A Comprehensive, Open-source Platform for Mass Spectrometry-based Glycoproteomics Data Analysis*[S]

Gang Liu‡, Kai Cheng‡§**, Chi Y. Lo‡, Jun Li¶‖, Jun Qu¶‖, and Sriram Neelamegham‡§‡‡**

Glycosylation is among the most abundant and diverse protein post-translational modifications (PTMs) identified to date. The structural analysis of this PTM is challenging because of the diverse monosaccharides which are not conserved among organisms, the branched nature of glycans, their isomeric structures, and heterogeneity in the glycan distribution at a given site. Glycoproteomics experiments have adopted the traditional high-throughput LC-MS^n proteomics workflow to analyze site-specific glycosylation. However, comprehensive computational platforms for data analyses are scarce. To address this limitation, we present a comprehensive, open-source, modular software for glycoproteomics data analysis called GlycoPAT (GlycoProteomics Analysis Toolbox; freely available from www.VirtualGlycome.org/glycopat). The program includes three major advances: (1) "SmallGlyPep," a minimal linear representation of glycopeptides for MS^n data analysis. This format allows facile serial fragmentation of both the peptide backbone and PTM at one or more locations. (2) A novel scoring scheme based on calculation of the "Ensemble Score (ES)," a measure that scores and rank-orders MS/MS spectrum for N- and O-linked glycopeptides using cross-correlation and probability based analyses. (3) A false discovery rate (FDR) calculation scheme where decoy glycopeptides are created by simultaneously scrambling the amino acid sequence and by introducing artificial monosaccharides by perturbing the original sugar mass. Parallel computing facilities and user-friendly GUIs (Graphical User Interfaces) are also provided. GlycoPAT is used to catalogue site-specific glycosylation on simple glycoproteins, standard protein mixtures and human plasma cryoprecipitate samples in three common MS/MS fragmentation modes: CID, HCD and ETD. It is also used to identify 960 unique glycopeptides in cell lysates from prostate cancer cells. The results show that the simultaneous consideration of peptide and glycan fragmentation is necessary for high quality MS^n spectrum annotation in CID and HCD fragmentation modes. Additionally, they confirm the suitability of GlycoPAT to analyze shotgun glycoproteomics data. ***Molecular & Cellular Proteomics 16: 10.1074/mcp.M117.068239, 2032–2047, 2017.***

Glycosylation regulates protein folding and cell-cell interactions in a variety of biological contexts (1, 2). This is an important post-translational modification (PTM)[1] in the context of protein therapeutics, development, normal physiology and diseases like inflammation and cancer (2). Unlike DNA and protein that are composed of a uniform set of nucleotide or amino acid building blocks across all organisms, monosaccharide composition is not uniform among species. To add to this complexity, glycans often contain branched structures, and they can be heterogeneous both in terms of whether a particular site is glycosylated (macroheterogeneity) and also in terms of the distribution of different glycans at a single site (microheterogeneity). This heterogeneity reflects the metabolic status of the cell, tissue or organ system at multiple levels, particularly the factors controlling mRNA transcription, protein translation and glycosylation reaction rates (3).

Tools to study glycosylation are rapidly being developed and recent years have witnessed the increasing use of mass spectrometry (MS) for the structural analyses of glycans (4, 5). In this regard, although classical glycomics methods first separate the glycans from proteins to determine either glycan structure or site of protein glycosylation, more recent glycoproteomics workflows focus on analyzing site-specific glycosylation by interrogating the intact glycopeptide (4, 6). Commonly, the latter applications use liquid chromatography (LC)

From the ‡Chemical and Biological Engineering; §Clinical & Translational Research Center; ¶Pharmaceutical Sciences; and ‖New York State Center for Excellence in Bioinformatics and Life Sciences, Buffalo, New York

Received March 7, 2017, and in revised form, August 23, 2017
Published, MCP Papers in Press, September 8, 2017, DOI 10.1074/mcp.M117.068239
Author contributions: G.L., K.C., J.Q., and S.N. designed the study. J.L. performed MS experiments. G.L., K.C., C.Y.L., and S.N. wrote the code, performed simulations, and analyzed data. G.L., K.C., and S.N. wrote the manuscript. All authors read, reviewed and approved the final manuscript.

