Published in final edited form as:

Anal Chem. 2018 July 03; 90(13): 8065–8071. doi:10.1021/acs.analchem.8b01137.

Reanalysis of Global Proteomic and Phosphoproteomic Data Identified a Large Number of Glycopeptides

Yingwei Hu, **Punit Shah**, **David J. Clark**, **Minghui Ao**, and **Hui Zhang**[*]

Department of Pathology, Johns Hopkins University, Baltimore, Maryland 21287, United States

Abstract

Protein glycosylation plays fundamental roles in many cellular processes, and previous reports have shown dysregulation to be associated with several human diseases, including diabetes, cancer, and neurodegenerative disorders. Despite the vital role of glycosylation for proper protein function, the analysis of glycoproteins has been lagged behind to other protein modifications. In this study, we describe the reanalysis of global proteomic data from breast cancer xenograft tissues using recently developed software package GPQuest 2.0, revealing a large number of previously unidentified N- linked glycopeptides. More importantly, we found that using immobilized metal affinity chromatography (IMAC) technology for the enrichment of phosphopeptides had coenriched a substantial number of sialoglycopeptides, allowing for a large-scale analysis of sialoglycopeptides in conjunction with the analysis of phosphopeptides. Collectively, combined tandem mass spectrometry (MS/MS) analyses of global proteomic and phosphoproteomic data sets resulted in the identification of 6 724 N-linked glycopeptides from 617 glycoproteins derived from two breast cancer xenograft tissues. Next, we utilized GPQuest 2.0 for the reanalysis of global and phosphoproteomic data generated from 108 human breast cancer tissues that were previously analyzed by Clinical Proteomic Analysis Consortium (CPTAC). Reanalysis of the CPTAC data set resulted in the identification of 2 683 glycopeptides from the global proteomic data set and 4 554 glycopeptides from phosphoproteomic data set, respectively. Together, 11 292 N-linked glycopeptides corresponding to 1 731 N-linked glycosites from 883 human glycoproteins were identified from the two data sets. This analysis revealed an extensive number of glycopeptides hidden in the global and enriched in IMAC-based phosphopeptide-enriched proteomic data, information which would have remained unknown from the original study otherwise. The reanalysis described herein can be readily applied to identify glycopeptides from already existing data sets, providing insight into many important facets of protein glycosylation in different biological, physiological, and pathological processes.

[*]Corresponding Author: huizhang@jhu.edu.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.anal-chem.8b01137.

Captions of the supporting tables (PDF)

Summary of search results of Breast Cancer Xenograft samples (XLSX)

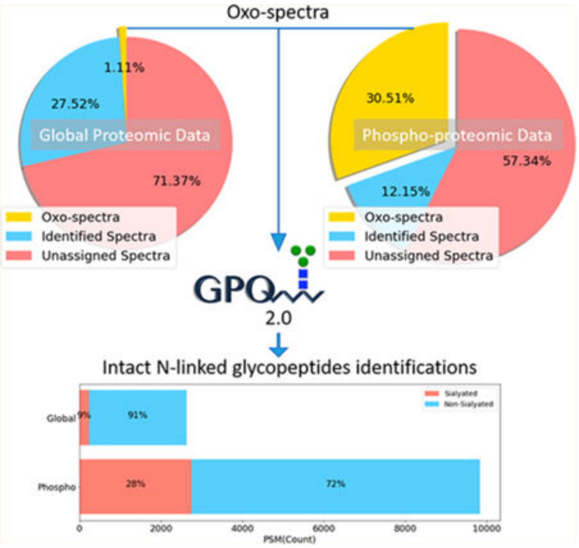
Search results (glycopeptide-spectrum matchings, GPSMs) of Breast Cancer Xenograft samples (XLSX) Summary of search results of Breast Cancer CPTAC samples (XLSX)

Search results (glycopeptide-spectrum matchings, GPSMs) of Breast Cancer CPTAC samples (ZIP)

Notes

The authors declare no competing financial interest.

Graphical Abstract



Protein glycosylation is one of the most common protein modifications universally present in all living organisms.[1] Glycosylation has diverse biological functions and plays critical roles in various biological activities, especially in cell recognition and adhesion.[1–7] With the development of high-throughput mass spectrometry-based proteomics, it is allowed for the identification and quantification of thousands of proteins and protein modifications for a comprehensive profiling of cells, tissues, or body fluids.[8–14]

Analysis of the glycoproteome is challenging. Due to the complicated nature of glycosylation, wherein a glycoprotein could have multiple sites of glycosylation, with each site displaying heterogeneity due to variability in the attached glycan structure, most studies of glycoproteins are limited in exploring either the glycosite[15] or released glycan structure separately.[7,16,17] However, information related to both glycosites and the attached glycans at specific glycosylation sites need be determined. More recently, analyses of intact glycopeptides have been used to study glycosylation heterogeneity from a specific glycoprotein or complex samples[18–24] Recently, several software tools[25] have been developed to assist the assignment of glycopeptides, such as Byonic,[26] Glycopeptide Search (GPS),[27] Glycopep grader,[28] SpectraST,[29] GPQuest,[30] and pGlyco,[31,32] allowing for high-throughput identification of intact glycopeptides.

Recently, it has been reported that unassigned mass spectrometry spectrum exists in proteomic data sets, and may be related to unidentified post-translational modifications.[33] With the preponderance of glycosylation as a common protein modification, we sought to explore the incidence of glycopeptides in nonglycocentric data sets. In this study, we reanalyzed data from global proteomics and phosphoproteomics using GPQuest 2.0 software tool to elucidate the presence of N-linked glycopeptides. In our analysis, we showed the existence of tens of thousands of glycopeptide spectra in nonglycoenriched data sets, yielding the identification of thousands of N-linked glycopeptides using GPQuest 2.0. Overall, these results further support the hypothesis of glycosylation as one of the most abundant modifications of proteins in human cells and that glycopeptides are widely present in global and immobilized metal affinity chromatography (IMAC)-enriched phosphorylated proteomic mass spectrometry data sets.

