCA protein assay was then performed to determine the amount of enriched glycoproteins.

Enzymatic Digestion of Lectin Affinity Enriched N-Linked Glycoproteins

Thermal denaturation was performed at 65 °C for 10 min. 200 mM DTT and 200 mM IAA solutions were prepared in 50 mM ammonium bicarbonate. A 1.25 μL aliquot of 200 mM DTT solution was then added to the samples prior to incubation at 60 °C for 45 min. The reduced glycoproteins were then alkylated through the addition of a 5 μL aliquot of 200 mM of IAA and incubation at 37.5 °C for 45 min. A second 1.25 μL aliquot of 200 mM DTT was added to consume excess IAA, followed by the incubation at 37.5 °C for 30 min. The samples were then tryptically digested and incubated at 37 °C overnight, followed by microwave digestion at 45 °C and 50 W for 30 min. The amount of trypsin added to the samples (enzyme/substrate ratio of 1:25 w/w) was determined based on the glycoprotein concentration values calculated from micro BCA protein assay. The enzymatic digestion was quenched by the addition of 0.5 μL of aliquots of neat formic acid to the samples. Then, the samples were lyophilized and resuspended in 0.1% formic acid prior to LC−MS/MS analyses.

Fetuin Enrichment Study

The samples for fetuin enrichment were prepared as follows. Hydrophilic interaction liquid chromatography (HILIC) was employed to enrich fetuin glycopeptides using cotton wool.[23] A 5-mg of cotton wool was weighed and packed into 200 μL microtips. The cotton-packed microtips were washed with 100 μL of 10% ACN/0.1% FA five times. Then, the tips were conditioned with 83% ACN. A 5 μL (1 μg) aliquot of fetuin digests was prepared in 95 μL of 83% ACN. The fetuin digests were applied to the cotton HILIC microtip and subjected to incubation for 1 h. The immobilized glycopeptides were washed with 83% ACN/0.1% FA by three times to remove peptides. The glycopeptides were collected using the elution buffer, 10% ACN/0.1% FA. The fetuin glycopeptides were lyophilized and suspended in 0.1%FA before LC−MS/MS analysis.

After fetuin HILIC, 1×, 2×, and 4× of fetuin glycopeptides volume were spiked into 2 μg of a four standard glycoprotein mixture. Those volumes were equivalent to 100, 200, and 400 ng of fetuin digests. The spiked samples were lyophilized and resuspended in 5 μL of 0.1% FA. LC−MS/MS analysis was performed with the same method of fetuin run.

LC−MS/MS Conditions for Patient Study

The LC−MS/MS conditions for the pooled samples are available from our prior work.[20] Here we describe the details of the analysis of patient samples. LC−MS/MS analysis was performed using Dionex 3000 Ultimate nano-LC system from Dionex (Sunnyvale, CA) interfaced to LTQ Orbitrap Velos mass spectrometer (Thermo Scientific, San Jose, CA) equipped with a nano-ESI source. The samples were online-purified using a PepMap 100 C18 cartridge (3 μm, 100 Å, Dionex) and separated using a PepMap 100 C18 capillary column (75 μm id × 150 mm, 2 μm, 100 Å, Dionex). Solvent A was an aqueous 2% ACN solution containing 0.1% formic acid, while solvent B is 98% ACN solution with 0.1% formic acid. Tryptically digested serum glycopeptides were analyzed with different LC gradients and scan times and separated based on 10% solvent B at 350 nL/min flow rate over 10 min, 10−35% over 260 min, 35−80% over 10 min, 80% over 18 min, 80−10% over 1 min, and 10% over 1 min. The separation and scan time was set to 180 min. The separate blank run (50 min) was performed to remove the carryover between each sample.

The LTQ Orbitrap Velos mass spectrometer was operated with three scan events. The first scan event was a full MS scan of 380−2000 m/z at a mass resolution of 15 000. The second scan event was CID (collision-induced dissociation). MS/MS of precursor ions selected from the first scan event with an isolation width of 3.0 m/z was subjected to a normalized collision energy (CE) of 35% and an activation Q value of 0.250. The third scan event was set to acquire HCD (higher-energy collision dissociation) MS/MS of the precursor ions selected from the first scan event. The isolation width of HCD experiment was set to 3.0 m/z, while the normalized CE was set to 45% with an activation time of 0.1 ms. The CID and HCD MS/MS were performed on the eight most intense ions observed from the first MS scan event. The LTQ Orbitrap Velos mass spectrometer was externally calibrated, permitting <2 ppm mass accuracy.

GlycoMapSera_Pooled Glycomap

The details of creating a glycomap from the pooled serum samples were given in our previous work.[20] In brief, the 12 data sets from the pooled samples were analyzed using Multi-Align[24] and GlycoFragwork.[20] MultiAlign identifies common ions in replicate experiments using mass and NET and also reports abundance in each experiment (elution peak maximum as default) along with start and stop scan numbers of elution within each data set. GlycoFragwork uses the reported scan to collect fragmentation spectra for each common ion across all data sets and scores them based on fragmentation methods. Glycopeptide searching was performed against a theoretical database of 101 glycoproteins, derived from a list of 142

dx.doi.org/10.1021/pr500242m | J. Proteome Res. XXXX, XXX, XXX−XXX

identified proteins after filtering for presence of N-glycosylation motif. A collection of representative-GSMs (glycopeptide-spectra matches) were acquired that contained glycopeptide assignments along with different fragmentation spectra (HCD, CID, and ETD). Using a combination of ETD scoring and CID glycan sequencing, FDR (false discovery rates) for assignments are estimated. Finally, a list of 33 glycoproteins with 103 glycopeptides on 53 glycosylation sites were identified from rep-GSMs with <0.05 FDR. These 103 glycopeptides were compiled into a serum glycomap (called GlycoMapSera_Pooled) and used for quantification across pooled samples. The maximum value of the elution peak was used as a measure of glycopeptide abundance.

GlycoMapSera_Individual Glycomap

Apart from glycopeptide quantification across pooled samples, data sets from individual samples across four groups (CA, HGD, BD, and DF) were also analyzed for quantitative differences. Because these data sets had only HCD/CID spectra, the glycopeptide identifications for these data sets were accomplished by alignment to GlycoMapSera_Pooled. The details of the alignment are given below. After these procedures, a list of 40 glycopeptides across 26 glycosylation sites from 22 glycoproteins were identified and compiled into GlycoMapSera_Individual. Similar to the pooled study, the elution profile maximum is used as metrics of abundance, although other measures such as peak area, sum of observed abundances, or total-intensity chromatogram can be also used.

Alignment Algorithms for Glycopeptide Identification

Consider a target input map F that contains LC−MS clusters with mass, NET, and CID fragmentation spectra. Such a map can be built using MultiAlign[24] and GlycoFragwork[20] as before, except GlycoFragwork is operated without using a reference database. Using a reference glycomap G (such as SerumGlycoMap_Pooled) that contains mass, NET, CID fragmentation spectra, and identified glycopeptides, the objective is to assign glycopeptide identifications to the input map based on matching through mass, NET, and CID spectra. We use the procedure described later to achieve this.

First, both target and reference maps are divided into bins of equal width (empirically chosen as 0.05) along the NET dimension (as shown in Supplementary Figure 1 in the Supporting Information). Each bin of target map is then matched with every bin of the reference based on the LC−MS features within two bins. Every feature, termed as a glycorecord, is essentially an LC−MS cluster represented as a mass and NET pair. A hit is defined as a glycorecord in a target NET bin with a matching counterpart in any reference NET bin within a mass tolerance (default 10 ppm) and a NET tolerance (default 0.025). A glycorecord that fails to match any feature in the reference bins is classified as a nonhit. In this case, a pseudo-glycorecord is introduced in the reference with the maximum tolerance from the target glycorecords, thereby penalizing the absence of a hit.