[1] The abbreviations used are: PTM, post-translational modification; API, Application Program Interface; *ES*, Ensemble Score; GUI, Graphical User Interface; GlycoPAT, GlycoProteomics Analysis Toolbox; GlyDB, Glycan database; GlyPepDB, GlycoPeptide database; XML, eXtensible Markup Language; SGP1.0, Small GlycoPeptide nomenclature, version 1.0.

to resolve a complex mixture of (glyco)peptides that are generated by the enzymatic digestion of proteins. In the most popular format, following electrospray ionization (ESI) and high-resolution precursor/MS^1 mass quantitation, tandem MS^n analysis is performed on selected ions following fragmentation using either vibrational dissociation methods like CID (collision induced/activated dissociation) and HCD (higher-energy collisional dissociation or beam-type CID), or activated electron dissociation methods like ETD (electron transfer dissociation) (4, 6, 7). Because of the high-throughput nature of the experiment, each LC-MS run results in tens of thousands of fragmentation spectra. Here, the CID mode is prone to producing B-/Y-ions because of glycan fragmentation while leaving the peptide backbone largely intact. Thus, it can assign glycan structure but not the site of glycosylation. HCD results in more extensive glycan fragmentation compared with CID, and peptide backbone b-y ion fragmentation. Although it does not provide detailed glycan-structure information, it identifies MS/MS spectra corresponding to glycopeptides because of the release of prominent low molecular mass mono- and disaccharide oxonium ions. Partial information on the site of glycosylation is also obtained (8). ETD predominantly results in N-Cα peptide bond cleavage to generate c-/z-type ions while leaving the glycan(s) intact (4). This is invaluable for the identification of glycosylation sites. Together, the complementary fragmentation data regarding the glycans and peptide backbone may be spliced together for comprehensive structural analysis.

Although several programs exist for the analysis of either one or a few glycoproteomics tandem MS spectra, the lack of programs that can handle high-throughput data is a major limitation in the field (reviewed by (5, 7, 9, 10)). Although a few programs for such data analysis have appeared, there is no gold-standard because the glycoproteomics experimental workflows are still evolving (4). Specifically, many of the currently available programs either only handle limited fragmentation modes or provide specific specialized analysis functions (11–13), cannot handle data in XML (eXtensible Markup Language) format (14, 15), lack a well-developed scoring algorithm (16), are protein centric in that they focus on the site of glycosylation and glycan composition rather than detailed carbohydrate structure (17, 18), lack a user-friendly Graphical User Interface (GUI) (11, 12, 14, 16, 18–20) or are proprietary (17). The program "Protein Prospector" has also been modified to handle organism-scale glycopeptide databases, particularly for ETD fragmentation mode data analysis (21). Although some of these programs are "freely available" on request, to our best knowledge none of these are open-source and modular, with comprehensive documentation that can enable expansion by the community. This is important because the number of ways in which tandem-MS runs can be performed with different fragmentation modes is large, as more than one mode of fragmentation may be applied in a single LC run. The analyses of such experiments can be even more complicated when the individual runs interrogate MS^3 and higher level. Such higher-level analysis is likely to be part of future glycoproteomics workflows because of the need to distinguish between different isomeric, complex glycans (22). Thus, different spectra refinement and scoring strategies need to be tested and this cannot be accomplished by existing programs. Additionally, to the best of our knowledge, existing programs do not handle custom monosaccharide definitions and they do not contain a systematic strategy to fragment glycopeptides at multiple locations, a requirement for efficient MS^n analysis. Because of these limitations, a majority of investigators in the field continue to rely on manual data post-processing and spectral interpretation, and this limits scientific progress (4, 6).

To address the above limitations, this manuscript introduces a new computational, open-source framework called GlycoProteomics Analysis Toolbox (GlycoPAT, available from www.VirtualGlycome.org/glycopat). This program is modular in design, and it is built around a new linear, minimal representation of glycopeptides called *SmallGlyPep* (SGP 1.0). SGP1.0 is well suited for MS^n data analysis because it allows the straightforward representation of multiple glycans on a single peptide backbone, and efficient *in silico* glycopeptide fragmentation at one or more locations. The incorporation of application programming interfaces (APIs) from a previous toolbox called GNAT (Glycosylation Network Analysis Toolbox, (23)), enables generation of candidate glycan search libraries based on existing knowledge of biochemistry. Data in both text and mzXML input formats are supported (24). The framework includes several additional, novel features including: (1) A scoring scheme to rank candidate glycopeptides based on an ensemble score (*ES*) which integrates multiple statistical parameters including cross-correlation and probability based scores; (2) A method to identify minimum acceptable *ES* scores ($ES_{cut\text{-}off}$) based on decoy libraries and glycopeptide false discovery rate (FDR) calculations; and (3) Parallel computing facilities to accelerate processing of bulky experimental data and search libraries. The program has been tested using single standard glycoproteins, simple mixtures of proteins, human blood plasma cryoprecipitate mixtures enriched in coagulation-related proteins, and complex prostate cancer cell lysates. Stand-alone GUIs, with basic functionality, are also provided to facilitate quick usage by those without access to the MATLAB software, persons unfamiliar with programming, and individuals looking for a ready-to-go, freely available application for data analysis. The results confirm the ability of GlycoPAT to perform scoring in multiple fragmentation modes, and analyze complex biological samples. It shows that the simultaneous consideration of peptide and glycan fragmentation enhances the quality of MS^n spectrum annotation, particularly following HCD and CID fragmentation.