RESULTS

The abundance of glycopeptides identified from the global and phosphoproteomic analyses of breast cancer xenograft tissues. To determine the abundance of glycopeptides in proteomic analysis, two human-in-mouse xenograft tissues were analyzed by proteomics and phosphoproteomics. This pair of xenograft tissues representing basal (P32) and luminal-B (P33) human breast cancer were used as reference materials to determine the longitudinal reproducibility of liquid chromatography–tandem mass spectrometry (LC–MS/MS) experiments during the analysis of cancer tissues in Clinical Proteomic Tumor Analysis Consortium (CPTAC) project.[8,11–14] The pair of xenograft tissues were tryptic digested, labeled by TMT, fractionated by bRPLC, and 5% of fractionated peptides were analyzed by LC– MS/MS for global proteomics. The remaining 95% of peptides were used for the enrichment of phosphopeptides using IMAC and analyzed by LC–MS/MS. We first utilized the database search tool MSGF+[34] to identify peptides without modifications. As shown in Figure 1A and Table S1, MSGF+ assigned over 27% of total 1 163 831 HCD MS/MS spectra from the 24 fractions of global proteomics data, with FDR ≤ 1%. We then determine whether the unassigned spectra could come from glycopeptides. It is well-known that the HCD fragmented oxonium ions in MS/MS spectra are reliable indicators of glycopeptides.[19] The spectra containing HexNAc oxonium ion (*m/z* 204.0966) in the top 10 most intense peaks are regarded as potential candidates of the glycopeptide-related spectra and named as "oxo-spectra". The search results of oxo-spectra in the global proteomics data sets are highlighted as yellow in Figure 1A, revealing that approximately 1% MS/MS spectra (12 896) were identified as "oxo-spectra".We employed the same search method to explore the phosphoproteomics data set, with the results shown in Figure 1B. MSGF+ assigned approximately 12.1% of the total 590 995 HCD MS/MS spectra from 13 fractions of phosphoproteomics data to phosphopeptides with FDR ≤ 1%. Of note, we found more than 30% of the MS/MS spectra identified as "oxo-spectra" in the phosphoproteomic data set. These results show that there was approximately about 30 times more potential glycopeptide-related spectra observed in phosphoproteomics data set than in the global proteomics data set. Surprisingly, we observed the number of "oxo-spectra" was higher than the number of identified spectra assigned as phosphopeptides by MSGF+ in the phosphoproteomics data set, indicating that the IMAC-based phosphopeptide affinity capture could also enrich glycosylated peptides. Taken together, these results showed that intact glycopeptides are widely present in these global and phosphoproteome data sets from human-in-mouse xenograft tissues.

Identification of N-Linked Glycopeptides of the Breast Tumor Xenograft by GPQuest.

To facilitate identification of intact N-linked glycopeptides from resulting "oxo-spectra", the global and IMAC-based phosphopeptide- enriched phosphoproteomics data sets of the breast cancer xenograft tissues were searched using GPQuest against a human database containing over 30 000 known N-linked glycopeptide sequences and 181 N-linked glycan compositions (unpublished data). The search results are summarized in Table 1. In the global proteomics data set, there were 2 631 MS/MS spectra assigned as intact N-linked glycopeptides by GPQuest with FDR ≤ 1%, corresponding to 1 660 N-linked glycopeptides, 427 N-linked glycosites, and 111 N-linked glycan compositions from 266 N-linked glycoproteins. Over 20% "oxo-spectra" (2 631/12 896) were successfully assigned as N-linked glycopeptides. In the phosphoproteomics data, there were 5% (9 836/180 330) MS/MS spectra assigned as intact N-linked glycopeptides by GPQuest 2.0 with FDR ≤ 1%, corresponding to 5 987 N-linked glycopeptides, 1 041 N-linked glycosites, and 163 N-linked glycan compositions from 589 N-linked glycoproteins (see Table 1 and Table S2). It is not surprising that there were more N-linked glycopeptides identified from the phosphoproteomics data relative to the global proteomics data based on the glycopeptide abundances estimated by the number of oxo-spectra (see Figure 1).

Sialylated glycopeptides were enriched in phosphoproteomics data set.

The compositions of identified glycans in both global and IMAC-based phosphopeptide-enriched proteomics data were further analyzed to determine the type of glycans identified in these two proteomics data sets. We observed a higher percentage of glycans containing sialic acid (28%) in the phosphoproteomics data compared to that of the global proteomics data (9%), suggesting that the IMAC-based phosphopeptide enrichment method may selectively enrich sialylated glycans (See Figure 2A). To determine the glycans specifically enriched in phosphoproteomics data set, we investigated the distribution of PSMs of different glycan types. The PSMs were grouped by the sample and the glycan type attached to the assigned *N*-linked glycopeptides. For classification of intact glycopeptides, glycans containing HexNAc = 2, Hex ≥5 and no other monosaccharides were classified as "Oligomannose", glycans containing HexNAc ≥ 2, Hex ≥3, Sialic Acid ≥1 grouped as "Sialylated", and finally all remaining glycans assigned in the "Others" group, which included nonsialylated hybrid or complex glycans. As shown in Figure 2B, the phosphoproteomics data were found to be highly enriched for sialylated *N*-linked glycopeptides compared to global proteomics data sets. We observed a less enrichment for nonsialylated hybrid or complex glycans, or oligomannose *N*-linked glycopeptides, which indicated that IMAC enrichment for phosphopeptides displayed a coenrichment for sialylated glycopeptides and less enrichment for glycopeptides with neutral glycans.

Intact Glycopeptide Analysis on CPTAC Breast Cancer Tumors.