The target glycorecord could have multiple hits in the reference glycomap within the given tolerances. To resolve this, we assign it to the closest feature in the 2D reference map. However, if the target and reference maps contain CID MS/MS spectra, a spectral correlation score is used for assigning the matching hit. First, the target and reference MS/MS spectra are binned into spectral bins of size 0.5 m/z between the range 400−2000 m/z. We then calculate the similarity between target and reference MS/MS spectra using the cosine-similarity between peaks in two MS/MS spectra

$$\text{cos\_sim} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

where A_i and B_i are the intensities of matching peaks (i.e., in the same bin) within the target and reference MS/MS spectra, respectively, and n is the total number of bins.

We extend the scoring model of Jaitly et al.[25] to create a similarity matrix S using the mass and NET differences between target and reference glycorecords

$$S(F_i, G_j) = -\left[\sum_{k=1}^{p}\left(\ln(\text{cos\_sim}_k) + \ln(\sigma_{\text{mass}}\sigma_{\text{NET}}2\pi) - \frac{\Delta^2_{\text{mass}_k}}{\sigma^2_{\text{mass}}} - \frac{\Delta^2_{\text{NET}_k}}{\sigma^2_{\text{NET}}}\right)\right]^{-1}$$

$$\forall\ 0 \le i < m,\ 0 \le j < n$$

where m is the number of bins in target map, n is the number of bins in reference map, p is the total number of hits and nonhits, σ_{mass} is the standard deviation of mass difference distribution, σ_{NET} is the standard deviation of NET difference distribution, Δ_{mass_k} is the mass difference between two glycorecords in the target and reference maps, and Δ_{NET_k} is the NET difference between two glycorecords in the target and reference map

The minimum score in the scoring matrix is selected as the gap penalty (denoted in the alignment equations as d) so as to assign the highest penalty for introducing a gap in the alignment. After computing the similarity matrix, the alignment matrix H is calculated by a dynamic programming algorithm, similar to the dynamic time warping algorithm.[26] Here we utilize the equations used in Finney et al.[27]

$$H(i, 0) = i*d,\ \forall\ 0 \le i < m$$

$$H(0, j) = j*d,\ \forall\ 0 \le j < n$$

$$H(i, j) = \max\begin{Bmatrix} H(i-1, j-1) + 2.1*S(i, j) & \\ H(i-1, j) + d, & \forall\ 1 \le i < m,\ 1 \le j < n \\ H(i, j-1) + d & \end{Bmatrix}$$

Finally in the trace-back step, the glycopeptide annotation of each reference glycorecord is transferred to its matched feature in the target map that corresponds to a hit along the alignment path. The efficacy of the alignment algorithm was tested by using simulated maps that involved matching with deletions, shift, and insertions of bins in the maps. (For details, see Supplementary Figures 1−3 in the Supporting Information.)

Statistical Model for Quantification of Site-Specific Glycosylations

To detect differential site-specific glycosylation in glycoproteins across different classes (e.g., cancer and control) of samples, an ANOVA-based model was developed, which extended the methods described in Hill et al.[6] and Clough et al.[9] for detecting differential peptide/protein abundances. ANOVA stands for analysis of variance and is used to determine the effect of explanatory variables on a measured value. While standard machine-learning prediction algorithms attempt to capture the relationship between a binary response (e.g., healthy = 0, cancer = 1) and a set of features, ANOVA models the dependence of the response variable on covariates (as experimental units) that were observed with the response. The

dx.doi.org/10.1021/pr500242m | J. Proteome Res. XXXX, XXX, XXX−XXX

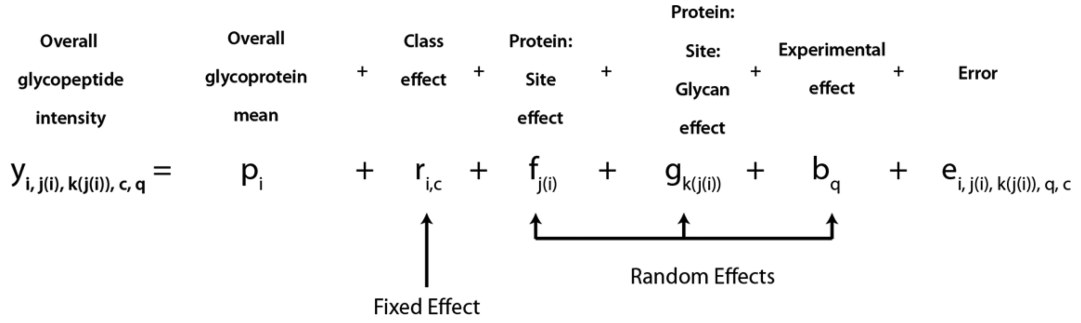


Figure 1. Mixed effects ANOVA model for intact glycopeptide quantification. The abundance of a glycopeptide can be broken down into contributions from a summation of factors. The class effect is denoted as fixed, whereas the rest of the factors are deemed to be random.

model here uses abundance as response variable, and covariates are categorical terms accounting for glycoproteins, their corresponding peptides (i.e., glycosylation sites, assuming each peptide containing one glycosylation site), and glycans attached to the sites, along with relevant interaction terms. In mixed-effects models such as ours, the terms (called effects) could be either "fixed" or "random". Fixed-effects parameters represent values from a discrete set of possibilities, where random-effect parameters represent values randomly sampled from an unknown distribution. For example, the covariate associated with class is fixed because this parameter can be one of two values: healthy or disease. The interaction between a glycan and a site is a good example of a random effect because the observed values represent a sample from the population of all peptides with all glycans. The aim of our model is to determine the source of variation in glycoprotein abundances between healthy and disease. In particular, we are interested in cases where the variation stems from not just the difference in classes but also in a site-specific manner; a particular glycosylation site may vary in abundance to a larger degree than another site (protein−site interaction effect), or a site may show an increased preference for a glycan (i.e., the glycan−site interaction effect).

Consider a glycomap (such as GlycoMapSera_Pooled) containing glycoprotein abundances across two classes: disease (c = 1) and healthy (c = 2). Let the glycomap contain I glycoproteins constructed from a set of N replicate (technical or biological) samples. We represent the observed abundance of a particular glycopeptide within glycoprotein i, with a corresponding site j, glycan k, and observed in data set q as a product term

$$Y_{i,j(i),k(j(i)),c,q} \approx [P_i \cdot R_{i,c} \cdot F_{j(i)} \cdot G_{k(j(i))} \cdot B_q \cdot E_{i,j(i),k(j(i)),c,q}] \quad (1)$$

where P_i is the average abundance of glycoprotein i across all samples, i = 1,...,I, where I is the number of glycoproteins in sample; $R_{i,c}$ is the abundance variation of glycoprotein i across two classes (disease or control); $F_{j(i)}$ is the abundance variation of all glycopeptides containing the same glycosylation site j from the glycoprotein I (but with different glycans attached); $G_{k(j(i))}$ is the abundance variation of the glycopeptide containing glycan k attached to the glycosylation site j from the same glycoprotein i; B_q is total abundance variation in a specific experiment q, q = 1,...,N, where N is the number of data sets; and $E_{i,j(i),k(j(i)),c,q}$ is the random error of the abundance measurement for a specific glycopeptide in a specific experiment.