EXPERIMENTAL PROCEDURES

Code Availability—GlycoPAT program source code, compiled GUIs and detailed instructional manuals and videos are available at the Sourceforge and Youtube repositories. These resources can be accessed from the software homepage: www.VirtualGlycome.org/glycopat.

Experimental Design and Statistical Rationale—Single standard protein runs include 3 LC-MS runs with fetuin (UniProt P12763), 1 with asialofetuin, and 2 with RNaseB (P61823). A total of 9 runs were performed for defined mixtures that contain a mixture of proteins: fetuin, fibronectin (P02751), RNaseB and human α1-acid glycoprotein 1 (AGP-1, P02763). In additional, 14 independent runs assayed the human plasma cryoprecipitome. Additional data for Basigin/CD147 and prostate cancer cells was downloaded from PRIDE. Statistical analysis methods are described as part of the software package. The mass spectrometry proteomics data have been deposited at the ProteomeXchange Consortium (PRIDE identifier: PXD006031).

Tandem-MS Experiment—Bovine asialofetuin, bovine fetuin, bovine RNase B, and AGP-1 were purchased from Sigma-Aldrich (St. Louis, MO). Human fibronectin was from Life Technologies (Grand Island, NY). Five milliliters human blood was drawn by venipuncture from an O-blood group individual into 1:9 sodium citrate following human subjects protocols approved by the University at Buffalo Health Science IRB. Platelet poor plasma (PPP) was isolated from this blood as described previously (25). This precipitate was rapidly frozen to −80 °C, and then slowly thawed at 3 °C. Protein precipitate thus formed was collected by centrifugation at 5000 × *g* for 15 min. The pellet was resuspended in 20 mM HEPES buffer and dissolved by warming to 37 °C. For MS, all protein samples were processed with a surfactant aided on-pellet digestion procedure (26, 27). Briefly, 100 μg protein of each sample was spiked with 0.5% SDS and then denatured and reduced using 4 mM tris(2-carboxyethyl)phosphine (TCEP) at 37 °C for 30 min. Following this, fresh 20 mM iodoacetamide was added for 30 min in the dark. Then, 6-fold cold acetone was added to the sample volume in two steps with vortexing, and the mixture was incubated overnight at −20 °C. Samples thus obtained were centrifuged at 20,000 × *g* at 4 °C for 30 min, the supernatant was discarded, and the pellet was washed with methanol and then air-dried for several min. This dried pellet was smashed to small particles in 50 mM Tris-FA (formic acid) buffer (pH 8.5) with a sonicator, 1:20 (w/w) enzyme-to-protein ratio of sequencing grade trypsin or Glu-C (Thermo-Pierce) was added to a total volume of 100 μl, and then the sample was incubated at 37 °C for 18 h. 6 μg of digested samples prepared in this manner were analyzed using either an LTQ-Orbitrap XL mass spectrometer or an Orbitrap-Fusion Tribrid mass spectrometer (Thermo Scientific, San Jose, CA), both with ETD module.

The nano-LC system used for the study featured low void volume and high chromatographic reproducibility (28, 29). Mobile phase A was 0.1% FA in 1% acetonitrile, and phase B was 0.1% FA in 88% acetonitrile. Samples were loaded onto a large-ID trap (300 μm ID × 0.5 cm, packed with Zorbax 5 μm C_{18} material) with 1% B at 10 μl/min. The trap was washed for 3min before the samples were back-flush onto the nano-LC column (75 μm ID × 75 cm, packed with Pepmap 3-μm C_{18} material). The typical elution run was 2 h long at 52 °C, with a flow rate of 250 nl/min. Separation followed the sequence: 3 to 8% B over 5min; 8 to 24% B over 85min; 24 to 38% B over 15min; 38 to 63% B over 10min; 63 to 97% B in 5min. Initial MS^1 analysis (*m/z* 300–2000) was performed at a mass resolution of 60,000 for OrbitrapXL (120,000 for Fusion).

