Published in final edited form as:
J Proteome Res. 2013 December 6; 12(12): 5609–5615. doi:10.1021/pr400575f.

Analysis of *N*-glycoproteins using Genomic *N*-glycosite Prediction

Shisheng Sun[†,#], **Bai Zhang**[†,#], **Paul Aiyetan**[†], **Jianying Zhou**[†], **Punit Shah**[†], **Weiming Yang**[†], **Douglas A. Levine**[‡], **Zhen Zhang**[†], **Daniel W. Chan**[†], and **Hui Zhang**[*,†]

[†]Department of Pathology, Johns Hopkins University, Baltimore, Maryland [‡]Memorial Sloan-Kettering Cancer Center, New York City, New York

Abstract

Protein glycosylation has long been recognized as one of the most common post-translational modifications. Most membrane proteins and extracellular proteins are N-linked glycosylated and they account for the majority of current clinical diagnostic markers or therapeutic targets. Quantitative proteomic analysis of detectable N-linked glycoproteins from cells or tissues using mass spectrometry has the potential to provide biological basis for disease development and identify disease associated glycoproteins. However, the information of low abundance but important peptides is lost due to the lack of MS/MS fragmentation or low quality of MS/MS spectra for low abundant peptides. Here, we show the feasibility of formerly *N*-glycopeptide identification and quantification at MS1 level using genomic *N*-glycosite prediction (GenoGlyco) coupled with stable isotopic labeling and accurate mass matching. The GenoGlyco Analyzer software uses accurate precursor masses of detected N-deglycopeptide peaks to match them to N-linked deglycopeptides which are predicted from genes expressed in the cells. This method results in more robust glycopeptide identification compared to MS/MS based identification. Our results showed that over three times the quantity of N-deglycopeptide assignments from the same mass spectrometry data could be produced in ovarian cancer cell lines compared to a MS/MS fragmentation method. Furthermore, the method was also applied to N-deglycopeptide analysis of ovarian tumors using the identified deglycopeptides from the two ovarian cell lines as heavy standards. We show that the described method has a great potential in the analysis of detectable *N*-glycoproteins from cells and tissues.

Keywords

glycosylation; prediction; genome-wide; SILAC; accurate mass matching; ovarian cancer; mass spectrometry

[*]Hui Zhang, Department of Pathology, Johns Hopkins University, 1550 Orleans Street, CRBII, Room 3M-03, Baltimore, MD 21231. Phone: (410) 502-8149; Fax: 443-287-6388 hzhang32@jhmi.edu.
[#]S.S. and B.Z contributed equally to this work.

Supporting Information Available: This material is available free of charge via the Internet at http://pubs.acs.org.

INTRODUCTION

N-linked glycoproteins play important roles in biological processes, including cell-to-cell recognition, growth, differentiation and programmed cell death, viral evolution and immune escape[1–5]. Specific N-linked glycoprotein changes are associated with disease progression and identification of these N-linked glycoproteins has a potential application in disease diagnosis, prognosis, and prediction of treatments[6,7].

Tandem mass spectrometry (MS/MS)-based shotgun proteomics technology combined with stable isotope dilution has become an effective approach for large-scale protein identification and quantification in complex biological or clinical samples[8–11]. A typical shotgun proteomic analysis consists of digestion of proteins to peptides and analyzing the peptides by liquid-chromatography-tandem mass spectrometry (LC-MS/MS). Prior to LC-MS/MS analysis, digested peptides are optionally labeled with isotopic or isobaric tags for peptide and protein quantitation[10,11]. Alternatively, proteins are metabolically labeled by stable-isotope labeling by amino acids in cell culture (SILAC) (typically lysine and/or arginine) before they are digested to peptides for mass spectromety analysis[12]. Each tandem spectrum is searched through the database against all possible peptide spectra with the same precursor mass and a peptide sequence based on the highest correlation of theoretical MS/MS spectrum to the acquired MS/MS spectrum. The assigned peptides are then quantified by the number of spectra assigned to each peptide or by using the isotope or isobaric tags for accurate quantification. The process is very convenient since many database search engines and software have been developed to allow the automated assignment of MS/MS spectra to the peptide sequences[13–15]. This workflow for the identification of specific peptide relies on two factors: 1) MS/MS spectra are generated from all peptides by mass spectrometer; and 2) MS/MS spectra are in high quality to generate enough information for correlation to the theoretical spectrum. However, some peptides have inherent sequences that do not generate high quality MS/MS, and precursor ions of low abundant peptides are either not selected for MS/MS acquisition or produce poor quality MS/MS spectrum. These factors have greater effect on the identification of protein modifications than the identification of the protein, as a protein can be identified by multiple peptides, while the modifications can be identified only when the modified peptide is identified. Meanwhile, LC-MS data contains a lot more information on the parent ions of peptides (accurate mass, retention time, abundance, etc.) and these have been employed for peptide identification[16,17]. However, in this approach, a database with a list of identified peptides by MS/MS is established prior to the peptide identification using LC-MS data (e.g. AMT tag database based on MS/MS identification) due to the high complexity of proteomes in biological and clinical samples.

Unlike global proteomics, N-linked glycoproteomics focuses on the known, well-defined N-linked glycopeptides[18,19]. The *N*-glycosylation of proteins occurs at N-X-S/T motif (where X is any amino acids except proline) and thus the potential *N*-glycopeptides from a genome or proteome can be predicted from genomic data by containing the N-linked motifs[20,21]. The number of *N*-glycopeptides in this case is significantly reduced compared to the tryptic peptides increasing the potential for identification based on their distinct masses.

The high throughput and accuracy of SILAC labeling has made it a widely used method in quantifying proteomes of cells, tissues and even blood[12,22–25]. The isotopic pairs and lysine and arginine count (K/R count) information containing in the mass difference between light and heavy SILAC peptides have the potential to enhance the identification accuracy of *N*-deglycopeptide based on accurate mass matching to predictive expressed peptides.