The term $j(i)$ indicates the glycosylation site j in glycoprotein i and $k(j(i))$ represents the index of glycan that is present at the site j that in turn is from glycoprotein, thereby indicating nesting effects. We note that "nested" interaction effects (e.g., the glycan effect indicated by $k(j(i))$) are different from "crossed" effects, which enumerates the effects of all possible glycans with all possible sites within a glycoprotein. Here we are interested in the alteration of the particular glycosylated sites with particular glycans between the disease and the healthy classes. To build the linear model, the log-2 transformation is conducted on the measured abundances because expression changes are typically multiplicative. The log-2 transformation also acts as a simple normalization scheme within a glycoprotein; additional normalization methods have been suggested such as using protein content[18] or most-abundant glycoform[19] as a normalization factor. After log-2 transformation, the previously described product term becomes

$$y_{i,j(i),k(j(i)),c,q} = p_i + r_{i,c} + f_{j(i)} + g_{k(j(i))} + b_q + e_{i,j(i),k(j(i)),c,q} \quad (2)$$

The model in eq 2 was solved using an ANOVA-mixed effects model. Each term was considered as a factor and defined as either fixed or random (Figure 1). The model can be then solved for estimating the model coefficients. We used the *lmer* command in the **lme4** package[28] within the statistics package R to solve the model. The lmer method was chosen so that parameters could be estimated through maximum likelihood estimates (MLEs). Additionally, lmer is also ideal for unbalanced, multiple-nested effects model data sets such as ours. The restrictive maximum likelihood criterion (REML) within the lmer command was set to FALSE to allow hypothesis testing using log-likelihood methods. Null models can be constructed by removing terms according to the nature of the test being conducted. For example, by constructing a null model containing class and experimental terms, we can test for differential site-specific glycosylation in each glycoprotein. Equations 3 and 4 show the full and null versions of linear mixed-effect modeling. The full model comprises fixed effects from class and random effects from site, glycan, and experimental effects, where the null model comprises fixed effects from class, but the random effects are from experimental effects only. This facilitates the hypothesis testing using log-likelihood ratio statistics, as shown in eq 5. The test statistic, which follows a chi-square distribution (indicated in eq 5 by $\overset{d}{\rightarrow}$), tests for significance of the site and glycan terms, thereby assessing the effects of site-specific glycosylations between the classes while accounting for experimental and overall class variation.

dx.doi.org/10.1021/pr500242m | J. Proteome Res. XXXX, XXX, XXX−XXX

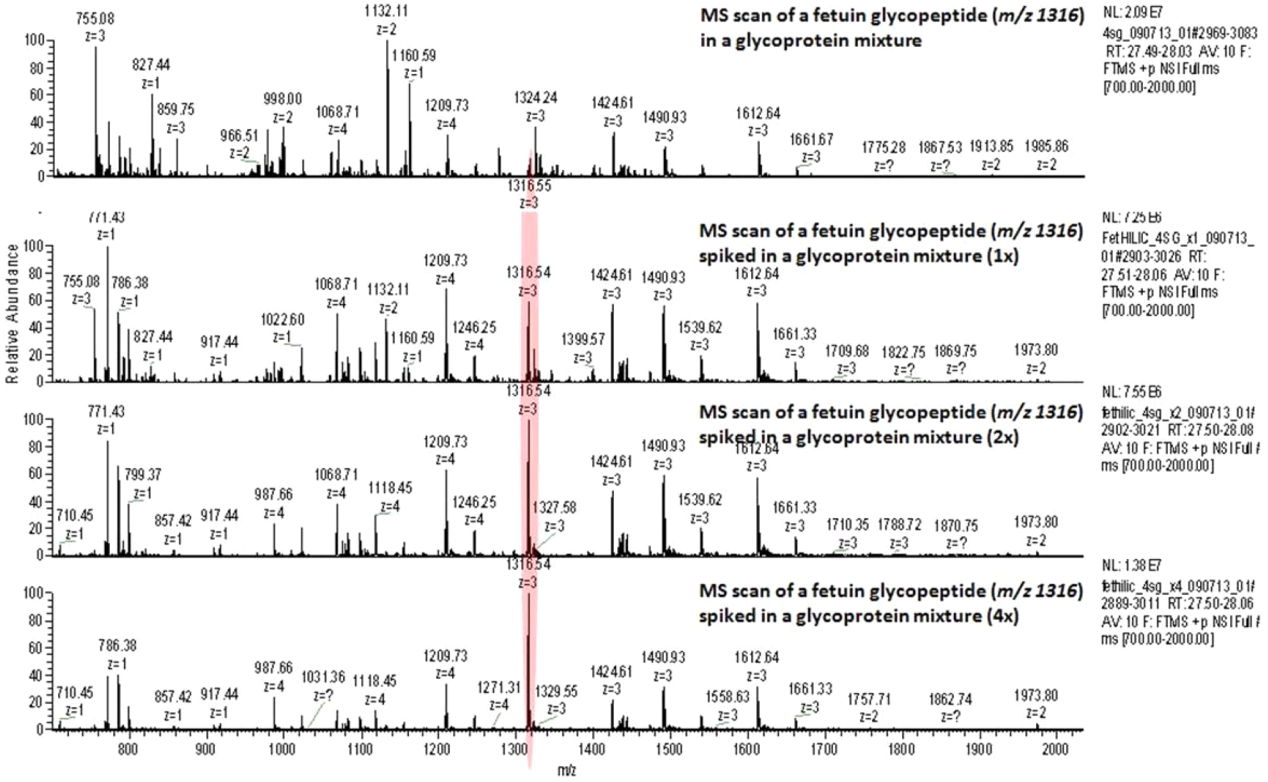


Figure 2. MS scan plots for data sets used in the fetuin study. The fetuin glycopeptide at 1316 m/z corresponds to the diattenary disialylated glycan at site N-156 on the fetuin protein. As can be seen, the intensity of elution peak corresponding to this glycopeptide increases steadily in abundance with increase in fetuin concentration as opposed to remaining peaks.

$$\Lambda_{\text{full}}: y \approx 1 + \text{Class} + (1|F) + (1|G) + (1|\text{experiment}) \quad (3)$$

$$\Lambda_{\text{null}}: y \approx 1 + \text{Class} + (1|\text{experiment}) \quad (4)$$

$$\lambda_{\text{LRT}} = -2\frac{\log \text{Lik}(\Lambda_{\text{null}})}{\log \text{Lik}(\Lambda_{\text{full}})} \xrightarrow{d} \chi_2^2 \quad (5)$$

RESULTS AND DISCUSSION

Quantification of Model Glycoprotein in Fetuin Spike-in Study

As a proof of concept, we tested the quantification model on a constructed glycomap (through a combination of GlycoFragwork and manual curation) from four data sets with increasing quantities of fetuin among a standard mixture of four glycoproteins: fetuin, alpha-1-acid glycoprotein, fibronectin, and pig thyroglobulin. As can be seen from Figure 2, the concentrations of fetuin glycopeptides steadily increase, while the concentrations of other glycoproteins and associated glycopeptides remain the same. We imposed a strict filtering condition that glycopeptides had to be present in three out of four data sets, and the missing values were imputed using nonzero average of the observed abundances. Abundance values of 14 glycopeptides from 9 sites across 4 glycoproteins were analyzed using the quantification model in eq 2 with a few changes: each of the four data sets is indicated to be a separate class, and there are no experiment terms.

We tested for class effects by log-likelihood ratio testing of the mixed effects model against a null model that contains the site and glycan random effects. As shown in Table 1, only fetuin shows a significant class effect difference, as expected. This result illustrates that our quantification model is effective in detecting glycoproteins with altered abundances across multiple samples.