MS^2 fragmentation was at a resolution of 30,000. In runs with standard proteins, the six most abundant precursors in the MS^1 spectra were MS/MS fragmented in CID, HCD, ETD, and/or alternating CID/ETD mode. In an additional 12 runs that assayed the human plasma cryoprecipitome, MS/MS CID and ETD mode fragmentation was triggered following the initial detection of glycoprotein product-ions in HCD mode (*m/z* = 204.0871; 138.0545; 366.14002; 163.0606; 292.1032; 657.2354, 168.09). Other instrument parameters were: 3 s cycle time; scan range (*m/z*) = 150–2000; AGC target = 2×10^4 ; MS: charge state = 2∼8; dynamic exclusion after n times = 2; decision: precursor priority = highest charge state then most intense (or most intense then highest charge state); $FTMS^n$ (HCD): Isolation mode = quadrupole; isolation window = 3; collision energy (%) = 35; resolution = 30K; AGC target = 1×10^5; maximum injection time = 60 ms; microscan = 1; Product Ion trigger: at least 2 product ions detected; Top N product ions = 30; Ion Trap (IT) MSn (CID): Isolation mode = quadrupole; isolation window = 3; collision energy (%) = 30; AGC target = 1×10^5; maximum injection time = 120 ms; microscan = 1; $ITMS^n$ (ETD): isolation mode = quadrupole; isolation window = 3; AGC target = $1*10^4$; maximum injection time = 150 ms; microscan = 2; reaction time = 200 ms.

MS Data Preprocessing—The .RAW files generated from the MS instruments were converted either to text format with .dta extension using Bioworks 3.3.1 (Thermo-Scientific) or to .mzXML format using the *msconvert* tool (ProteoWizard 3.0.5759, (30)). In the .dta files, the first row contains the precursor MS^1 mass ($M+H^+$) and charge assignment inferred by Bioworks. The remaining rows list MS^2 fragment *m/z* values along the first column and corresponding intensity (*I*) data on the second column. The .mzXML file presents the same data and also additional experimental information, like fragmentation mode, in XML format.

Theoretical Glycopeptide Database Generation—The GlyDB was generated for N- and O-linked glycans as explained in Results. GlyPepDB was then synthesized *in silico* by digesting one or more proteins supplied in FASTA format using the specified proteolytic enzyme(s), and then appending both fixed and variable PTM modifications. Protein UniProt i.d. for simple mixtures is provided above, and protein accession i.d. for plasma proteins is listed in Supplemental Tables. The MATLAB function used in this step is called *digestSGP* and it outputs the GlyPepDB in SGP1.0 format. When generating this database, in this current manuscript, either 2 or 3 missed cleavages, fixed cysteine carbamidomethyl modification (+57.02146 Da) and variable methionine oxidation (+15.99492 Da) was allowed. N-glycans could appear at Asn in the N-X-S/T sequon and O-glycans were allowed at Ser/Thr. Although there was no limit on the number of fixed modifications on any peptide, the number of variable modifications was limited to two, and occasionally three. These variable modifications include both glycan and nonglycan PTMs.

Scoring Experimental Spectra to Obtain Ensemble Score (ES)—GlycoPAT scoring follows two-steps shown in Fig. 1*B*. Although the parameters used in the current manuscript are stated below and the text corresponds to the case of MS/MS fragmentation, this can be changed for other applications that require MS^n data analysis.

In the first step, the experimental MS^1 precursor mass was compared with the theoretical mass of all (glyco)peptides in GlyPepDB. GlyPepDB members with mass difference less than tolerance (typically 10ppm) are termed "candidate (glyco)peptides."