In this study, we describe a new method, named GenoGlyco, to show the feasibility of quantitative *N*-glycoproteome analysis based on genomic *N*-glycosite prediction. The peptides that are N-linked glycosylated in the native proteins were selectively isolated using solid-phase extraction of N-linked glycopeptides (SPEG) which is based on hydrazide chemistry[18,26], and the deglycosylated forms of these peptides from metabolic labeled cells were analyzed by liquid chromatography–mass spectrometry (LC-MS) and tandem mass spectrometry (MS/MS). The *N*-deglycopeptide pairs were identified by accurate mass matching of isolated peptides to the *N*-glycosites from the genome-wide prediction of the expressed glycoproteins in the cells (GenoGlyco). The SILAC peptide peak pairs containing K/R count information allows the accurate measurement of the deglycopeptide from potential deglycopeptides with the same mass. We showed that this method allowed the assignments of 1772 deglycopeptides detectable by LC-MS using accurate mass mapping of detected peptide peaks in SKOV-3 cell line. Among these, 501 deglycopeptides were verified by the traditional shotgun approach using MS/MS spectra and database search, and additional 184 deglycopeptides were verified using at least 3 fragment ions from MS/MS spectra. The SILAC labeled and assigned deglycopeptides were used as standards for the *N*-glycoprotemic analysis of ovarian cancer tissue.

MATERIALS AND METHODS

SILAC labeling of SKOV-3 and OVCAR-3 cells

Human ovarian carcinoma cell lines, SKOV-3 (ATCC, HTB77) and OVCAR-3 (ATCC, HTB161), were obtained from ATCC (Rockville, MD). Both cell lines were cultured in SILAC RPMI 1640 medium containing heavy isotope-labeled $^{13}C_6$-L-lysine and $^{13}C_6{}^{15}N_4$-L-arginine (Cambridge Isotope Laboratories, Andover, MA) supplemented with 10% dialyzed fetal bovine serum (diFBS) (Invitrogen, Carlsbad, CA) or in normal RPMI 1640 medium with 10% FBS to generate the SILAC-labeled and normal cell proteins, respectively. Cells were cultured for approximately 10 doublings in the SILAC medium to make sure complete labeling. After removing medium, the cells were washed 3 times with PBS buffer and then lysed by 8M urea/0.5% SDS/1M NH_4HCO_3 buffer directly. Lysates were briefly sonicated till the solutions were clear. Protein concentrations were determined by BCA protein assay reagent (Pierce, Rockford, IL).

Ovarian carcinoma tissue

An optimal cutting temperature-embedded (OCT) frozen high-grade serous ovarian cancer tissue was obtained from Memorial Sloan-Kettering Cancer Center, New York. The tissue was cut into small pieces and lysed by 8M urea/0.5% SDS/1M NH_4HCO_3 buffer as described for cell lines.

Formerly *N*-glycopeptides isolation

For each cell line, the *N*-deglycopeptides were isolated from both light cell proteins and protein mixtures of the same amount of light and heavy cell proteins. For tissue samples, the SILAC labeled cell proteins from SKOV3 and OVCAR-3 were first mixed together with same amount to prepare SILAC mix. The same amount of SILAC cell protein mix was added into each tissue samples as internal standards. Then N-deglycopeptides were isolated from whole cell extracts of both cells and tumor tissue sample directly by employing solid-phase extraction of *N*-glycopeptides method as described previously with minor modifications[26]. Briefly, One milligram of proteins were reduced by 5mM of DTT at 37°C for 1h and alkylated by 20mM iodoacetamide at room temperature in the dark for 30 min. After the solutions were diluted 8-fold with 0.1M NH_4HCO_3 buffer, 20μg of trypsin (Promega, Madison, WI) were added and incubated at 37°C overnight with shaking. Samples were centrifuged at 13,000g for 10min to remove any particulate matter and cleaned by C18 column (Milford, MA, Waters). Peptides were eluted by 400μL of 60% ACN/0.1% TFA. Glycopeptides were oxidized by 10mM $NaIO_4$ solution at room temperature for 1h in the dark; then the samples were cleaned by C18 column again. The eluted solution was collected into 25μL of equilibrated hydrazide beads (Bio-Rad, Richmond, CA) directly and incubated with 100mM aniline at room temperature for 3h[27]. The beads were washed 3 times each with 50%ACN, 1.5M NaCl, water and PBS buffer. *N*-deglycopeptides were released via 3μL PNGase F (New England Biolab, Beverly, MA) in PBS buffer at 37°C overnight with shaking. *N*-deglycopeptides were collected in supernatants and wash solutions and cleaned by C18 column. The *N*-deglycopeptides were dried via SpeedVac and resuspended in 20μL 0.1% TFA solution for mass spectrometry analysis.

Mass spectrometry analysis

All deglycopeptide samples from light cells, mixed cells (H:L=1:1) and tissues (H:L=1:1) underwent 3 replications of LC-MS/MS analysis for one biological and technical replicate of each sample with ~1μg (2μL) per run. Peptides were separated on a Dionex Ultimate 3000 RSLC nano system (Thermo Scientific, Bremen, Germany) with a 75μm × 15cm Acclaim PepMap100 separating column (Thermo Scientific) protected by a 2cm guarding column (Thermo Scientific). Mobile phase flow rate was 300 nL/min and consisted of 0.1% formic acid in water (A) and 0.1% formic acid 95% acetonitrile (B). The gradient profile was set as follows: 4–35% B for 70 min, 35–95% B for 5 min, 95% B for 10 min and equilibrated in 4% B for 15 min. MS analysis was performed using an Orbitrap Velos Pro mass spectrometer (Thermo Scientific). The spray voltage was set at 2.2 kV. Orbitrap spectra (AGC 1×106) were collected from 400–1800 m/z at a resolution of 60K followed by data-dependent HCD MS/MS (at a resolution of 7500, collision energy 45%, activation time 0.1ms) of the ten most abundant ions using an isolation width of 2.0Da. Charge state screening was enabled to reject unassigned and singly charged ions. A dynamic exclusion time of 35s was used to discriminate against previously selected ions.