Table 1. Glycoprotein p Values for the Fetuin Study from Log-Likelihood Ratio Testing for Class Effect[a]

proteins	p values
sp|P12763|FETUA_BOVIN	0.003176557
sp|P02763|A1AG1_HUMAN	0.122103951
sp|P02751|FINC_HUMAN	0.402179912
sp|P01267|THYG_BOVIN	0.555350130

[a]Only fetuin appears to be significant (p value <0.005), which agrees with the experimental setup.

Quantification of Intact Glycopeptides in GlycoMapSera_Pooled

Within GlycoMapSera_Pooled, the data sets containing ETD in both cancer and control classes were ignored because these samples were at a higher concentration. Glycopeptides that were observed in at least three (out of five) replicates in each group are retained. The missing values (either from random chance or insufficient isotopic envelope information for accurate deisotoping) were imputed using the average of the observed nonzero abundances in the remaining replicates. This filtering criterion will preclude detection of glycopeptides present only in one class; however, our intention here is to depict the efficacy of the quantification model for biomarker discovery through proof-of-concept. Additional methods of accounting for missing values can be found in work by Karpievitch.[8] After these filtering and imputation steps, 27

dx.doi.org/10.1021/pr500242m | J. Proteome Res. XXXX, XXX, XXX−XXX

glycopeptides from 21 sites across 17 glycoproteins were retained for quantitative analysis. Figure 3 depicts a scatter plot

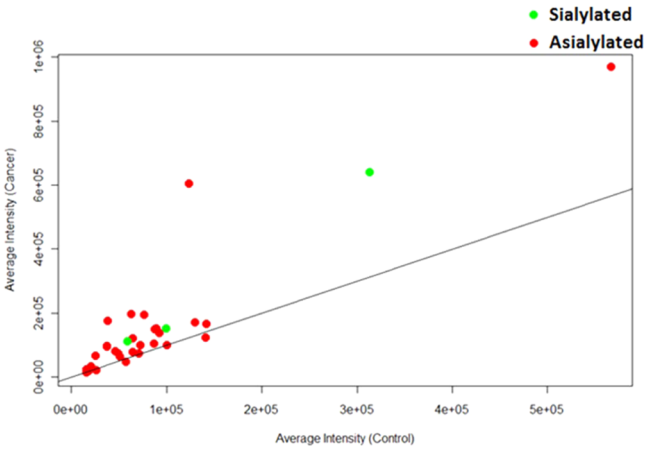


Figure 3. Glycopeptide abundance scatter plots between cancer and control. Green circles are glycopeptides with nonsialylated glycans and red circles are those with sialic acid terminations. Glycopeptides on the linear intercept-0 slope-1 line do not show abundance chance between cancer and control. Off-line glycopeptides indicate abundance change.

of average abundances of glycopeptides between cancer and control groups. The green points are nonsialylated glycopeptides, which, on examination, were glycopeptides with high-mannose attached glycans from the complement C3 glycoprotein (site N-85). All points on the straight line do not show a change in abundance, but the off-line glycopeptides show a differential change between cancer and control.

To detect differential abundances at site-specific levels within a glycoprotein in the pooled samples, we analyzed the 17 glycoproteins in GlycoMapSera_Pooled, one protein at a time, using the quantification model previously described. The hypothesis testing for significance for the site and glycan effects in the linear mixed effects model was performed, as shown in eq 2. Table 2 describes the 17 glycoproteins along

Table 2. Glycoprotein *p* Values from Log-Likelihood Ration Testing for Site-Specific Glycosylation in GlycoMapSera_Pooled[a]

protein	*p* values
sp|P04004|VTNC_HUMAN	8.80×10^{-17}
sp|P02790|HEMO_HUMAN	1.29×10^{-11}
sp|P01024|CO3_HUMAN	1.21×10^{-6}
sp|P00738|HPT_HUMAN	0.000372122
sp|P00450|CERU_HUMAN	0.606276481
sp|P02749|APOH_HUMAN	0.999999995
sp|P04196|HRG_HUMAN	0.999999998
sp|P00734|THRB_HUMAN	0.999999999
sp|P01023|A2MG_HUMAN	0.99999999
sp|P01008|ANT3_HUMAN	1
sp|P01009|A1AT_HUMAN	1
sp|P01011|AACT_HUMAN	1
sp|P02751|FINC_HUMAN	1
sp|P02763|A1AG1_HUMAN	1
sp|P04217|A1BG_HUMAN	1
sp|P05546|HEP2_HUMAN	1
sp|P08603|CFAH_HUMAN	1

[a]Glycoproteins with *p* values <0.005 are deemed to be significant.

with the *p* values arising from the log likelihood test comparison of the mixed effects against the null model. A low *p* value (<0.005) indicates that a significant difference in levels of site-specific glycosylation occurred in the glycoprotein between cancer and control. In a large-scale glycoproteomics study, multiple hypothesis testing must be undertaken, as the probability of making a Type-I error (i.e., detecting a glycopeptide of "significance" when the null hypothesis that there is no variation between two groups is true) increases with greater number of tested hypotheses.[29] Typical genomic and proteomic experiments contain a large number of multiple hypotheses, for which the probability of making at least one Type-I error (defined as family wise error rate (FWER)) is adjusted through methods such as Bonferroni or Holm's. Similar approaches can be used to control the FWER in large-scale glycoproteomics experiments.

The utility of using the mixed-effect quantification model is illustrated in Figure 4, which depicts glycopeptide abundance profiles between cancer and control for the glycoproteins that were found to be significant (i.e., *p* value < 0.005). The distribution of glycopeptide abundance across replicates within a group is expressed as a box plot. Cancer and control replicates are stacked right next to each other. This is done for all glycopeptides identified for a particular glycoprotein, thereby indicating abundance profiles at all sites and glycan levels for a particular glycoprotein within a single plot. The glycan sequences for each glycopeptide were obtained from the sequencing algorithm and then matched to structures pertaining to humans that are observed in the CFG database.[30] As can be seen from Figure 4, the ANOVA model can detect site-specific glycosylation behavior between classes. Supplementary Table 1 in the Supporting Information compares the *p* values for proteins considered significant using our ANOVA model with a standard *t* test performed per glycoprotein. Outside of haptoglobin, the *t* test fails to detect any significance (*p* value <0.005) in glycoprotein variation between cancer and control, which contradicts observed behavior in Figure 4.

In Figure 4a, abundance profile for glycopeptides from haptoglobin is shown with two sites and two sialylated glycans. Although haptoglobin is one of the proteins targeted for depletion by the MARS column, it is not completely depleted on account of secondary interactions with other proteins or homologous proteins. Moreover, if this protein is in higher abundance in one sample then it will not be effectively depleted relative to another sample in which it is less abundant. All haptoglobin glycopeptides are observed to increase in abundance in cancer, but the disialylated glycopeptides at site N-184 and N-241 showed a larger extent of increase than the monosialylated one at N-184. Haptoglobin is a common biomarker and has been previously reported as being increased in inflammatory processes and also in ovarian[31] and pancreatic cancer.[32] From the spectral count data collected from the Mascot identification of nonglycosylated peptides (shown in Supplementary Table 2 in the Supporting Information), we observe that an elevated spectral count for the protein correlating with the increased abundance is shown across all quantified glycopeptides.