In the second step, the ensemble score (*ES*) was calculated by comparing the experimental MS/MS spectrum with the same spectrum generated for the candidate glycopeptide *in silico*. The nature of this scoring was differed among the different fragmentation modes: CID, HCD and ETD. In this regard, one of two noise reduction methods was applied to delete low intensity peaks in the experimental data that are because of instrument noise. "Global noise reduction" was applied in CID and HCD modes to remove all peaks less than 2-times median peak intensity provided these are <1% of the most intense peak. "Local noise reduction" was applied in the ETD mode to delete the unfragmented precursor ion peak, and peaks below the median

value in local m/z windows that span 100Th. Following this, the theoretical MS/MS spectrum of the candidate glycopeptide was generated in CID/HCD modes by fragmenting glycosidic bonds to form B-Y type ions in the case of glycopeptides or peptide b-y ions when the glycan is absent. In CID mode, two glycan fragmentations were allowed when the number of monosaccharides in the glycopeptide exceeded 4 because multiple fragmentations on a single glycan can occur in the case of *N*-linked glycopeptides. A mix of 1 glycan and 1 peptide fragmentation was permitted when the glycans had 4 or fewer monosaccharides because this is commonly observed in the case of O-GalNAc type glycopeptides (31). HCD mode analysis focused on mapping the underlying peptide backbone rather than analyzing the attached glycan. Thus, glycopeptide fragments in the theoretical MS/MS spectrum included selected oxonium ions (m/z = 138.05496, 204.08665; 292.10269 and 274.09213, if the corresponding monosaccharides are present in the candidate), and peptide b/y-ions that contain glycans separated from the underlying peptide by up to 2 glycosidic linkages. In ETD mode, only c-z peptide ions were included. In all cases, oxonium ion charge state (z) = 1. Theoretical peptide fragments had $z\leq$ precursor charge state for CID and HCD, and $z<$ precursor charge state for ETD. Next, the ability of the theoretical spectrum to match experimental MS/MS data was quantified using four statistical scores:

I. Pearson Cross-correlation Analysis (X_{corr})—This procedure follows previous literature (32) with some modifications. Specifically, the theoretical MS/MS spectra was simplified such that only one peak was included for each (glyco)peptide fragment. The specific charge state for that fragment corresponded to the charge state of the theoretical peak that had a corresponding experimental MS/MS peak match. In case more than one theoretical peak had matching experimental peaks, the theoretical charge state corresponding to the most intense experimental peak was chosen. If no match was found, the fragment with charge state of +1 was retained. The intensity I of each peak in the theoretical spectra was arbitrarily set to 50. Next, the intensity data for both the theoretical and experimental fragmentation spectra were binned according to the instrument resolution. In the case where the MS/MS tolerance was say p ppm, consecutive bins at a given m/z values were separated by $(m/z)\cdot p/10^6$ Da. When the MS/MS tolerance was q Da units, consecutive bins were separated by q Da. In this manuscript, the MS/MS tolerance was often 1Da. Thus, for a theoretical or experimental peak at say m/z = 331.5, a peak with corresponding intensity was placed at m/z of both 331 and 332. The intensity of multiple peaks appearing at a given integer m/z value were then summed to determine the final intensity. The theoretical MS/MS spectrum was then offset/translated over the corresponding experimental data over a range (τ), and the cross-correlation score ($X_{corr}(\tau)$) was calculated using (32).

$$X_{corr}(\tau) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_{i-\tau} - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_{i-\tau} - \bar{y})^2}}$$

x_i and y_i correspond to the intensity of the i^{th} m/z value of the processed experimental and theoretical spectrum respectively. $\bar{x}$ and $\bar{y}$ are corresponding mean values averaged over all n possible peaks. τ ranged from −50 to +50. During the correlation analysis, two parameters were recorded: i. Peak lag, or the τ values where $X_{corr}(\tau)$ was maximum; and ii. Height Center (HC_{corr}), which quantifies the normalized height of $X_{corr}(\tau = 0)$ with respect to the mean $X_{corr}(\tau)$ value: $HC_{corr} = Xcorr(\tau = 0) \times [max(\tau) - \min(\tau) + 1]/\sum_{min(\tau)}^{max(\tau)} Xcorr(\tau)$. Thus, for a good match, peak lag should lie between +1 and −1, and HC_{corr} should be large. This is captured in the scoring parameter s_1 below:

$$s_1 = \begin{cases} 0 & \text{if } |peak\ lag| > 1 \\ \frac{HC_{corr}}{0.65} & \text{if } |peak\ lag| \leq 1\ \&\ HC_{corr} \leq 0.65 \\ 1 & \text{if } |peak\ lag| \leq 1\ \&\ HC_{corr} > 0.65 \end{cases}$$