Database search

All LC-MS/MS data was searched against corresponding cell expression databases by both Sequest (proteome discoverer) and MaxQuant (v1.3.0.5). The search parameters for Sequest were set as follows: up to two missed cleavage were allowed for trypsin digestion, 10ppm precursor mass tolerance and 0.06Da fragment mass tolerance; Carbamidomethylation (C) was set as a static modification, while oxidation (M), deamination (N), $^{13}C_6$ label of lysine (K) and $^{13}C_6{}^{15}N_4$ arginine (R) were set as dynamic modifications; SILAC 2plex (Arg10, Lys6) was selected as the quantification method and the results were filtered with 1% FDR. For MaxQuant, the search parameters were as follows: two missed cleavages were allowed for trypsin digestion. Carbamidomethylation (C) was set as static modification, oxidation (M) and deamination (N) were set as dynamic modifications. Two multiplicity with “Arg 10” and “Lys6” were selected as heavy labels. 1% FDR was employed for peptide identification and all peptides were used for protein quantification. All other settings were set as default values.

Identifying formerly *N*-glycopeptides by GenoGlyco method

GenoGlyco Analyzer is an algorithm that matches mass spectrometric peak masses to those predicted deglycopeptide masses from a sample. GenoGlyco Analyzer predicts potential N-deglycopeptides with consensus N-linked glycosylation sequons (N-X-S/T, where X is not P) in a sample and matches the experimental detected peaks (or peak pairs) by mass spectrometer to the predicted deglycopeptides using accurate mass matching, and the K/R count information based on the mass difference if the peptides are derived from SILAC labeled samples. Potential *N*-deglycopeptides of cells were predicted from the cell expression database[28] by searching N-X-S/T (X can’t be proline) motifs (supporting information). The SILAC peak pair information of *N*-deglycopeptides of both cell lines was extracted from raw MS data into ‘allpeptides.txt’ file when searching database using MaxQuant as described above. The SILAC peak pairs were matched to the cell line specific *N*-deglycopeptide based on the monoisotopic mass of the light peptides with 10ppm mass error tolerance and then the K/R count of the peak pairs was employed to filter the results. The matched unique or multiple peptides to each SILAC peak pair with the same mass and K/R count were considered as the *N*-deglycopeptides or *N*-deglycopeptide candidates of the peak pair. The replicates of deglycopeptides resulting from different runs or different charge states of the same peptide within 2min retention time range were then removed. The script for this analysis was written and ran in the R programming environment.

RESULTS AND DISCUSSION

GenoGlyco method for N-linked glycoproteomics

Here we describe the GenoGlyco method for quantitative *N*-glycoproteome analysis based on genome-wide glycosite prediction coupled with accurate mass measurement of SILAC-labeled deglycopeptide pairs (Figure 1). The GenoGlyco method in this paper includes the following steps: 1) *N*-deglycopeptides were specifically isolated from light and heavy SILAC labeled cell proteins by the solid-phase extraction of N-linked glycopeptides (SPEG) based on hydrazide chemistry. 2) The accurate mass of the mixture of light and heavy labeled deglycopeptides was measured by LC-MS. 3) Potential *N*-deglycopeptides from the

SILAC labeled cell lines were predicted using the *N*-glycosylation motif and the gene expression data of the cell. 4) The deglycopeptide peak pairs were matched to predicted deglycopeptides of the cell line using accurate mass matching, and the K/R count in the peptides derived from the mass difference of SILAC pairs. We verified the assigned *N*-deglycopeptides by MS/MS. We showed that the heavy SILAC labeled *N*-deglycopeptides identified using this method could be used for the analysis of ovarian cancer tissues.

Prediction of potential *N*-glycopeptides from a cell line using *N*-glycosylation motif and the gene expression data

First, we determined the theoretical feasibility of the GenoGlyco method by employing the SKOV3 ovarian cancer cell line as a model. We created a SKOV3 cell expressed protein database based on mRNA microarray data from the cell-miner repository developed by the NCI genomics and bioinformatics group[28] and human IPI 3.87 database[29], and then predicted all potential *N*-glycopeptides using the gene expression data from the cells. Compared to the tryptic peptides in the entire IPI database (3,565,543 unique tryptic peptides with up to 2 missed cleavage sites), only ~4.5% of the tryptic peptides (159,028 peptides) were potential *N*-glycopeptides predicted by consensus N-X-T/S motif (X denotes any amino acids except proline)[20] and gene expression data in SKOV3 cells (14,652 genes)[28]. Among 159,028 potential N-linked glycopeptides containing the consensus N-X-T/S motif from peptides from SKOV3 cells, about one third (~50,000 with up to 2 missed cleavage sites) are from transmembrane proteins, cell surface proteins, or secreted proteins that were potentially *N*-glycosylated[21]. These results showed that the targeted analysis of *N*-deglycopeptides dramatically reduced the complexity of the peptides over the total tryptic peptides in the protein database (Table 1).

The reduced complexity of potential *N*-deglycopeptides reduced the mass overlap of peptides and allowed the separation of these deglycopeptides by their mass. We calculated the number of unique deglycopeptides to each mass. With a high mass accuracy of 1ppm, which is achievable using the current mass spectrometer, over 65% of potential *N*-deglycopeptides (104,748) contained distinct masses. With a decrease of instrument mass accuracy (10ppm), the percentage of unique *N*-deglycopeptides with distinct masses was reduced to 13.6% (21,575) of all potential *N*-deglycopeptides. In addition, due to the different mass shifts introduced by SILAC labeling for the K and R containing peptides, the mass difference of the paired light and heavy labeled peptides could be used to determine the number of K or R in the peptides. The K/R count information contained in SILAC peak pairs makes 91.4% of *N*-deglycopeptides (145,384) unique within 1ppm mass error and 65% of *N*-deglycopeptides (103,357) unique within 10ppm mass error (Figure 2A). The reduced number of potential *N*-deglycopeptides and the K/R count information in SILAC peak pairs dramatically increased the unique *N*-deglycopeptides for each distinct mass detected by mass matching using LC-MS spectra.