Figure 4b depicts the similar results for the glycoprotein CO3_HUMAN (complement C3 protein) that was found with one glycosylation site (N-85). All sites were attached to glycans of high-mannose type. Again, all glycopeptides showed increased abundance in cancer with a slight variation among them. High-mannose glycans have been associated regularly

dx.doi.org/10.1021/pr500242m | J. Proteome Res. XXXX, XXX, XXX−XXX

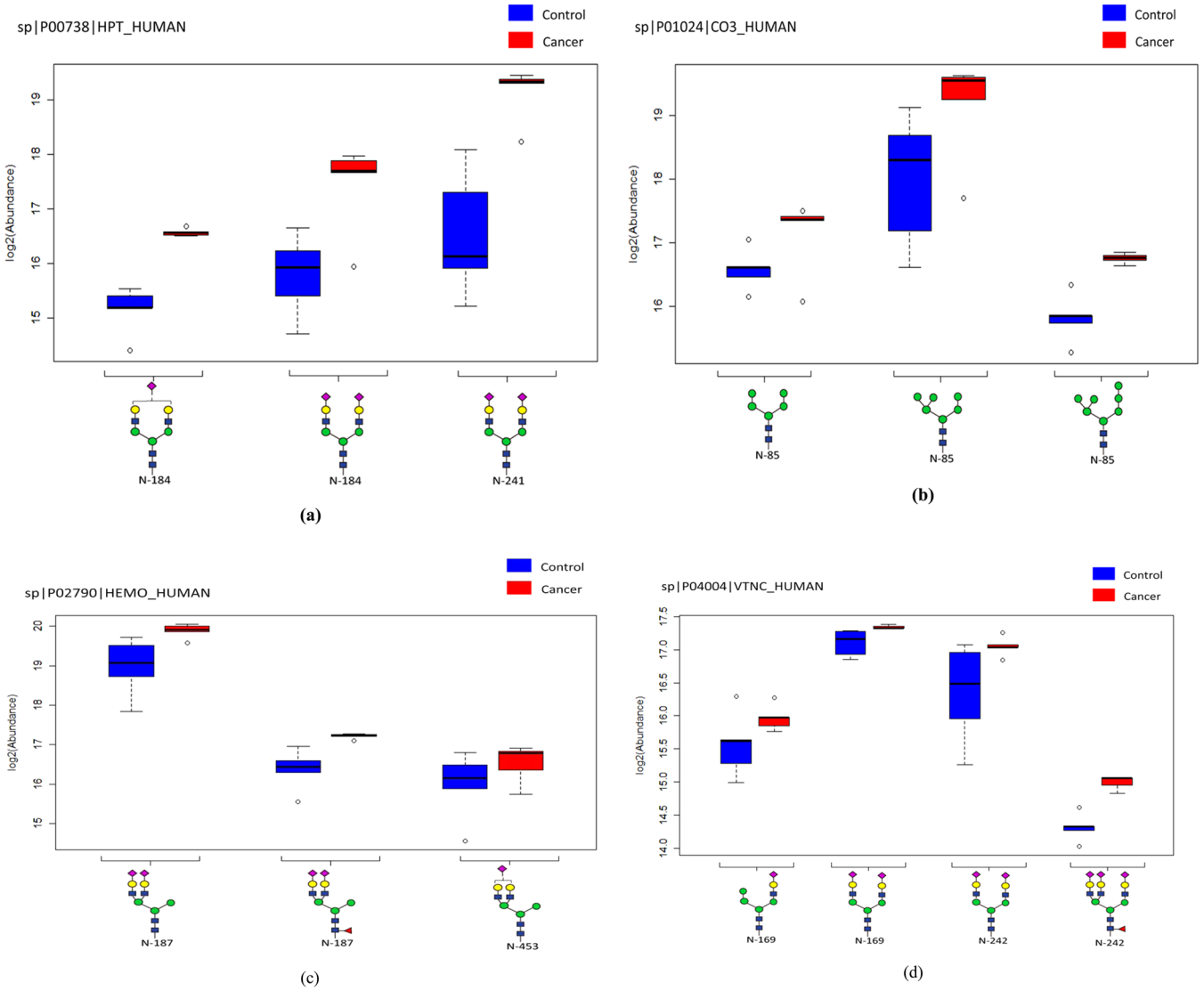


Figure 4. Box plots depicting glycopeptide abundance profiles of proteins that were found to be significant (p value < 0.005): (a) haptoglobin, (b) complement C3 protein, (c) hemopexin, and (d) vitronectin.

with elevation in cancer, for example, breast cancer.[33] Incidentally, the complement C3 glycoprotein is part of the complement cascade system that has been observed to play a role in cellular proliferation, especially in cancer.[34] The increased abundance in esophageal cancer was observed to be consistent at protein level as well from the observations in the spectral count data (Supplementary Table 2 in the Supporting Information). Because this is a highly abundant protein in human sera, it is currently being targeted as one of the depletion proteins in Agilent's latest top 10 depletion system. (Note that we used seven-protein depletion system, and thus the protein was retained). We show here that although the protein is of high abundance, the intact glycopeptides are low-abundant and show differential glycosylation between cancer and control samples.

Figure 4c depicts the abundance profile of the protein HEMO_HUMAN (hemopexin), which indicates an overall increase in glycosylation with monosialylated structures at site N-453 showing a lower increase compared with disialylated structures at site N-187. The glycoprotein was observed to be abundant, as observed from the protein spectral count. Hemopexin is a heme-binding protein whose fucosylated N-linked glycans have been observed to be significantly abundant in hepatocellular carcinoma.[35] Figure 4d shows the abundance profile of VTNC_HUMAN (vitronectin), which shows site-specific differential glycosylation. The increase in the fucosylated trisialylated glycopeptide at N-242 between cancer and control seems to be much more than the increase in the disialylated glycopeptide at N-169. It is also worth noting that in a separate study (manuscript in preparation) that performed MRM (multiple reaction monitoring) quantification blood serum glycoproteins enriched by lectin affinity chromatography and hydrazide chemistry also identified hemopexin and vitronectin as highly expressed glycoproteins in cancer with p value < 0.05. While it remains to be seen whether these glycoproteins that show site-specific differential glycosylation are biomarkers of esophageal cancer, the utility of the framework and the quantification model for selecting biomarker candidates are illustrated here.

Construction of GlycoMapSera_Individual

As previously mentioned, GlycoMapSera_Individual is the glycomap built from data sets corresponding to serum samples

dx.doi.org/10.1021/pr500242m | J. Proteome Res. XXXX, XXX, XXX−XXX

from 15 esophageal cancer patients, 15 disease-free individuals, 7 patients suffering from Barrett's disease, and 12 patients suffering from high-grade dysplasia. The details of building this glycomap are as follows. First, an input map containing mass and NET values for ions was built using MultiAlign. Ions had to be present in at least 2 out of 49 data sets to be counted. The map was then aligned to GlycoMapSera_Pooled using alignment algorithms previously described. Annotations from verified and tentative glycorecords were transferred from a reference to target. The verified records lead to 40 glycopeptides across 26 glycosylation sites from 22 glycoproteins. These numbers are lower than GlycoMapSera_Pooled, but this is expected due to lack of ETD. In a separate study, we built a new glycomap by combining the pooled and the individual data sets using GlycoFragwork and found significant overlap with the identifications in GlycoMap_Individual (data not shown). This further proves the accuracy of our alignment algorithms and raises the possibility of using them as a standard means of identifying glycopeptides using mass, NET, and CID spectra.