II. % Ion Match—The total number of peaks in the full theoretical spectrum is N_1. The number of these peaks that have corresponding experimental matches is K_1. % ion match = $100 \times K_1/N_1$. High % ion match values thus indicate superior spectrum matches. Thus, the score s_2 is specified as:

$$s_2 = \begin{cases} \%\ Ion\ match/80 & \text{if } \%\ Ion\ match \leq 80 \\ 1 & \text{if } \%\ Ion\ match > 80 \end{cases}$$

III. Top 10 Peaks—This parameter quantifies how many of the 10 most intense experimental peaks were matched during the % ion match calculation, after excluding the unfragmented precursor in ETD mode. Here:

$$s_3 = \frac{Top10}{10}$$

IV. Poisson Probability—The probability based scoring strategy determined if the predicted match between the experimental data and candidate glycopeptide is a chance event (33, 34). For this, a set of "decoy" glycopeptides were generated. This was done by randomly scrambling the peptide sequence in the glycopeptide, and at the same time arbitrarily adding or subtracting a molecular mass between −50 to +50 for each monosaccharide while keeping the total glycan mass unaltered (see Fig. 4 for example). Twenty-five such decoys were generated for each candidate glycopeptide. The theoretical fragment ions for individual candidate and decoy glycopeptides was then compared with the experimental spectrum. The p value for the candidate was then computed by:

$$P(K_1, p) = \frac{(N_1 p)^{K_1}}{K_1!} e^{-N_1 p};\ p = K/N$$

where K_1 and N_1 denote the number of matched and total number of fragment ions for the candidate glycopeptide. K and N denote corresponding values for the entire database that includes both candidate and decoy glycopeptides. Low p values indicate a better match. Thus:

$$s_4 = \begin{cases} 0 & \text{if } P\ value > 2 \times 10^{-2} \\ 1 - \frac{(\log(P\ value) - \log(10^{-5}))}{(\log(2 \times 10^{-2}) - \log(10^{-5}))} & \text{if } 10^{-5} < P\ value < 2 \times 10^{-2} \\ 1 & \text{if } P\ value < 10^{-5} \end{cases}$$

Ensemble Score (ES)—The above four parameters were weighted according to the following equation in order to arrive at an ensemble score: $ES = s_1\cdot w_1 + s_2\cdot w_2 + s_3\cdot w_3 + s_4\cdot w_4$. Although the individual weights can be varied depending on the fragmentation modes, the following were the settings for the current manuscript: i. CID mode, $w_1 = w_2 = w_3 = w_4 = 0.25$; ii. HCD mode, $w_1 = w_2 = w_4 = 1/3$; $w_3 = 0$; iii. ETD, mode, $w_1 = 0.2$; $w_2 = 0$; $w_3 = 0.1$; $w_4 = 0.7$. In this regard, the CID mode equally weights all the scoring parameters. HCD excludes Top10 as the high-peaks in the MS/MS spectrum are often dominated by oxonium ions which do not inform glycopeptide identification. ETD has a high probability-based weighting because glycopeptide fragmentation is often incomplete, with a large precursor peak remaining.

Glycopeptide False Discovery Rate (FDR) Calculation and Structure Assignments—The *ES* for the candidate glycopeptide was generated as described in the previous section. For each candidate glycopeptide, a "decoy glycopeptide" was also generated by scrambling the peptide backbone and adding/subtracting up to 50Da mass to each

monosaccharide while conserving the overall glycan mass. The *ES* for this decoy is called ES_{decoy}. The glycopeptide FDR at any $ES_{\text{cut-off}}$ value was then calculated using:

$$\text{FDR}(ES_{cut\text{-}off}) = 100\frac{\text{\# decoy glycopeptides with } ES_{decoy} > ES_{cut\text{-}off}}{\text{\# candidate glycopeptides with } ES > ES_{cut\text{-}off}}$$

To confirm assignments in this manuscript, $ES_{cut\text{-}off}$ was determined for glycopeptide FDR = 1%. All candidate glycopeptides with $ES > ES_{cut\text{-}off}$ were then manually inspected using the Browse Results GUI (*browsegui*) of GlycoPAT. This program presents an annotated MS/MS spectrum detailing the assignments that could be made for the candidate glycopeptide, including an ion map summarizing the identified hits and a more elaborate table showing the details of each assignment. Thus, all assignments were manually inspected and validated.