Identification of the SILAC-labeled *N*-deglycopeptide pairs

Next, we assessed the feasibility of the GenoGlyco method by applying it towards the identification of *N*-deglycopeptides from SKOV3 cells. SKOV3 cells were labeled by light and heavy SILAC. Equal amounts of proteins from labeled cells were mixed and the *N*-

deglycopeptides were isolated by solid-phase extraction of *N*-glycopeptides (SPEG) and analyzed by triplicate LC-MS using orbitrap velos mass spectrometer. A total of 8621 peak pairs (average 2874 per run) were detected from SKOV3 cells with a mass accuracy of 10ppm. Out of these peak pairs, 3218 pairs only contained K (mass difference of 6 Da, 12 Da or 18 Da), 3439 pairs only contained R (mass difference of 10 Da, 20 Da or 30 Da) and 1964 pairs contained both K and R (mass difference of 16 Da, 22 Da or 26 Da). After matching the light mass of all peak pairs with potential *N*-deglycopeptides of the proteins expressed in SKOV3 cells using K/R count information with10ppm mass error, we were able to assign 1,772 unique peak pairs with a specific mass and retention time to *N*-deglycopeptides, 64.3% (1,139) peak pairs matched to unique *N*-deglycopeptide mass, and 97.3% peak pairs (1,725) matched to no more than three deglycopeptides (Figure 2B and Table S1). The results showed that most of the *N*-deglycopeptides (~2/3 of identified deglycopeptides) from a cell line could be distinguished and assigned by their mass alone within 10ppm mass error. These are consistent with results from the predicted *N*-deglycopeptides. The separation will be increased when mass accuracy reaches 1ppm (Figure 2A).

Verification of identified *N*-deglycopeptides by tandem spectra

To determine the specificity of the identified *N*-deglycopeptides using the GenoGlyco method, we evaluated the assigned deglycopeptides using the MS/MS spectra collected from the same raw files. The MS/MS spectra were searched against the SKOV3 cell expressed protein database using Sequest[13] with up to 2 missed tryptic cleavage sites and 10ppm mass error. The assigned peptide sequences using MS/MS spectra (5565 spectra with 1% FDR rate) showed that 93.2% (561 unique peptides) of assigned unique sequences contained consensus N-linked glycosylation motif, showing high specificity of the isolated deglycopeptides from ovarian cancer cells (Figure 3A and Table S2). Among the deglycopeptides identified by MS/MS spectra, 503 of them (89.7%) were included in the 1,772 matched deglycopeptides and could be used to verify the results of deglycopeptide assignment. (Figure 3B and Table S1).

We further investigated the 1269 peak pairs that were assigned by GenoGlyco, but not identified by MS/MS spectra and the database search (Figure 3B). Examination of the MS and MS/MS spectra of these peak pairs showed three reasons for the failure of identification of observed MS peaks by MS/MS and database search using MS/MS spectra. *First, a Lack of MS/MS spectra due to the complexity of the samples*. The majority of peak pairs (781) were not identified due to the lack of MS/MS acquisition for the peak pairs. Analysis of these peptides by additional three additional LC-MS/MS spectra of *N*-deglycopeptides allowed the identification of an additional 103 peaks (Figure 3B and Table S1). *Second, low number of fragment ions from the selected peptides*. The remaining 488 peak pairs were selected by data-dependent acquisition for tandem mass spectrometry. However, the quality of the MS/MS spectra was not informative enough to allow the assignments of the peptide sequences from these MS/MS spectra using the database search algorithm. Among the 488 peak pairs with MS/MS spectra, we identified 81 peak pairs containing at least 3 fragment ions correctly assigned to the fragment ions from the predicted *N*-deglycopeptide sequences identified by GenoGlyco (Figure 3B). These peptide sequence–dependent MS/MS spectra may cause the lack of identification of these peptides. Additional evidence for the failure of peptide identification from these peptides caused by their sequence-dependent fragmentation pattern came from the analysis of these peptides by 2D-LC-MS/MS. The data showed that similar spectra were generated from these peptides by increased level of these peptides, but the assignment of MS/MS spectra to peptide sequence by database search algorithm still failed. *Third, low quality of MS/MS spectrum caused by high background fragment ions*. These spectra (407 peak pairs) were mostly generated from low abundant MS peaks with high background MS/MS ions. The quality of MS/MS spectra are known to related to be the abundance of peptides, therefore, low abundant deglycopeptides can only detected by MS, but not being able to generate quality MS/MS spectra for sequence assignment. Our results showed that assigning *N*-deglycopeptide sequences by GenoGlyco and mass spectrometric analysis of *N*-deglycopeptides reduces the obstacles of traditional shotgun proteomics and allows the assignment of detectable *N*-deglycopeptides from samples.

Examination of precursor ions of 58 MS/MS only deglycopeptides showed that 19 identified *N*-deglycopeptides contained other modifications (deamidation), two contained more than three K/R and one contained both light lysine and heavy arginine (false identification). Sixteen *N*-deglycopeptides peak pairs didn’t contain K/R count information (isotopic type) in MaxQuant results and 13 *N*-deglycopeptides were not detected as pairs, while another 6 peak pairs contained different K/R count from the identified ones (Table S3). The undetected peak pairs and the deviation of H/L ratios of many deglycopeptides (Table S1, S2 and S3) in the study might be caused by the different components in normal and SILAC medium as well as normal and dialyzed FBS used for cell culture in this study. This situation should be improved by using same medium and dialyzed FBS with normal and heavy isotope labeled amino acids for culturing light and heavy SILAC cells, respectively.