For comprehensive proteome and phosphoproteomic analyses of breast cancer, a total of 108 breast tumors were analyzed with a total of 72 iTRAQ data sets (36 global proteomics and 36 IMAC-enriched phosphoproteomics data sets).[11] The data sets were analyzed using MSGF+, "oxo-spectra", and GPQuest as described in the analysis of breast xenograft tissues. On average, we observed 4 857 "oxo-spectra" in each global proteomics data set (Table S3 and Figure 3), whereas the average number of PSMs of phosphopeptides identified by MSGF+ in each global data set was 3 617. This result suggests that glycopeptides in global samples without any enrichment are abundant, the total number of glycan containing spectra ("oxo-spectra") is similar to the number of the spectra identified as phosphopeptides, which is currently considered as one of the most abundant protein modifications in eukaryotes.[35] The median number of intact N-linked glycopeptides, glycosites, glycans, and glycoproteins identified by GPQuest with FDR ≤ 1% in each global proteomics data set was about 527, 238, 61, and 164, respectively. In sum, 26994 PSMs, 2 683 N-linked glycopeptides from 727 glycosites, 420 glycoproteins, and 114 glycans were identified from 36 global proteomic data sets. In each phosphoproteomics data set, there were averagely 23 100 "oxo-spectra". The median number of N-linked glycopeptides, glycosites, glycans, and glycoproteins identified by GPQuest with FDR≤ 1% (glycopeptide-spectrum matching GPSM level) in each phosphoproteomics data set was around 611, 247, 69, and 163, respectively. In sum, 44865 PSMs, 4 554 N-linked glycopeptides from 978 glycosites, 514 glycoproteins, and 144 glycans were identified from 36 phosphoproteomic data sets (Table S4).

The number of assigned phosphopeptides and N-linked glycopeptides from all 72 data sets are summarized in Figure 3. We plotted the log (PSM/spectral count) of identified phosphopeptides, "oxo-spectra", N-linked glycopeptides, and sialylated N-linked glycopeptides against the phosphoproteomics data sets (odd sample ID number) and the corresponding global proteomics data sets from the identical breast tumors (even sample ID number).The results showed the abundance of "oxo-spectra" and N-linked glycopeptides, especially sialylated N-linked glycopeptides were strongly positively correlated to the spectral counting of phosphopeptides in nearly all the data sets except 3 cases (nos. 5, 13, and 47), which illustrated that the IMAC-based phosphopeptide affinity capture method coenriched glycopeptides from the original samples in most of the cases. In total, 71 859 Glycopeptide-Spectrum Matchings (GPSMs), 5 625 N-linked glycopeptides, 1 163 glycosites, and 146 glycans were identified from 590 glycoproteins of the 72 data sets. The complete results of all data sets are shown in Table S4.

DISCUSSION

Glycosylation is one of the most abundant and important modifications of protein. However, the identification of glycosylation is often omitted in the previous studies due to a multitude of factors, including their low abundance, dynamic expression, and lack of complementary approaches to investigate multiple protein modifications in parallel. However, with the rapid advances of mass spectrometry technologies, sample preparation, and novel search algorithm strategies, there has been a growing interest in analyzing a wide array of protein/ peptide modifications including intact glycopeptides. In this study, we investigated the presence of N-linked glycoproteins in the global and IMAC-based phosphopeptide-enriched proteomics data sets of the breast cancer xenograft and CPTAC breast cancer samples, revealing a large number of N-linked glycopeptides in a total of 84 326 GPSMs, 11 292 N-linked glycopeptides, 1731 glycosites, and 166 glycans identified from 883 glycoproteins that were omitted in the original studies. This observation has provided further evidence that glycosylation is one of the most abundant protein modifications, illustrating its significant role in biology and the importance of its inclusion in future analyses.

The development of glyco-specific algorithms and tools to interrogate MS/MS spectra is critical to identify glycopeptides. GlycopeptideSearch (GPS) is a semiautomated tool for the identification of N-linked glycopeptides.[27] It can generate a list of most likely candidates of the combinations of peptides and glycans but requires additional manual validation to report and ascertain a false discovery rate (FDR). GlycoPep grader is a web-based utility for N-linked glycopeptide identification, which considers fragments from complexity composition of glycan structures to achieve more accurate glycan structure assignment.[28] It does not support high-throughput searching. Another software tool, pGlyco, can be utilized for intact N-linked glycopeptide analysis with an FDR estimation.[31,32] The most recent version eliminated the dependency on the multiple fragmentation strategies (i.e., combining MS3 and HCD/ETD spectra), instead, required high quality of HCD MS/MS spectra using a stepped fragmentation method. Byonic is another software tool for intact glycopeptide analysis.[26] It is a commercial software with user-friendly interface. Finally, the GPQuest used in this study is a freely available software focused on the high-throughput intact glycopeptide analysis, utilizing HCD MS/MS spectra.[30] GPQuest 2.0 was rewritten in Python 2.7 to apply a novel ion-index based searching strategy for intact glycopeptide identification, which supports parallel-computing for much larger search space than last version. Many other new features were implemented in GPQuest 2.0, such as the new method of precursor mass correction, N-/O- linked glycan type prediction, more customized options of modifications on glycans and peptides. GPQuest 2.0 also provided a new spectrum viewer for users to review the search results manually. The software provides FDR estimation for each assignment of the identified glycopeptide as well as predicting glycosylation types based on generated oxonium ions.[36]