Quantification of Intact Glycopeptides in GlycoMapSera_Individual

The quantification model was applied to GlycoMapSera_Individual for detecting site-specific variation across four classes (CA, DF, BD, and HGD). First, filtering was done on GlycoMapSera_Individual using a minimum observance in 5 data sets out of a total of 49 data sets. Missing values were left as such; they were not imputed due to the biological replicate nature of the data. Quantification was done using the mixed-effect models in eq 3, and likelihood-ratio test statistic, shown in eq 5, was used for detecting the presence of site-specific differential glycosylation.

Table 3 describes the 14 glycoproteins (obtained after filtering) along with the p values arising from the log-likelihood test comparison of the mixed effects against the null model. As before, a significant p value (<0.005) indicates that significantly different levels of a site-specific glycosylation occurred in the glycoprotein between all four classes (as shown in Figure 5). Interestingly, vitronectin is once again detected to display site-specific differential glycosylation over all four classes, as can be observed from Figure 5a. There is a distinct increase in the triantenarary trisialylated, and fucosylated glycan at site N-242 in cancer as compared with other sites, which is in accordance to the trend observed in the pooled study. Furthermore, compared with cancer, levels in BD and HGD disease groups are much closer to DF. While this behavior remains to be explored in the context of esophageal cancer etiology, the utility of the quantitation model for selecting biomarker candidates is highlighted here. Other glycoproteins that are involved in site-specific differential behavior across four groups are ceruloplasmin (CERU_HUMAN, Figure 5b), alpha-2-macroglobulin (A2MG_HUMAN, Figure 5c), and complement factor I (CFAI_HUMAN, Figure 5d). Multiple glycopeptides were observed for ceruloplasmin in the pooled study without any evidence of site-specific differential glycosylation. However, on testing for a class effect, ceruloplasmin was found to be significant in the pooled study (data not shown). Ceruloplasmin, a nonspecific disease marker, is a copper-carrying protein involved in intracellular transport and is an established diagnostic marker for cancer.[36,37] Both A2MG_HUMAN and CFAI_HUMAN were observed in the pooled study, but no evidence of class, site, or glycan effects was observed. Reciprocally, glycopeptides from hemopexin, complement C3, and haptoglobin were not detected in the individual analysis for two reasons, either because the calibrated mass and NET used for glycopeptide detection was outside tolerance limits or because the particular glycopeptide was not observed in more than five data sets.

Table 3. Glycoprotein p Values from Log-Likelihood Ratio Testing for Site-Specific Glycosylation in Individual Patient Samples (GlycoMapSera_Individual)[a]

protein	p values
sp|P01023|A2MG_HUMAN	2.85×10^{-52}
sp|P00450|CERU_HUMAN	2.41×10^{-30}
sp|P05156|CFAI_HUMAN	1.31×10^{-11}
sp|P04004|VTNC_HUMAN	2.57×10^{-10}
sp|P08603|CFAH_HUMAN	0.809568567
sp|P01011|AACT_HUMAN	0.862884817
sp|P04196|HRG_HUMAN	0.883306338
sp|P13671|CO6_HUMAN	0.957992486
sp|P01008|ANT3_HUMAN	0.994443905
sp|P00734|THRB_HUMAN	1
sp|P02763|A1AG1_HUMAN	1
sp|P03952|KLKB1_HUMAN	1
sp|P04217|A1BG_HUMAN	1
sp|Q14624|ITIH4_HUMAN	1

[a]Glycoproteins with p values <0.005 are deemed to be significant.

The disease markers detected from quantification of GlycoMapSera_Pooled and GlycoMapSera_Individual are mostly high-abundant serum proteins and were routinely implicated in other cancer studies and thus may not be specific to esophageal cancer. However, the site-specific glycosylations associated with these proteins reported here are novel; herein lies the utility of the model presented here. It is now possible to study site-specific glycosylation behavior within a protein for purposes of biomarker discovery.

CONCLUSIONS

We illustrate that site-specific glycosylation can be characterized through analysis of intact glycopeptides. A novel ANOVA-based mixed effects model is presented for label-free glycopeptide quantification. Through the use of null models and log-likelihood testing, site-specific differential glycosylation between groups can be detected. The power of the model is in the selection of biomarker candidates that then can be verified using targeted studies. We applied our statistical model on glycomaps from an esophageal cancer study on both pooled (that were built by GlycoFragwork) and individual (wherein glycopeptide identification was made through alignment algorithms) data sets. The objective was to detect aberrant glycosylations in cancer samples with respect to control, Barrett's disease, and high-grade dysplasia samples. Among our candidates, vitronectin was consistently observed to show an aberrant site-specific glycosylation across many samples. A further study is required on accurately characterizing glycosylation changes in reported glycoproteins, not only for the sites mentioned in the quantification study but also across all sites. Further improvements in the model can be done particularly on the issue of missing data. Because the model is a linear framing of mixed effects, it can be expanded in several directions. Information on nonglycosylated peptides of the glycoprotein, the total abundance of the glycoprotein, and