Comparison with SEQUEST and Byonic—Some studies compared proteomics analysis results from GlycoPAT with a similar single protein search result from SEQUEST (32). Both programs analyzed tryptic digested peptides with two missed cleavages, 20 ppm MS^1 peptide tolerance, 1Da MS/MS fragment ion tolerance, 3 maximum PTMs, variable methionine oxidation, variable serine/threonine/tyrosine sulfation, and fixed cysteine carboxyamidomethyl. The GlycoPAT ensemble score (*ES*) was calculated using default program parameters described above. In SEQUEST (1.4.1.14), after scoring, standard filters were applied (Xcorr $\geq$ 1.5, 2.0, 2.5, 3.5 for z = 1, 2, 3, 4+, respectively and peptide probability $\leq$ 0.05) to determine acceptable matches in CID fragmentation mode.

For comparison between GlycoPAT and Byonic (17), raw HCD MS/MS data for tryptic digested Basigin/CD147 was downloaded from PRIDE (data set identifier: PXD004243) (35). Identical settings were used for both programs: precursor mass tolerance = 10 ppm, MS/MS tolerance = 20 ppm and cysteine carbamidomethylation fixed modification. All variable modifications (methionine oxidation, asparagine/glutamine deamidation, and N-glycan library from (35)) were set to "common 1" in Byonic (version 2.10.47). Maximum variable modifications = 2 and missed cleavage = 2 in both programs. Tryptic digested bovine fetuin MS/MS data from LTQ-Fusion was analyzed in CID mode. Here, precursor mass tolerance = 10 ppm; MS/MS tolerance = 1 Da; fixed modification = cysteine carbamidomethylation; variable modification = methionine oxidation; "standard" N-Glycan library described in Results.

Proteomics and Glycoproteomics Analysis of Human Plasma Cryoprecipitate—Proteomics analysis was performed using Proteome Discoverer™ (2.1) embedded with search engine SEQUEST HT to identify proteins in two cryoprecipitate samples digested with trypsin and two more samples that were Glu-C digested. These samples were subjected to the above described LC-MS/MS (HCD) experiments. The search parameters were: MS^1 tolerance = 10 ppm, MS^2 tolerance = 0.20 Da, fixed carboxyamidomethyl modification, variable methionine oxidation, max missed cleavage = 2 for trypsin and = 3 for Glu-C. The reviewed Swiss-Prot human FASTA database of "Uniprot Release 2015_01" (Homo sapiens subset with forward-decoys) was used for the search. These analysis results were then combined using Scaffold (version 4.4.3, Portland, OR) using 0.1% peptide decoy FDR and 4.3% protein FDR. GlycoPAT search was then conducted on the *N*-linked glycopeptides of the top 7 proteins identified in this manner. The GlyDB search library used here was identical to the "standard N-glycan GlyDB" described in Results, only it also included O-type blood group antigen terminal modifications.

Glycoproteome of Prostate Cancer Cells—LNCaP and PC-3 prostate cancer cell line glycoproteomics data were downloaded from ProteomeXchange (PXD002107) (36). Twenty-two of the 24 .RAW files could be converted to mzXML. The monoisotopic mass corresponding to each MS/MS product was determined using the "averagine" method (37), using trypsin digested fetuin glycopeptide results presented in this publication as a model. Briefly, we calculated the molecular composition of each of the identified fetuin glycopeptides (*i.e.* $C_aH_bN_cO_dS_e$), and determined the average unit glycopeptide composition by dividing by the overall molecular mass. The isotopic distribution of this typical glycopeptide was determined using the Bioinformatics toolbox of MATLAB. Next, for each experimental MS/MS spectrum, we determined the precursor isotopic distribution by adding the local MS^1 spectra ($\pm$4Da) surrounding the parent ion in a 1 min chromatographic window ($\pm$ 30 s) (13). The monoisotopic peak was then determined at 10 ppm resolution by translating the experimental MS^1 distribution over the distribution of the theoretical unit glycopeptide scaled based on the parent ion mass, and determining the position at which the cross-correlation was maximum.