Reducing the complexity by capturing *N*-deglycopeptides is a prerequisite for the identification accuracy of the method. The database of potential deglycopeptides expressed from cell lines contains only 4.5% of unique peptides in the entire human database. SILAC labeling and detecting peptides by peak pairs in LC-MS help the detection of these peptides. Additional peaks present in MS spectra are due to chemical noise and other contamination, but metabolic labeling of cell glycopeptides with SILAC peak pairs can exclude these contaminants. Moreover, the MS/MS based identification indicated that all methionine residues in identified *N*-deglycopeptides isolated by SPEG method were in the oxidized state, which might due to the oxidation of sodium periodate[30]. All these reduced the complexity of deglycopeptide samples and thus enhanced the identification accuracy of deglycopeptides at the MS1 level. The specificity of the method can be further enhanced by an additional 2 mass units by incorporating ^{18}O mass tags on *N*-glycosites of formerly heavy *N*-glycopeptides during the removal of *N*-glycans by PNGase F. The ^{18}O mass tag can provide the glycosite count information in SILAC peak pairs and distinguish the deglycopeptide from non-glycopeptide peak pairs.

In addition, our theoretical data showed that the GenoGlyco method also works when the potential deglycopeptide database was generated from the human entire protein database (RefSeq as July 29th, 2013[31]) instead of the cell expressed database and deamidation was set as variable modification from additional potential *N*-glycosites if a peptide contains more than one N-X-S/T motifs. 87.7% (298806) and 49.5% (168,630) of potential *N*-deglycopeptides contained distinct masses within 1ppm and 10ppm, respectively (Figure S1).

Analysis of ovarian cancer tissue using SILAC labeled *N*-deglycopeptides from multiple ovarian cancer cell lines

The identified *N*-deglycopeptides from SILAC labeled cells could be used for the analysis of *N*-deglycopeptides from a variety of biological or clinical samples such as cells, tissues, and body fluids (Figure 1). In this study, we used the SILAC labeled *N*-deglycopeptides from ovarian cancer cells as universal standards for the analysis of ovarian tumor.

To increase the coverage of SILAC labeled *N*-deglycopeptides as standards for ovarian tissue analysis, we applied the same procedure to the analysis of another ovarian cancer cell line – OVCAR-3 cell line. Theoretically, the potential *N*-deglycopeptides from OVCAR-3 cells (162,177 potential *N*-deglycopeptides with 2 missed cleavage sites) also contain 4.5% of tryptic peptides from human IPI database (Table 1). The unique *N*-deglycopeptides with distinct peptide mass from potential *N*-deglycopeptides increased from 13.4% by mass matching only to 64.6% by both mass and K/R count information with mass tolerance of 10ppm (Figure S2). From the LC-MS analysis of SILAC labeled OVCAR-3 cells, we found 60.4% (1055) peak pairs were assigned to unique *N*-deglycopeptides when coupled with K/R information (Figure S3, S4-A). Among 1748 peak pairs assigned to *N*-deglycopeptides using SILAC peak pairs, 436 of them were also identified by Sequest at 1% FDR (Table S4, Figure S4-B and Table S5). Another 133 peak pairs were verified by MS/MS spectra with at least 3 fragment ions and 168 peak pairs were verified by three additional LC-MS/MS spectra of light *N*-deglycopeptides (Figure 4B and Table S4). These results indicated that assigning SILAC peak pairs of *N*-deglycopeptides to unique *N*-deglycopeptide sequences should be applicable to most cell lines.

We then pooled the heavy labeled *N*-deglycopeptides extracted from both SILAC labeled ovarian cancer cell lines and made SILAC *N*-deglycopeptide mix as heavy isotope labeled standards with 4075 non-redundant assigned *N*-deglycopeptides from the two cell lines. The mix of SILAC labeled *N*-deglycopeptides from both cells increases the number of *N*-deglycopeptide standards for analysis of ovarian tumors (Figure 4A). The SILAC-labeled *N*-deglycopeptide standards were used to analyze ovarian tumor.

Tumor tissues are often stored frozen embedded in the optimal cutting temperature (OCT) medium. We applied the SILAC labeled *N*-deglycopeptide standards to analyze the OCT embedded ovarian cancer tissues. The SILAC deglycopeptide standards were added to *N*-deglycopeptides isolated from ovarian tumor stored frozen with OCT embedding. All detected peak pairs were matched with the assigned SILAC *N*-deglycopeptides based on mass and the K/R count of the deglycopeptides. We were able to identify 835 *N*-deglycopeptides in OCT-embedded tumor tissues (Figure 4B, Table S6). 326 assigned peak pairs were validated via MS/MS based identification (Figure 4B and Table S6–7). Additionally 509 *N*-glycosites were identified by SILAC peak pair mapping (Figure 4B). Furthermore, MS/MS based method identified an additional 396 *N*-glycosites: 319 of them were only identified in tissues and most of them did not originate from ovarian cancer cells but rather from plasma glycoproteins (Figure 4B and Table S6–7). These results demonstrated that hundreds of *N*-deglycopeptides from tumor tissues could also be analyzed by LC-MS-MS/MS using the GenoGlyco method.

CONCLUSIONS

We described the theoretical feasibility of *N*-glycoprotein analysis using genome-wide mapping of *N*-glycosites and tested the feasibility by mapping the deglycopeptide peak pairs from SILAC labeled ovarian cancer cell lines to identify *N*-glycosites using both accurate mass and K/R count information. By employing the GenoGlyco Analyzer software, which identifies the N-deglycopeptides by accurate mass matching of precursor masses to that of predicted N-deglycopeptides, we identified more than three times the quantity of N-deglycopeptides in two ovarian cancer cell lines compared to a MS/MS fragmentation based identification method. The identification specificity and accuracy of this method can theoretically be further increased by incorporating ^{18}O on *N*-glycosites of *N*-deglycopeptide and increasing mass accuracy of SILAC labeled lysine/arginine peptides. This method has a potential to identify all detectable *N*-deglycopeptides and *N*-glycoproteins from a cell line or even tissue and it promises to be helpful in biomarker discovery. By applying super SILAC labeled *N*-deglycopeptides for each clinical specimen, this method can be used to identify and quantify thousands of *N*-deglycopeptides from clinical specimens of a small size (1 mg). This is especially useful for the proteomic analysis of some important specimens with limited available materials. The candidate glycoproteins identified from clinical samples using this method can be further validated by independent methods, such as applying a MudPIT for peptide separation and MS/MS identification, targeted MS/MS analysis, MRM experiment and western blot.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work is partially supported by National Institutes of Health, National Cancer Institute, Clinical Proteomic Tumor Analysis Consortium (U24CA160036), The Early Detection Research Network (EDRN, U01CA152813), R01CA112314, and Chia Family Foundation.