dx.doi.org/10.1021/pr500242m | J. Proteome Res. XXXX, XXX, XXX−XXX

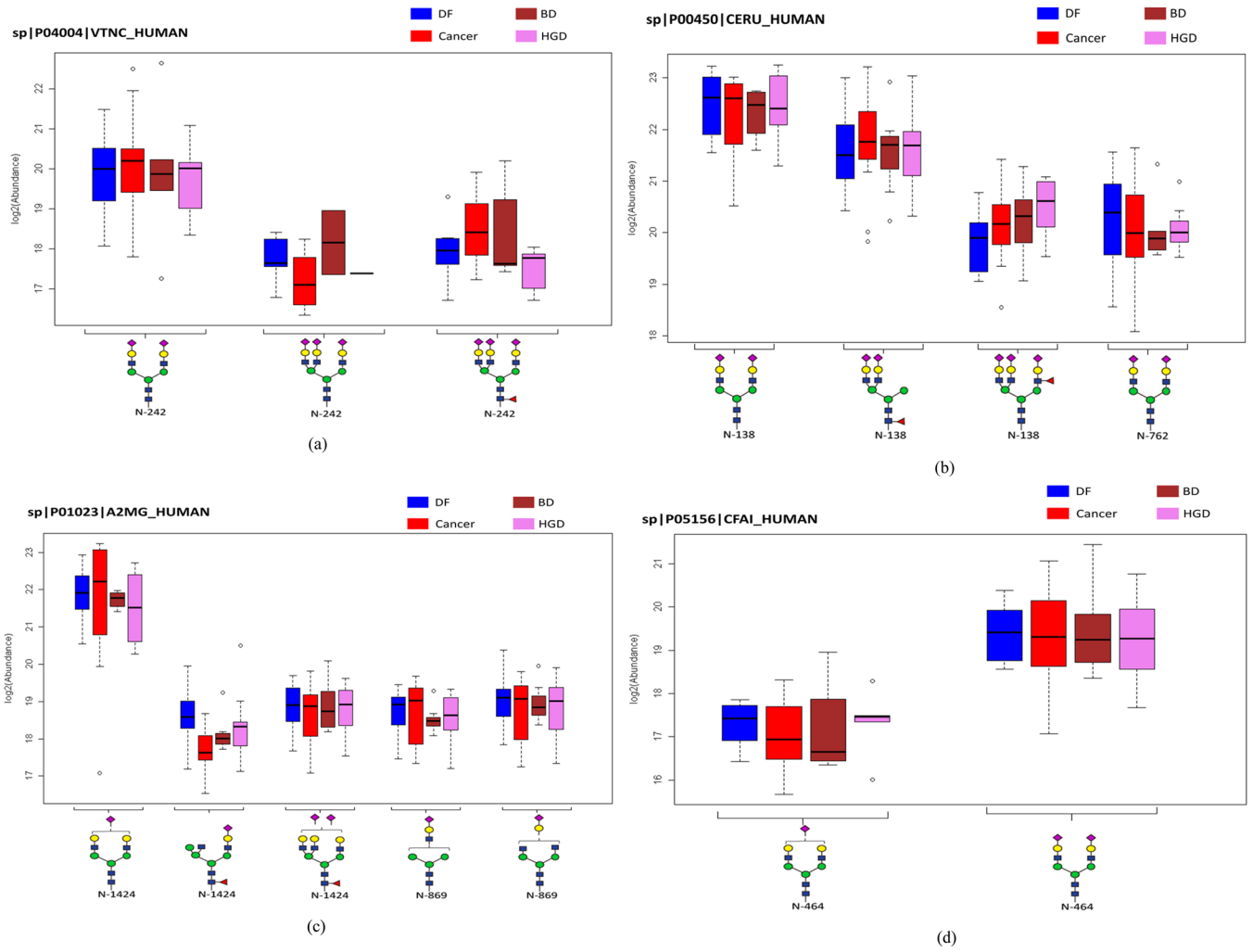


Figure 5. Box plots depicting glycopeptide abundance profiles of proteins that were found to be significant (p value < 0.005) in individual patient samples (a) vitronectin, (b) ceruloplasmin, (c) alpha-2-macroglobulin, and (d) complement factor I.

abundance information from a separate analysis of cleaved glycans can all be incorporated into the model, leading to a systems-biology approach to studying glycosylation effects in diseases.

The quantification model presented here allows for the detection of site-specific glycosylation events, which may be biologically important. Additionally, our model can be utilized to study other modifications such as phosphorylation and ubiquitination, provided the proper terms are constructed. In particular, detection of site-specific variation of a particular modification (or even between modification) is possible using this approach. On a separate note, the alignment algorithms presented here allow glycopeptide identifications for input maps that do not have any fragmentation spectra (or have only CID spectra) so that identifications can be made based on mass and NET (and CID spectra, if present).

■ ASSOCIATED CONTENT

Supporting Information

Supplementary Figure 1: GlycoMap alignment in the scenario when the target and reference are the same. Supplementary Figure 2: GlycoMap alignment in the scenario when the reference is the same as target map but with points missing. Supplementary Figure 3: GlycoMap alignment in the scenario when the reference is the same as target map but with both missing and shifted points. Supplementary Table 1: Comparison of p values of four significant proteins from the ANOVA model to those acquired from a simple t test. The t test fails to capture site-specific variation within a glycoprotein among classes. Supplementary Table 2: Average total protein spectral count in pooled cancer and control data sets. The spectral count was derived from doing a MASCOT search on nonglycosylated peptides of the glycoprotein. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Authors

*A.M.: E-mail: anmayampurath@uchicago.edu. Phone: (773) 702-5457. Fax: (773) 834-6818
*Y.M.: E-mail: yehia.mechref@ttu.edu. Phone: 806-834-8246. Fax: 806-742-1289.
*H.T.: E-mail: hatang@indiana.edu. Phone: (812)-856-1859. Fax: (812)-856-4764

Present Address

‖A.M.: Computation Institute, University of Chicago, Chicago, Illinois 60637, United States.

dx.doi.org/10.1021/pr500242m | J. Proteome Res. XXXX, XXX, XXX−XXX

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The work is supported by NSF grant DBI-0642897 (H.T.) and 1R01GM093322-03 (Y.M.). A.M. was supported by a fellowship from Persistent Systems.

■ REFERENCES

(1) Varki, A.; Cummings, R. D.; Esko, J. D.; Freeze, H. H.; Stanley, P.; Bertozzi, C. R.; Hart, G. W.; Etzler, M. E. *Essentials of Glycobiology*, 2nd ed.; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, 2008.

(2) Dwek, R. A. Glycobiology: Toward Understanding the Function of Sugars. *Chem. Rev.* **1996**, *96* (2), 683−720.

(3) An, H. J.; Kronewitter, S. R.; Leoz, M. L. A. d.; Lebrilla, C. B. Glycomics and disease markers. *Curr. Opin. Chem. Biol.* **2009**, *13* (5−6), 601−607.

(4) Miyoshi, E.; Nakano, M. Fucosylated haptoglobin is a novel marker for pancreatic cancer: Detailed analyses of oligosaccharide structures. *Proteomics* **2008**, *8* (16), 3257−3262.

(5) Nahnsen, S.; Bielow, C.; Reinert, K.; Kohlbacher, O. Tools for Label-free Peptide Quantification. *Mol. Cell. Proteomics* **2012**, *12* (3), 549−556.

(6) Hill, E. G.; Schwacke, J. H.; Comte-Walters, S.; Slate, E. H.; Oberg, A. L.; Eckel-Passow, J. E.; Therneau, T. M.; Schey, K. L. A Statistical Model for iTRAQ Data Analysis. *J. Proteome Res.* **2008**, *7* (8), 3091−3101.

(7) Oberg, A. L.; Mahoney, D. W.; Eckel-Passow, J. E.; Malone, C. J.; Wolfinger, R. D.; Hill, E. G.; Cooper, L. T.; Onuma, O. K.; Spiro, C.; Therneau, T. M.; Bergen, I. I. I. H. R. Statistical Analysis of Relative Labeled Mass Spectrometry Data from Complex Samples Using ANOVA. *J. Proteome Res.* **2008**, *7* (1), 225−233.

(8) Karpievitch, Y.; Stanley, J.; Taverner, T.; Huang, J.; Adkins, J. N.; Ansong, C.; Heffron, F.; Metz, T. O.; Qian, W.-J.; Yoon, H.; Smith, R. D.; Dabney, A. R. A statistical framework for protein quantitation in bottom-up MS-based proteomics. *Bioinformatics* **2009**, *25* (16), 2028−2034.

(9) Clough, T.; Key, M.; Ott, I.; Ragg, S.; Schadow, G.; Vitek, O. Protein Quantification in Label-Free LC-MS Experiments. *J. Proteome Res.* **2009**, *8* (11), 5275−5284.

(10) Hongsachart, P.; Huang-Liu, R.; Sinchaikul, S.; Pan, F.-M.; Phutrakul, S.; Chaung, Y.-M.; Chen, S.-T. Glycoproteomic analysis of WGA-bound glycoprotein biomarkers in sera from patients with lung adenocarcinoma. *Electrophoresis* **2009**, *30* (7), 1206−1220.

(11) Jung, K.; Cho, W.; Regnier, F. E. Glycoproteomics of Plasma Based on Narrow Selectivity Lectin Affinity Chromatography. *J. Proteome Res.* **2008**, *8* (2), 643−650.

(12) Pan, S.; Tamura, Y.; Chen, R.; May, D.; McIntosh, M. W.; Brentnall, T. A. Large-scale quantitative glycoproteomics analysis of site-specific glycosylation occupancy. *Mol. BioSyst.* **2012**, *8* (11), 2850−2856.

(13) Parker, B. L.; Palmisano, G.; Edwards, A. V. G.; White, M. Y.; Engholm-Keller, K.; Lee, A.; Scott, N. E.; Kolarich, D.; Hambly, B. D.; Packer, N. H.; Larsen, M. R.; Cordwell, S. J. Quantitative N-linked Glycoproteomics of Myocardial Ischemia and Reperfusion Injury Reveals Early Remodeling in the Extracellular Environment. *Mol. Cell. Proteomics* **2011**, *10* (8), M110.006833.