The presence of unassigned spectra existing in shotgun proteomics experiment has been observed previously, even with proper search parameter settings.[37] The missed assignment of glycopeptides is one potential contributor to the unassigned spectra, resulting from the limitation of protein database search space or with defined protein modifications.[38] With knowledge of oxonium ions indicating the fragmentation of a glycan structure, we can exploit this occurrence and search for “oxo-spectra” that can then be used to roughly estimate the potential expression of glycopeptides in proteomic data sets. As a proof-of-concept, we showed that “oxo-spectra” occupied about 1% of a global proteomics data set and over one-third of the unassigned MS/MS spectra from an IMAC-based phosphopeptide-enriched proteomics data set (Figure 1). The percentages of the “oxo-spectra” indicated that adding glycopeptides to the search space is essential, especially in the IMAC-based phosphopeptide-enriched samples. In fact, thousands of N- linked glycopeptides were identified from the global proteomics and IMAC-enriched phosphoproteomics data sets as shown in Table 1, which is sufficient to be regarded as a large-scale profiling of protein N-linked glycosylation of breast cancer samples. Additional unassigned oxo-spectra in this study could be contributed by glycosylation other than N-linked glyco-sylation, which were not evaluated in this study for glycopeptide identifications. There were some other factors for failure of spectral assignments, such as incompleteness of search space of peptides or glycan in database, unknown or unconsidered modificatoins on the peptides or glycans, loss of informative MS/MS fragment ions, unknown adducts, mixed spectra, and multiglycosylated peptides. The identification rate was lower in phosphoproteomic data probably because some glycosylated peptides enriched by IMAC technology were also phosphorylated or carried sulfated glycans that were not considered in the searching strategy of this study. With increased understanding of spectral patterns from different glycosylation and further development of GPQuest software tool to include algorithms for other types of glycosylation, there should be additional glycopeptides identified from the “oxo-spectra” that were not identified as N-linked glycopeptides.

According to the search results of “oxo-spectra” and intact N-linked glycopeptides by GPQuest, phosphopeptide affinity capture using IMAC enrichment technology enriches not only phosphopeptides but also sialoglycopeptides, possibly due to the negative charges carried by both phosphopeptides and sialoglycopeptides. As shown in Figure 2A, 28% of the intact N-linked glycopeptides identified in the IMAC-based phosphopeptide-enriched samples contained a sialic acid residue but only 9% of those in global samples contain sialic acids. It is reasonable to believe that phosphopeptide affinity capture using IMAC enrichment technology may selectively enrich sialylated glycosylated peptides compared to other intact glycopeptides. Moreover, the identification of 72% glycopeptides without sialic acids could be related to the loss of sialic acids during the LC- MS/MS analysis after initial IMAC enrichment procedure, due to the labile nature of sialic residues during ionization.[39] This latter prospect is supported by the specific enrichment of hybrid and complex glycopeptides but not glycopeptides with neutral oligomannose which would not contain the sialic residue. This enrichment preference on sialylated glycosylation might result in less informative spectra due to large glylcan modification to the peptides for peptide backbone fragmentation and low ionization efficiency, which could influence the identification rate of intact glycopeptides in phosphoproteomic data. Further investigation of the influence of IMAC-enrichment on intact glycopeptides, specifically sialylated intact glycopeptides is warranted, as well as the scheme of the coenrichment of both phosphopeptides and intact glycopeptides.

CONCLUSIONS

With the rapid development of mass spectrometry and computational glycoproteomics tools, we can conduct a large-scale analysis of glycoproteome without specific glycopeptide enrichment and mass spectrometry analysis. In this study, we used our recently developed and improved intact glycopeptide analysis tool, GPQuest 2.0, to investigate the glycopeptide expression in two large data sets: two breast cancer xenograft samples and 108 CPTAC breast cancer tissues. The search results of “oxo-spectra” and intact N-linked glycopeptides analysis demonstrated the feasibility of profiling glycoproteome utilizing global proteomics or the phosphoproteomics data sets. It was also shown that the IMAC-based phosphopeptide enrichment technology has a specific preference for glycopeptides containing sialylated glycans. There is no doubt that the information on glycoproteins from the additional intact glycopeptide analysis can provide a new dimension of in viewing the glycoproteome, as well as help us further understand the complicated biological role of glycosylation in nature.

METHODS

Sample Preparation of “Breast Cancer Xenograft” Sample.

The xenograft tumor samples acquired from CPTAC program were generated from primary or metastatic breast tumors.[11–14] The 50 mg tissue of each tumor sample was lysed with sonication in 8 M urea and 1 M NH_4HCO_3 pH 8, containing 75 mM NaCl. Inhibitors of phosphatase and O-GlcNAcase were added in the lysis buffer. After lysis, proteins were reduced with 5 mM DTT, alkylated with 10 mM IAA, and digested with LysC and trypsin (Promega) at 37 °C. The digested peptides were desalted on C18 SepPak columns (Waters). 400 *μ*g of desalted peptides were then labeled by an individual channel of TMT10plex (Thermo Fisher Scientific). After TMT label reaction, all ten channels were combined and desalted on a C18 SepPak column. Basic reversed phase fractionation was performed on Agilent 1100 HPLC analytical system, generating 24 fractions for global proteome analysis and 13 combined fractions for phosphopeptide enrichment. Phosphopeptides were enriched from each of the 13 fractions using IMAC method and desalted by stage-tips. The global and phosphoproteomic study of CPTAC breast cancer tissues was described in a previous publication.[11]

Mass Spectrometry Analysis.

Each fraction of global proteomics data and phosphoproteomics data was analyzed on a Lumos instrument (Thermo Fisher Scientific) once with Data-Dependent Acquisition (DDA). The DDA run consisted of one MS survey scan (60 000 resolution; AGC target, 4.0 × 10^5; mass range, 350–1800 *m/z*; charge state include, 2–6) followed by 20 MS/MS scans (15 000 resolution; AGC target, 5.0 × 10^4; HCD collision energy, 37 eV), with former precursors excluded for 45 s after being selected once (dynamic exclusion option).

Data Sets.

The “Xenograft” data set of the breast cancer cells generated above was called as “Breast Cancer Xenograft” or “Xenograft” data set in this study. The other data sets of CPTAC breast cancer cells were downloaded from the CPTAC Data Portal[11] (https://cptac-data-portal.georgetown.edu/cptacPublic/). The whole data repository was generated for proteogenomic analysis of TCGA breast cancer samples. It contains 72 data sets of 108 TCGA samples (36 global proteomic data sets and 36 phosphoproteomic data sets). IMAC-enriched phosphoproteomics and global proteomics data sets were marked in odd and even data set ID numbers, respectively. There were 13 fractions of each phosphopeptide-enriched data set and 25 fractions of each global proteomics data set in each group of 3 TCGA samples and one pooled reference sample. All the mzML files of the 72 data sets were downloaded and studied as “Breast Cancer CPTAC” or “CPTAC” data set in the following analysis.