REFERENCES

1. Rudd PM, Elliott T, Cresswell P, Wilson IA, Dwek RA. Glycosylation and the Immune System. Science. 2001; 291:2370–2376. [PubMed: 11269318]
2. Lis H, Sharon N. Protein glycosylation. Euro. J. Biochem. 1993; 218:1–27.
3. Tian Y, Zhang H. Glycoproteomics and clinical applications. Proteom. Clin. Appl. 2010; 4:124–132.
4. Haltiwanger RS, Lowe JB. Role of glycosylation in develooment. Annu. Rev. Biochem. 2004; 73:491–537. [PubMed: 15189151]
5. Sun S, Wang Q, Zhao F, Chen W, Li Z. Glycosylation site alteration in the evolution of influenza A (H1N1) viruses. 2011; 6:e22844.
6. Ohtsubo K, Marth JD. Glycosylation in Cellular Mechanisms of Health and Disease. Cell. 2006; 126:855–867. [PubMed: 16959566]

7. Durand G, Seta N. Protein Glycosylation and Diseases: Blood and Urinary Oligosaccharides as Markers for Diagnosis and Therapeutic Monitoring. Clin. Chem. 2000; 46:795–805. [PubMed: 10839767]

8. Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature. 2003; 422:198–207. [PubMed: 12634793]

9. Link AJ, Eng J, Schieltz DM, Carmack E, Mize GJ, Morris DR, Garvik BM, Yates JR. Direct analysis of protein complexes using mass spectrometry. Nat. Biotech. 1999; 17:676–682.

10. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat. Biotech. 1999; 17:994–999.

11. Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, Khainovski N, Pillai S, Dey S, Daniels S, Purkayastha S, Juhasz P, Martin S, Bartlet-Jones M, He F, Jacobson A, Pappin DJ. Multiplexed Protein Quantitation in Saccharomyces cerevisiae Using Amine-reactive Isobaric Tagging Reagents. Mol. Cell. Proteomics. 2004; 3:1154–1169. [PubMed: 15385600]

12. Ong S-E, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M. Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics. Mol. Cell. Proteomics. 2002; 1:376–386. [PubMed: 12118079]

13. Eng J, McCormack A, Yates J. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am Soc. Mass Spectrom. 1994; 5:976–989. [PubMed: 24226387]

14. Li, X-j; Zhang, H.; Ranish, JA.; Aebersold, R. Automated Statistical Analysis of Protein Abundance Ratios from Data Generated by Stable-Isotope Dilution and Tandem Mass Spectrometry. Anal. Chem. 2003; 75:6648–6657. [PubMed: 14640741]

15. Perkins DN, Pappin DJC, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis. 1999; 20:3551–3567. [PubMed: 10612281]

16. Smith RD, Anderson GA, Lipton MS, Pasa-Tolic L, Shen Y, Conrads TP, Veenstra TD, Udseth HR. An accurate mass tag strategy for quantitative and high-throughput proteome measurements. Proteomics. 2002; 2:513–523. [PubMed: 11987125]

17. Zimmer JSD, Monroe ME, Qian W-J, Smith RD. Advances in proteomics data analysis and display using an accurate mass and time tag approach. Mass Spectrom. Rev. 2006; 25:450–482. [PubMed: 16429408]

18. Zhang H, Li X-j, Martin DB, Aebersold R. Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry. Nat. biotech. 2003; 21:660–666.

19. Kaji H, Saito H, Yamauchi Y, Shinkawa T, Taoka M, Hirabayashi J, Kasai K-i, Takahashi N, Isobe T. Lectin affinity capture, isotope-coded tagging and mass spectrometry to identify N-linked glycoproteins. Nat. Biotech. 2003; 21:667–672.

20. Bause E. Structural requirements of N-glycosylation of proteins. Studies with proline peptides as conformational probes. Biochem. J. 1983; 209:331–336. [PubMed: 6847620]

21. Zhang H, Loriaux P, Eng J, Campbell D, Keller A, Moss P, Bonneau R, Zhang N, Zhou Y, Wollscheid B, Cooke K, Yi EC, Lee H, Peskind ER, Zhang J, D Smith R, Aebersold R. UniPep - a database for human N-linked glycosites: a resource for biomarker discovery. Genome Biol. 2006; 7

22. Ong S-E, Mittler G, Mann M. Identifying and quantifying in vivo methylation sites by heavy methyl SILAC. Nat. Meth. 2004; 1:119–126.

23. Geiger T, Cox J, Ostasiewicz P, Wisniewski JR, Mann M. Super-SILAC mix for quantitative proteomics of human tumor tissue. Nat. Meth. 2010; 7:383–385.

24. Monetti M, Nagaraj N, Sharma K, Mann M. Large-scale phosphosite quantification in tissues by a spike-in SILAC method. Nat. Meth. 2011; 8:655–658.