(14) Tang, Z.; Varghese, R. S.; Bekesova, S.; Loffredo, C. A.; Hamid, M. A.; Kyselova, Z.; Mechref, Y.; Novotny, M. V.; Goldman, R.; Ressom, H. W. Identification of N-Glycan Serum Markers Associated with Hepatocellular Carcinoma from Mass Spectrometry Data. *J. Proteome Res.* **2009**, *9* (1), 104−112.

(15) Mechref, Y.; Hu, Y.; Desantos-Garcia, J. L.; Hussein, A.; Tang, H. Quantitative Glycomics Strategies. *Mol. Cell. Proteomics* **2013**, *12* (4), 874−884.

(16) Kohler, J. J.; Patrie, S. M.; Orlando, R. Quantitative Analysis of Glycoprotein Glycans. In *Mass Spectrometry of Glycoproteins*; Humana Press: New York, 2013; Vol. *951*, pp 197−215.

(17) Ruhaak, L. R.; Miyamoto, S.; Lebrilla, C. B. Developments in the Identification of Glycan Biomarkers for the Detection of Cancer. *Mol. Cell. Proteomics* **2013**, *12* (4), 846−855.

(18) Hong, Q.; Lebrilla, C. B.; Miyamoto, S.; Ruhaak, L. R. Absolute Quantitation of Immunoglobulin G and Its Glycoforms Using Multiple Reaction Monitoring. *Anal. Chem.* **2013**, *85* (18), 8585−8593.

(19) Hua, S.; Hu, C. Y.; Kim, B. J.; Totten, S. M.; Oh, M. J.; Yun, N.; Nwosu, C. C.; Yoo, J. S.; Lebrilla, C. B.; An, H. J. Glyco-Analytical Multispecific Proteolysis (Glyco-AMP): A Simple Method for Detailed and Quantitative Glycoproteomic Characterization. *J. Proteome Res.* **2013**, *12* (10), 4414−4423.

(20) Mayampurath, A.; Yu, C.-Y.; Song, E.; Balan, J.; Mechref, Y.; Tang, H. A Computational Framework for Identification of Intact Glycopeptides in Complex Samples. *Anal. Chem.* **2014**, *86* (1), 453−463.

(21) Spechler, S. J. Barrett esophagus and risk of esophageal cancer: A clinical review. *JAMA, J. Am. Med. Assoc.* **2013**, *310* (6), 627−636.

(22) Yousef, F.; Cardwell, C.; Cantwell, M. M.; Galway, K.; Johnston, B. T.; Murray, L. The Incidence of Esophageal Cancer and High-Grade Dysplasia in Barrett's Esophagus: A Systematic Review and Meta-Analysis. *Am. J. Epidemiol.* **2008**, *168* (3), 237−249.

(23) Selman, M. H. J.; Hemayatkar, M.; Deelder, A. M.; Wuhrer, M. Cotton HILIC SPE Microtips for Microscale Purification and Enrichment of Glycans and Glycopeptides. *Anal. Chem.* **2011**, *83* (7), 2492−2499.

(24) LaMarche, B.; Crowell, K.; Jaitly, N.; Petyuk, V.; Shah, A.; Polpitiya, A.; Sandoval, J.; Kiebel, G.; Monroe, M.; Callister, S.; Metz, T.; Anderson, G.; Smith, R. MultiAlign: a multiple LC-MS analysis tool for targeted omics analysis. *BMC Bioinf.* **2013**, *14* (1), 1−14.

(25) Jaitly, N.; Monroe, M. E.; Petyuk, V.; Clauss, T. R. W.; Adkins, J. N.; Smith, R. D. Robust Algorithm for Alignment of Liquid Chromatography-Mass Spectrometry Analyses in an Accurate Mass and Time Tag Data Analysis Pipeline. *Anal. Chem.* **2006**, *78*, 7397−7409.

(26) Wang, W.; Zhou, H.; Lin, H.; Roy, S.; Shaler, T. A.; Hill, L. R.; Norton, S.; Kumar, P.; Anderle, M.; Becker, C. H. Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Anal. Chem.* **2003**, *75* (18), 4818−4826.

(27) Finney, G. L.; Blackler, A. R.; Hoopmann, M. R.; Canterbury, J. D.; Wu, C. C.; MacCoss, M. J. Label-Free Comparative Analysis of Proteomics Mixtures Using Chromatographic Alignment of High-Resolution LC-MS Data. *Anal. Chem.* **2008**, *80* (4), 961−971.

(28) Bates, D.; Bolker, B.; Maechler, M.; Walker, S. *lme4: Linear Mixed-Effects Models Using Eigen and S4*, 2014. http://cran.r-project.org/web/packages/lme4/index.html.

(29) Noble, W. S. How does multiple testing correction work? *Nat. Biotechnol.* **2009**, *27* (12), 1135−1137.

(30) Raman, R.; Venkatraman, M.; Ramakrishnan, S.; Lang, W.; Raguram, S.; Sasisekharan, R. Advancing glycomics: Implementation strategies at the Consortium for Functional Glycomics. *Glycobiology* **2006**, *16* (5), 82R−90R.

(31) Ye, B.; Cramer, D. W.; Skates, S. J.; Gygi, S. P.; Pratomo, V.; Fu, L.; Horick, N. K.; Licklider, L. J.; Schorge, J. O.; Berkowitz, R. S.; Mok, S. C. Haptoglobin-Î± Subunit As Potential Serum Biomarker in Ovarian Cancer: Identification and Characterization Using Proteomic Profiling and Mass Spectrometry. *Clin. Cancer Res.* **2003**, *9* (8), 2904−2911.

(32) Lin, Z.; Simeone, D. M.; Anderson, M. A.; Brand, R. E.; Xie, X.; Shedden, K. A.; Ruffin, M. T.; Lubman, D. M. Mass Spectrometric Assay for Analysis of Haptoglobin Fucosylation in Pancreatic Cancer. *J. Proteome Res.* **2011**, *10* (5), 2602−2611.

(33) de Leoz, M. L. A.; Young, L. J. T.; An, H. J.; Kronewitter, S. R.; Kim, J.; Miyamoto, S.; Borowsky, A. D.; Chew, H. K.; Lebrilla, C. B. High-Mannose Glycans are Elevated during Breast Cancer Progression. *Mol. Cell. Proteomics* **2011**, *10*, 1.

(34) Rutkowski, M. J.; Sughrue, M. E.; Kane, A. J.; Mills, S. A.; Parsa, A. T. Cancer and the Complement Cascade. *Mol. Cancer Res.* **2010**, *8* (11), 1453−1465.
(35) Debruyne, E. N.; Vanderschaeghe, D.; Van Vlierberghe, H.; Vanhecke, A.; Callewaert, N.; Delanghe, J. R. Diagnostic Value of the Hemopexin N-Glycan Profile in Hepatocellular Carcinoma Patients. *Clin. Chem.* **2010**, *56* (5), 823−831.
(36) Knekt, P.; Aromaa, A.; Maatela, J.; Rissanen, A.; Hakama, M.; Aaran, R. K.; Nikkari, T.; Hakulinen, T.; Peto, R.; Teppo, L. Serum ceruloplasmin and the risk of cancer in Finland. *Br. J. Cancer* **1992**, *65* (2), 292−296.
(37) Senra Varela, A.; Bosco Lopez Saez, J. J.; Quintela Senra, D. Serum ceruloplasmin as a diagnostic marker of cancer. *Cancer Lett.* **1992**, *121* (2), 139−145.

dx.doi.org/10.1021/pr500242m | *J. Proteome Res.* XXXX, XXX, XXX−XXX