Searching “oxo-spectra” and N-Linked Glycopeptides by GPQuest 2.0.

The .RAW files from the Lumos instrument were converted to .mzML format by ProteoWizard 3.0[40] with the “Peak Picking” option selected for all MS levels and read by pyteomics package in Python.[41] GPQuest 2.0 was applied to investigate the expression of protein glycosylation on the unidentified MS/MS spectra in two approaches: the searching of spectra containing oxonium ions (“oxo-spectra”) and the identification of intact N-linked glycopeptides. The oxonium ions were regarded as the signature features of the glycopeptides assigned to the MS/MS spectra, which were caused by the fragmentation of glycans attached to intact glycopeptides in the mass spectrometer. In this study, the MS/ MS spectra containing the oxonium ions *(m/z* 204.0966) in the top 10 abundant peaks after removing reporter ions (TMT or iTRAQ) were considered as the potential glycopeptide candidates and named as “oxo-spectra”. The number of “oxo- spectra” was used as a preliminary estimation of protein glycosylation expression. The intact N-linked glycopeptides were identified by using GPQuest to search against a customized database containing over 30 000 known N-linked glycopeptide sequences of human cells and a database containing 181 N-linked glycan compositions. The database of N-linked glycopeptides was collected from the previous literature and unpublished results in our lab. The glycan database was collected from the public database of Glyco- meDB[42] (http://www.glycome-db.org). The theoretical b/y ions fragments of all the target and decoyed candidates of glycopeptides were calculated and built as a fragment ion index similar in MSFragger[33] but applied in a different searching strategy without precursor mass restriction. The binned *m/z* values of fragments and their belonged peptides were stored as ⟨key,value⟩ pairs in a hashed map as the search space. Each tandem mass spectrum was first processed in a series of preprocessing procedures, including removing reporter ions (TMT or iTRAQ), spectrum denoising, intensity square root transformation,[43] oxonium ions evaluation, and glycan types prediction. The top 100 peaks in each qualified preprocessed spectrum were matched to the fragment ion index to find all the candidate peptides. The candidate peptides theoretically containing the MS2 peaks of the spectrum were merged and sorted descendingly according to the number of their shared peaks. If the number of shared peaks is less than a threshold (6 peaks in this study), these candidate peptides were filtered. All the remaining candidate peptides were compared with the spectrum again to calculate the Morpheus scores[44] by considering all the peptide fragment, glycopeptide fragments, and their isotope peaks. The peptide having highest Morpheus score was assigned to the spectrum finally. The mass gap between the assigned peptide and the precursor mass was searched in the glycan database to find the associated glycan. The best hits of all “oxo-spectra” were ranked by the Morpheus score descendingly, in which those under FDR = 1% and covering over 20% total intensity of the spectrum were reserved as qualified identifications. In this strategy, no precursor mass or Y ion was required in searching the peptide which decreases the loss of identification due to incorrect precursor mass or charge assignment. The precursor mass tolerance was set as 10 ppm, and the fragment mass tolerance was 20 ppm.

Search Phosphopeptides by MSGF+.

To compare the abundances of glycopeptides and phosphopeptides in both global proteomics and IMAC-enriched phosphoproteomics data sets, MSGF+(v9949) was applied to perform the searching of phosphopeptides. MSGF+ is a commonly used search engine in the conventional proteomics approach. In the searching of Xenograft samples of breast cancer, the protein database contains 32 799 proteins of human and 32 800 proteins of mouse. The missed cleavage is 2. The in *silico* cleavage was performed by using the trypsin rule. All the other parameters were set by default.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This work was supported by funding from the National Cancer Institute, the Clinical Proteomic Tumor Analysis Consortium (CPTAC, Grant U24CA210985), the Early Detection Research Network (EDRN, Grant U01CA152813), National Heart Lung and Blood Institute, Programs of Excellence in Glycosciences (PEG, Grant P01HL107153). We also appreciate Dr. Weiming Yang, Dr. Shisheng Sun, and Dr. Shuang Yang's suggestions for the development of GPQuest 2.0 in this study. The copyright to the GPQQuest 2.0 image in the abstract/TOC graphics is retained by Hui Zhang.

REFERENCES

(1). Varki A; Cummings RD; Esko JD; Freeze HH; Stanley P; Bertozzi CR; Hart GW; Etzler ME Essentials of Glycobiology; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, 2009.

(2). Ohtsubo K; Marth JD Cell 2006, 126 (5), 855–867. [PubMed: 16959566]

(3). Schachter H; Freeze HH Biochim. Biophys. Ada, Mol. Basis Dis. 2009, 1792 (9), 925–930.

(4). Zhao Y-Y; Takahashi M; Gu J-G; Miyoshi E; Matsumoto A; Kitazume S; Taniguchi N Cancer Sci 2008, 99 (7), 1304–1310. [PubMed: 18492092]

(5). Durand G; Seta N Clin. Chem. 2000, 46 (6), 795–805. [PubMed: 10839767]

(6). Varki A Trends Mol. Med. 2008, 14 (8), 351–360. [PubMed: 18606570]

(7). Hart GW; Copeland RJ Cell 2010, 143 (5), 672–676. [PubMed: 21111227]

(8). Zhang B; Wang J; Wang X; Zhu J; Liu Q; Shi Z; Chambers MC; Zimmerman LJ; Shaddox KF; Kim S; et al. Nature 2014, 513 (7518), 382–387. [PubMed: 25043054]

(9). Aebersold R; Mann M Nature 2016, 537 (7620), 347–355. [PubMed: 27629641]

(10). Larance M; Lamond AI Nat. Rev. Mol. Cell Biol. 2015, 16 (5), 269–280. [PubMed: 25857810]

(11). Mertins P; Mani DR; Ruggles KV; Gillette MA; Clauser KR; Wang P; Wang X; Qiao JW; Cao S; Petralia F; et al. Nature 2016, 534 (7605), 55–62. [PubMed: 27251275]