25. Boersema PJ, Geiger T, Wiśniewski JR, Mann M. Quantification of the N-glycosylated Secretome by Super-SILAC During Breast Cancer Progression and in Human Blood Samples. Mol. Cell. Proteomics. 2013; 12:158–171. [PubMed: 23090970]

26. Tian Y, Zhou Y, Elliott S, Aebersold R, Zhang H. Solid-phase extraction of N-linked glycopeptides. Nat. Protoc. 2007; 2:334–339. [PubMed: 17406594]

27. Zeng Y, Ramya TNC, Dirksen A, Dawson PE, Paulson JC. High-efficiency labeling of sialylated glycoproteins on living cells. Nat. Meth. 2009; 6:207–209.

28. Shankavaram UT, Reinhold WC, Nishizuka S, Major S, Morita D, Chary KK, Reimers MA, Scherf U, Kahn A, Dolginow D, Cossman J, Kaldjian EP, Scudiero DA, Petricoin E, Liotta L, Lee JK, Weinstein JN. Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study. Mol. Cancer Ther. 2007; 6:820–832. [PubMed: 17339364]

29. Kersey PJ, Duarte J, Williams A, Karavidopoulou Y, Birney E, Apweiler R. The International Protein Index: An integrated database for proteomics experiments. Proteomics. 2004; 4:1985–1988. [PubMed: 15221759]

30. Yamasaki RB, Osuga DT, Feeney RE. Periodate oxidation of methionine in proteins. Anal. Biochem. 1982; 126:183–189. [PubMed: 6295208]

31. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. 2007; 35:D61–D65.

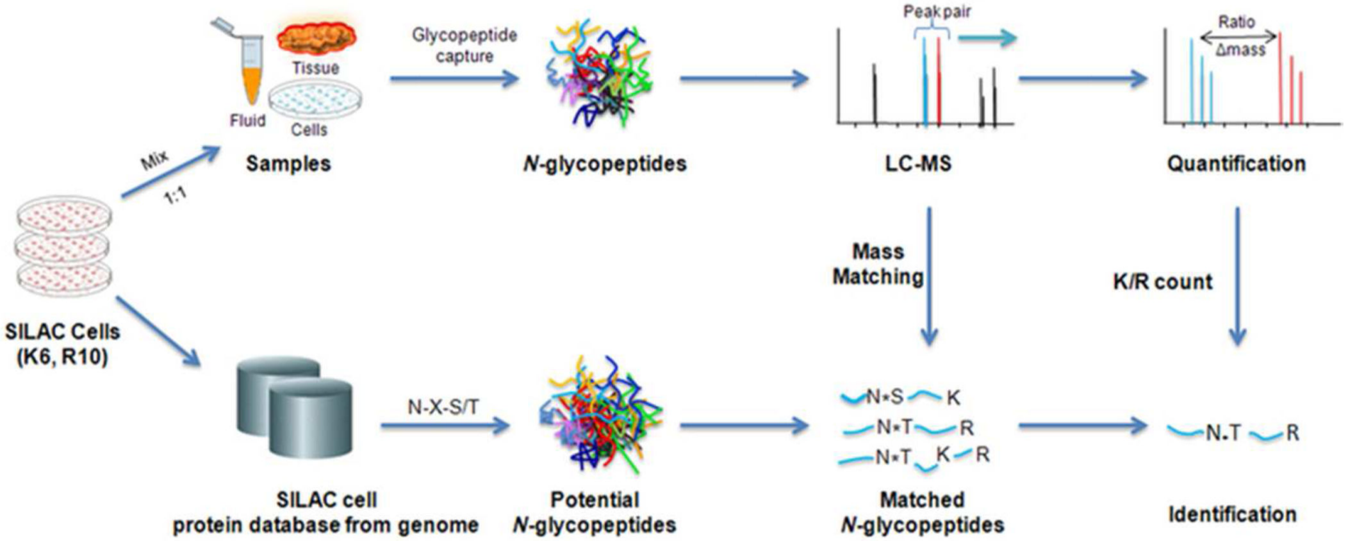

Figure 1.
Workflow of the genomic *N*-glycosite prediction (GenoGlyco) method for glycoproteomic analysis.

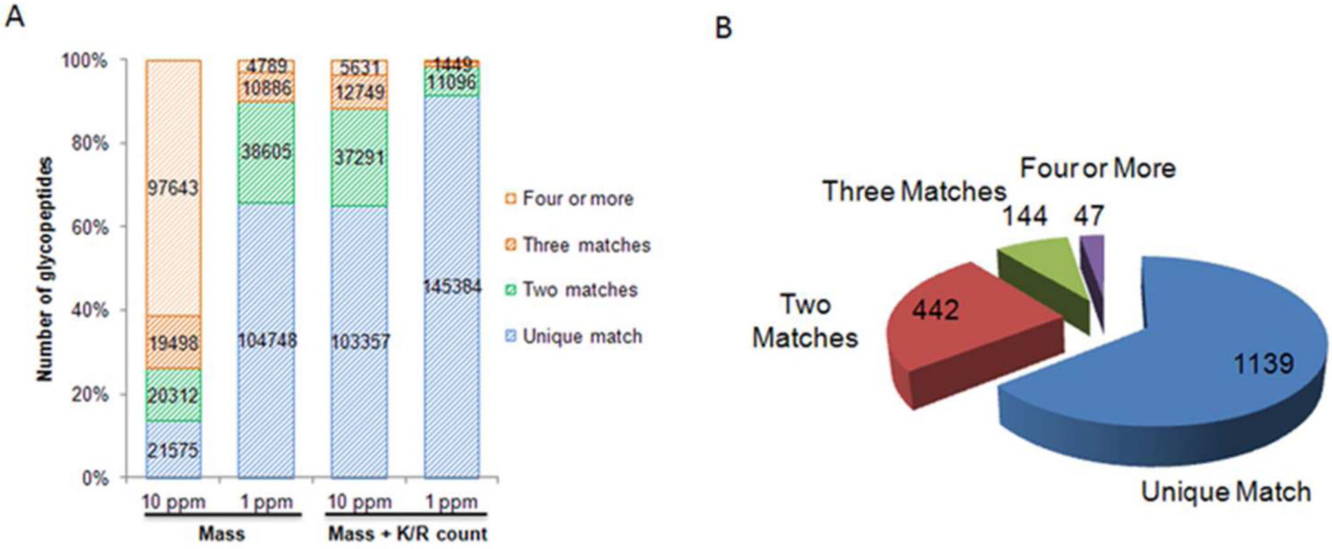

Figure 2.