(12). Zhang H; Liu T; Zhang Z; Payne SH; Zhang B; McDermott JE; Zhou J-Y; Petyuk VA; Chen L; Ray D; et al. Cell 2016, 166 (3), 755–765. [PubMed: 27372738]

(13). Ntai I; LeDuc RD; Fellers RT; Erdmann-Gilmore P; Davies SR; Rumsey J; Early BP; Thomas PM; Li S; Compton PD; et al. Mol. Cell. Proteomics 2016, 15 (1), 45–56. [PubMed: 26503891]

(14). Tabb DL; Wang X; Carr SA; Clauser KR; Mertins P; Chambers MC.; Holman JD; Wang J; Zhang B; Zimmerman LJ; et al. J. Proteome Res. 2016, 15 (3), 691–706. [PubMed: 26653538]

(15). Zhang H; Loriaux P; Eng J; Campbell D; Keller A; Moss P; Bonneau R; Zhang N; Zhou Y; Wollscheid B; et al. Genome Biol. 2006, 7 (8), R73–R73. [PubMed: 16901351]

(16). Yang S; Zhang H Curr. Protoc. Chem. Biol. 2014, 6 (3), 191–208. [PubMed: 25205566]

(17). Yang S; Rubin A; Eshghi ST; Zhang H Proteomics 2016, 16 (2), 241–256. [PubMed: 26390280]

(18). Pompach P; Brnakova Z; Sanda M; Wu J; Edwards N; Goldman R Mol. Cell. Proteomics 2013, 12 (5), 1281–1293. [PubMed: 23389049]

(19). Yang W; Shah P; Toghi Eshghi S; Yang S; Sun S; Ao M; Rubin A; Jackson JB; Zhang H Anal. Chem. 2014, 86 (14), 6959–6967. [PubMed: 24941220]

(20). Go EP; Herschhorn A; Gu C; Castillo-Menendez L; Zhang S; Mao Y; Chen H; Ding H; Wakefield JK; Hua D; et al. J. Virol. 2015, 89 (16), 8245–8257. [PubMed: 26018173]

(21). Shah P; Wang X; Yang W; Eshghi ST; Sun S; Hoti N; Chen L; Yang S; Pasay J; Rubin A; et al. Mol. Cell. Proteomics 2015, 14 (10), 2753–2763. [PubMed: 26256267]

(22). Sun S; Shah P; Eshghi ST; Yang W; Trikannad N; Yang S; Chen L; Aiyetan P; Hoti N; Zhang Z; et al. Nat. Biotechnol 2016, 34 (1), 84–88. [PubMed: 26571101]

(23). Zhou J; Yang W; Hu Y; Hoti N; Liu Y; Shah P; Sun S; Clark D; Thomas S; Zhang H Anal. Chem. 2017, 89 (14), 7623–7630. [PubMed: 28627880]

(24). Medzihradszky KF; Kaasik K; Chalkley RJ Mol. Cell. Proteomics 2015, 14 (8), 2103–2110. [PubMed: 25995273]

(25). Tsai P-L; Chen S-F Mass Spectrom. 2017, 6 (2), S0064.

(26). Bern M; Kil YJ; Becker C Curr. Protoc. Bioinformatics 2012, Unit13.20.

(27). Pompach P; Chandler KB; Lan R; Edwards N; Goldman R J. Proteome Res. 2012, 11 (3), 1728–1740. [PubMed: 22239659]

(28). Woodin CL; Hua D; Maxon M; Rebecchi KR; Go EP; Desaire H Anal. Chem. 2012, 84 (11), 4821–4829. [PubMed: 22540370]

(29). Pai P-J; Hu Y; Lam H Anal. Chim.Acta 2016, 934, 152–162. [PubMed: 27506355]

(30). Toghi Eshghi S; Shah P; Yang W; Li X; Zhang H Anal. Chem. 2015, 87 (10), 5181–5188. [PubMed: 25945896]

(31). Zeng W-F; Liu M-Q; Zhang Y; Wu J-Q; Fang P; Peng C; Nie A; Yan G; Cao W; Liu C; et al. Sci. Rep. 2016, 6, 25102. [PubMed: 27139140]

(32). Liu M-Q; Zeng W-F; Fang P; Cao W-Q; Liu C; Yan GQ; Zhang Y; Peng C; Wu J-Q; Zhang X-J; et al. Nat. Commun. 2017, 8 (1), 438. [PubMed: 28874712]

(33). Kong AT; Leprevost FV; Avtonomov DM; Mellacheruvu D; Nesvizhskii AI Nat. Methods 2017, 14 (5), 513–520. [PubMed: 28394336]

(34). Kim S; Pevzner PA Nat. Commun. 2014, 5, 5277. [PubMed: 25358478]

(35). Khoury GA; Baliban RC; Floudas CA Sci. Rep. 2014, 1, 90.

(36). Toghi Eshghi S; Yang W; Hu Y; Shah P; Sun S; Li X; Zhang H Sci. Rep. 2016, 6, 37189. [PubMed: 27869200]

(37). Ning K; Fermin D; Nesvizhskii AI Proteomics 2010, 10 (14), 2712–2718. [PubMed: 20455209]

(38). Chick JM; Kolippakkam D; Nusinow DP; Zhai Bo; Rad R; Huttlin EL; Gygi SP Nat. Biotechnol. 2015, 33 (7), 743–749. [PubMed: 26076430]

(39). Yang S; Jankowska E; Kosikova M; Xie H; Cipollo J Anal. Chem. 2017, 89 (17), 9508–9517. [PubMed: 28792205]

(40). Chambers MC; Maclean B; Burke R; Amodei D; Ruderman DL; Neumann S; Gatto L; Fischer B; Pratt B; Egertson J; et al. Nat. Biotechnol. 2012, 30, 918. [PubMed: 23051804]

(41). Goloborodko AA; Levitsky LI; Ivanov MV; Gorshkov MV J. Am. Soc. Mass Spectrom. 2013, 24 (2), 301–304. [PubMed: 23292976]