The feasibility of the glycoprotein identification in SKOV-3 cell line using the GenoGlyco method. (A) Theoretical analysis of potential *N*-deglycopeptides with distinct mass with or without K/R count information in SKOV-3 cell line. (B) The analysis of *N*-deglycopeptides isolated from SKOV-3 cells using the GenoGlyco method.

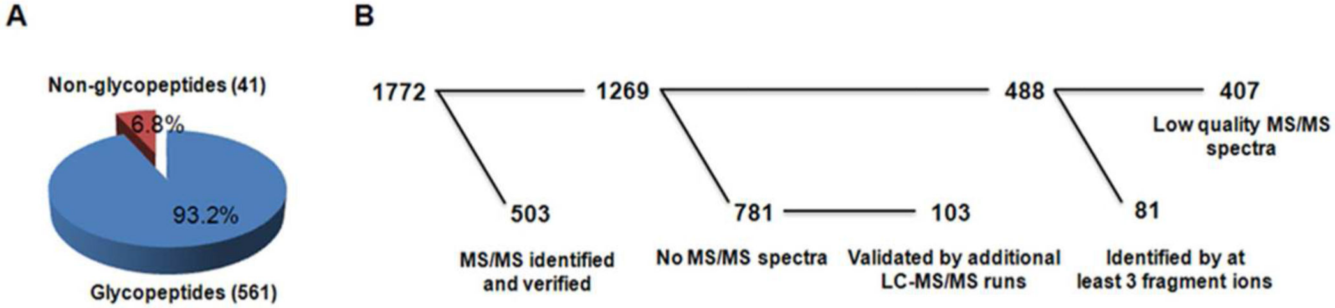

Figure 3.

Analysis of *N*-deglycopeptides identified in SKOV-3 cell line by GenoGlyco and MS/MS based identification. (A) The specificity of the isolated *N*-deglycopeptides identified by LC-MS/MS and database search. (B) Examination of identified 1,772 *N*-glycosites by GenoGlyco. 503 *N*-glycosites were verified by MS/MS identification, of the remaining 1,269*N*-glycosites, 781 were not subjected for data dependent MS/MS analysis during the same LC-MS-MS/MS analysis, but additional 3 LC-MS/MS runs of the light *N*-deglycopeptide samples identified and verified 103 of these. MS/MS spectra were generated for the remaining 488 SILAC labeled deglycopeptide pairs. Of these, 81 MS/MS spectra contained at least 3 fragment ions matched to theoretical fragment ions using targeted spectrum investigation, and the remaining 407 deglycopeptides were of low quality MS/MS spectra.

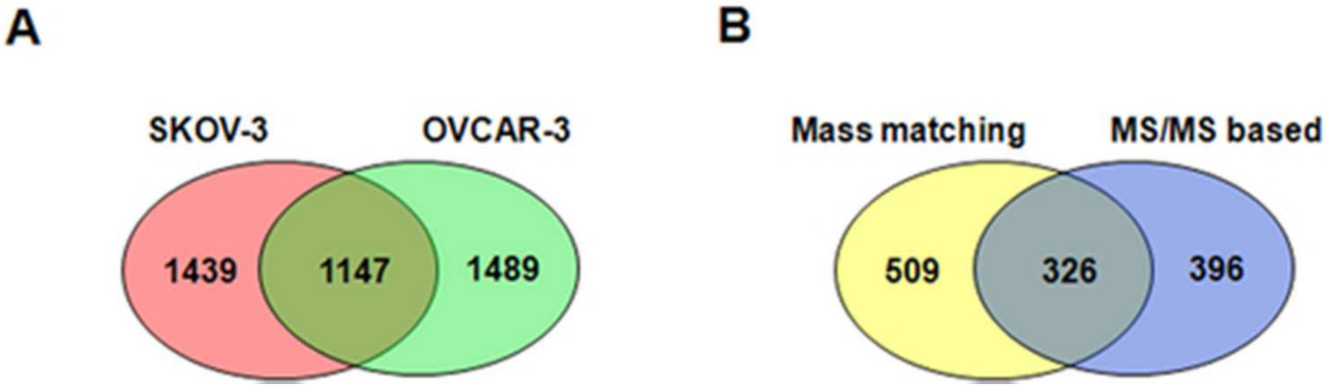

Figure 4.

Identification of *N*-deglycopeptides in ovarian cancer tissue by GenoGlyco using SILAC labeled deglycopeptides from multiple ovarian cancer cells and MS/MS based identification. A) Identification of deglycopeptides from both SKOV-3 and OVCAR-3 cells. B) Identification of *N*-deglycopeptides by GenoGlyco method and MS/MS based method.

Table 1

Number of proteins and potential *N*-deglycopeptides expressed in SKOV3 and OVCAR-3 cell lines.

Database	**Proteins**	**Peptides**[b]	**Potential *N*-deglycopeptides**[c]
HUMAN IPI3.87	91,491	3,565,543	360,665
Human RefSeq	36,430	2,884,481	298,511
SKOV3 cell expressed proteins	14,652	1,446,496	159,028
OVCAR3 cell expressed proteins	14,888	1,471,285	162,177

[a] Human IPI3.87 protein database were downloaded from IPI database[29] (total 91,491 protein entries). Human Refseq protein database was downloaded from NCBI website[31] as July 29th, 2013. The expressed protein database for SKOV-3 and OVCAR-3 cell lines were derived from RNA expression data[28]

[b] Number of tryptic peptides with up to 2 missed cleavage sites.

[c] Number of tryptic peptide containing N-X-S/T motif (X is any amino acids except Pro) with up to 2 missed cleavage sites.