(42). Ranzinger R; Herget S; von der Lieth C-W; Frank M Nucleic Acids Res. 2011, 39 (Database), D373–D376. [PubMed: 21045056]

(43). Liu J; Bell AW; Bergeron JJM; Yanofsky CM; Carrillo B; Beaudrie CEH; Kearney RE Proteome Sci. 2007, 5, 3. [PubMed: 17227583]

(44). Wenger CD; Coon JJ J. Proteome Res. 2013, 12 (3), 1377–1386. [PubMed: 23323968]

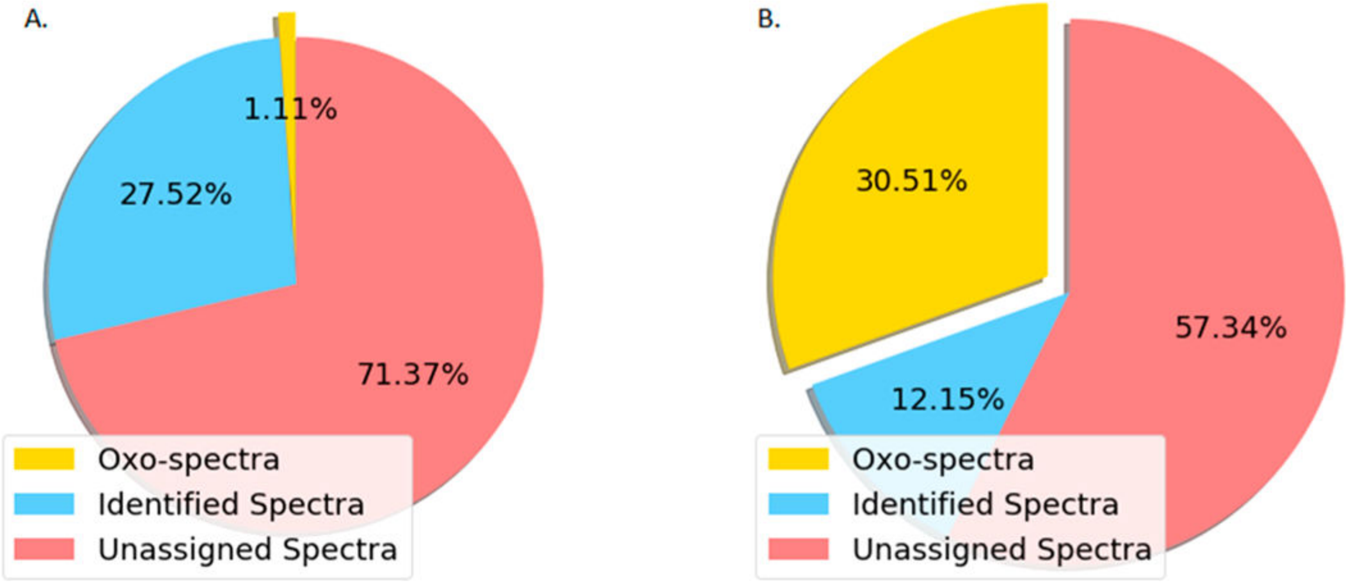


Figure 1.

Preliminary estimation of potential glycopeptides in the global and phosphoproteomic data sets from breast cancer xenograft tissues. The search results of global and phosphopeptide-enriched data sets are shown in parts A and B pie charts, respectively. The blue chart represents the proportion of identified peptides (A) and phosphopeptides (B). The red part presents the proportion of unidentified spectra. The yellow part represents the proportion of “oxo-spectra”.

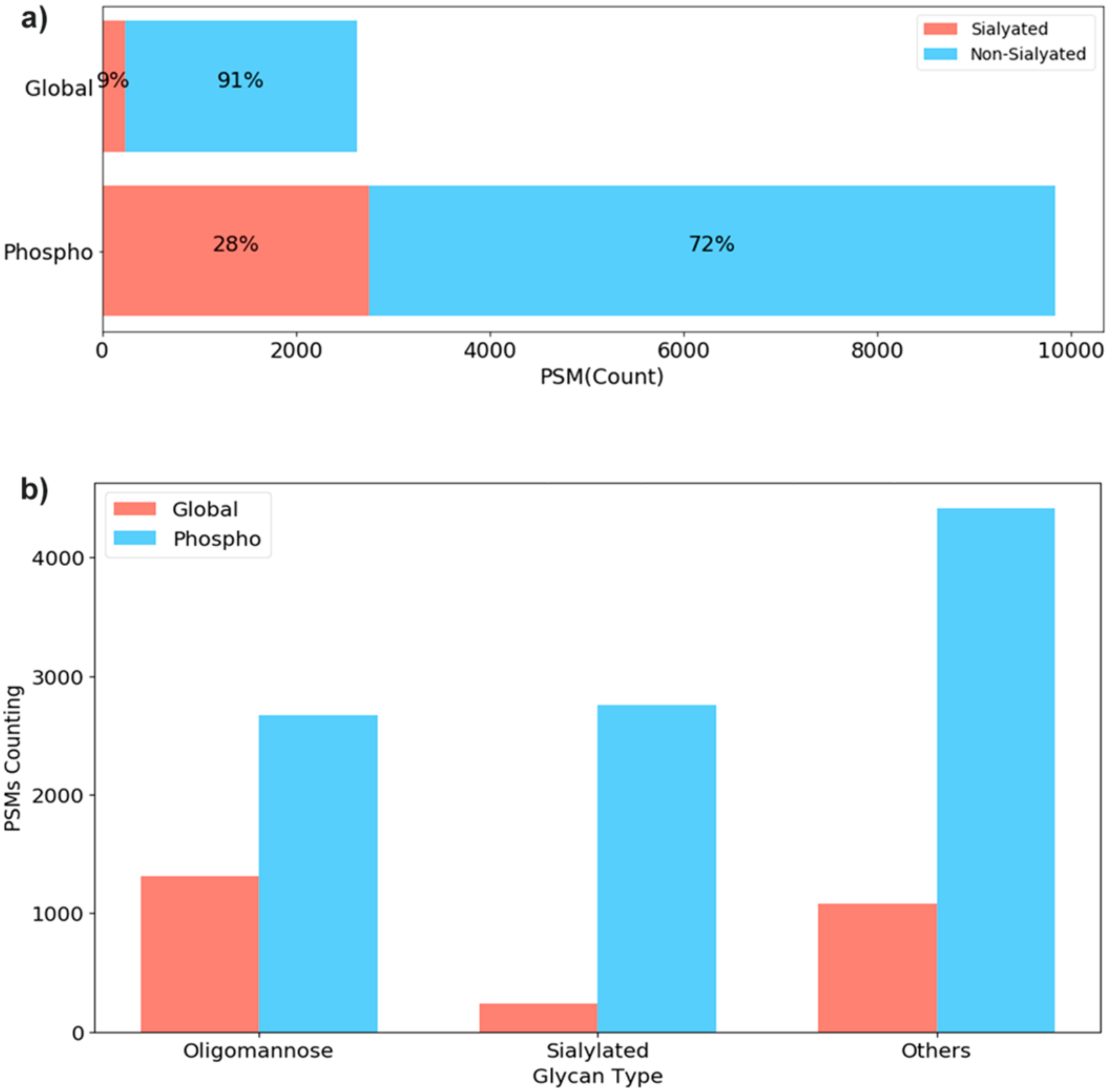


Figure 2.

Classified glycopeptide-spectrum matchings (PSMs) of different glycopeptides from global proteomics and IMAC-enriched phosphoproteomics data sets of breast cancer xenograft tissues. (A) The red bars are the percentage of PSMs assigned as N-linked sialoglycopeptides. The blue bars are the percentage of PSMs from nonsialylated glycopeptides. (B) Total number of glycopeptide-spectrum matchings (PSMs) of different glycan types in global proteomics and IMAC-enriched phosphoproteomics data sets.

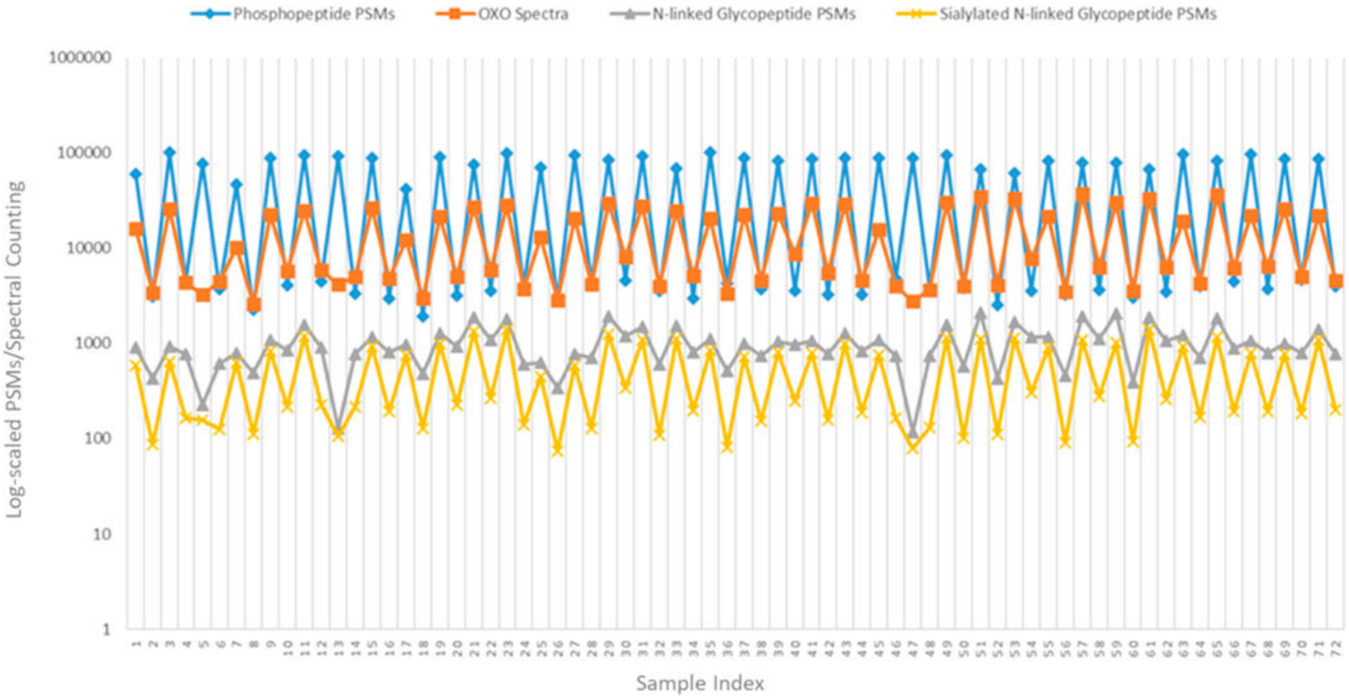


Figure 3.

Abundance of phosphopeptides and glycopeptides in all the samples of "Breast Cancer CPTAC" data. All the 72 data sets (36 global and 36 phosphopeptide-enriched data sets) of the CPTAC breast cancer cells were searched for phosphorylation (blue), oxo-spectra (red), intact N-linked glycopeptides (gray), and sialylated N-linked glycopeptides (yellow). The results of PSMs/spectral counting were plotted together. All the odd index numbers represent phosphoproteomic data sets. All the even index numbers represent the global proteomic data sets of the sample with an index of $N-1$, where N is the even index number.

Table 1.

Summary of Glycoproteomic Results from GPQuest Search of Xenograft and CPTAC Data Sets

	xenograft global	xenograft phospho	CPTAC global	CPTAC phospho
PSM	2631	9836	26994	44865
glycopeptide	1660	5987	2683	4554
glycosite	427	1041	727	978
glycoprotein	266	589	420	514
glycan	111	163	114	144