Once the experimental monoisotopic mass was determined, each of the MS files with ~40,000 spectra were searched against a theoretical GlyPep library with 429,841 members using a 36-core cluster (Intel Xeon-E5645 processor, 12 cores per node, 3 nodes). This library was generated using the 1793 unique peptides reported using the SPEG (Solid Phase Extraction of Glycopeptides) method (36), and 172 unique glycan masses including high mannose, bi-, tri-, and tetra-antennary structures, core- and terminal fucosylated carbohydrates, and sialylated glycans similar to previous work (38). Additional variable modifications included methionine oxidation. Only one N-glycan was permitted on each glycopeptide, whereas there was no limit on the number of oxidation sites. Fixed modifications included cysteine carbamidomethylation, and iTRAQ labeling (114, 115, 116, 117) at N terminus and lysine. For search parameters, tolerance for MS^1 and MS^2 was 10 ppm and 0.06 Da. All other parameters used were program defaults. During the final data processing steps, the candidate with highest ES was selected, provided ES$\geq$ 0.5. Additionally, all accepted results had at least two glycan oxonium ions, and it was verified that the intensity of the highest oxonium ion exceeded the intensity of the iTRAQ reporter. All ES results were compared with Byonic scores reported previously (36).

RESULTS

SmallGlyPep (SGP1.0)—The SGP1.0 nomenclature is designed for the minimal representation of glycopeptides in linear text format for MS based glycoproteomics data analysis (Fig. 1 supplemental Movie SA). Its design enables the straightforward *in silico* MS^n fragmentation of glycopeptides at one or more locations that may reside either on the peptide backbone or glycan/nonglycan PTMs. Here, upper and lowercase letters represent amino acid and PTM modifications, respectively. Glycan PTMs are described within braces or curly brackets. Nonglycan PTMs are enclosed within chevrons or angle brackets. The list of currently available monosaccharides and nonglycan modifications in GlycoPAT are provided in supplemental Table S1. Additional members can be added by modifying class definitions as described in the software manual. In addition to single letters, arbitrary monosaccharides can also be represented by numbers corresponding to their molecular mass. During the representation of glycans, each monosaccharide is enclosed within a single pair of braces, with the open bracket ("{") just prior to the residue representing the glycosidic bond that links it to the rest of the molecule and the paired closing bracket ("}") indicating the nonreducing end of the antenna on which this monosaccharide resides.

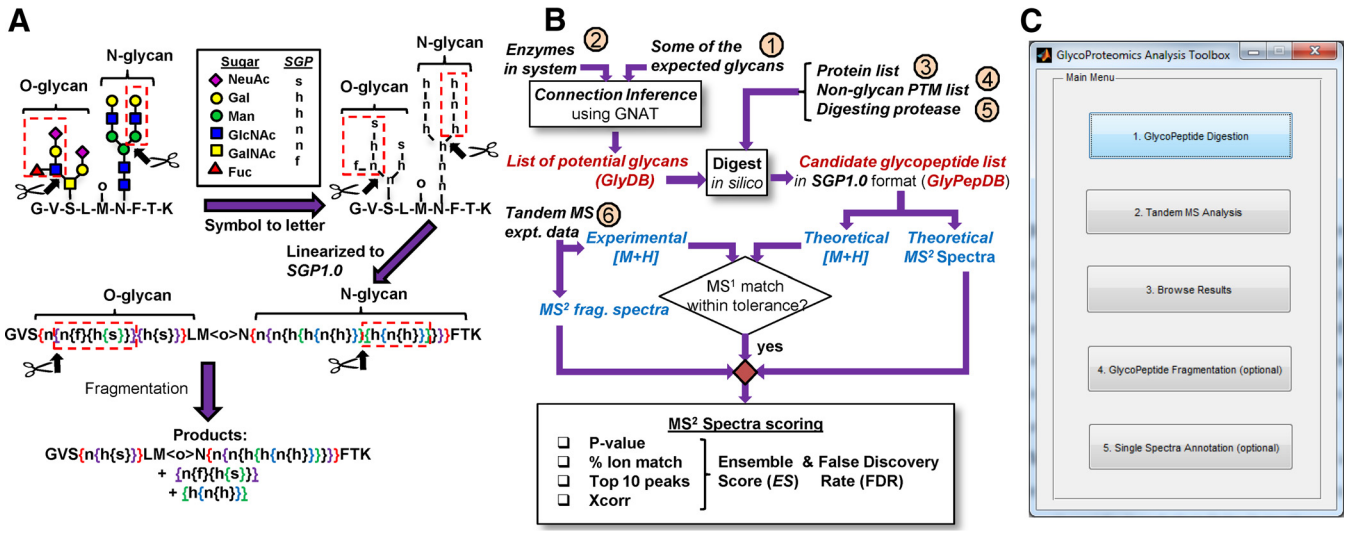
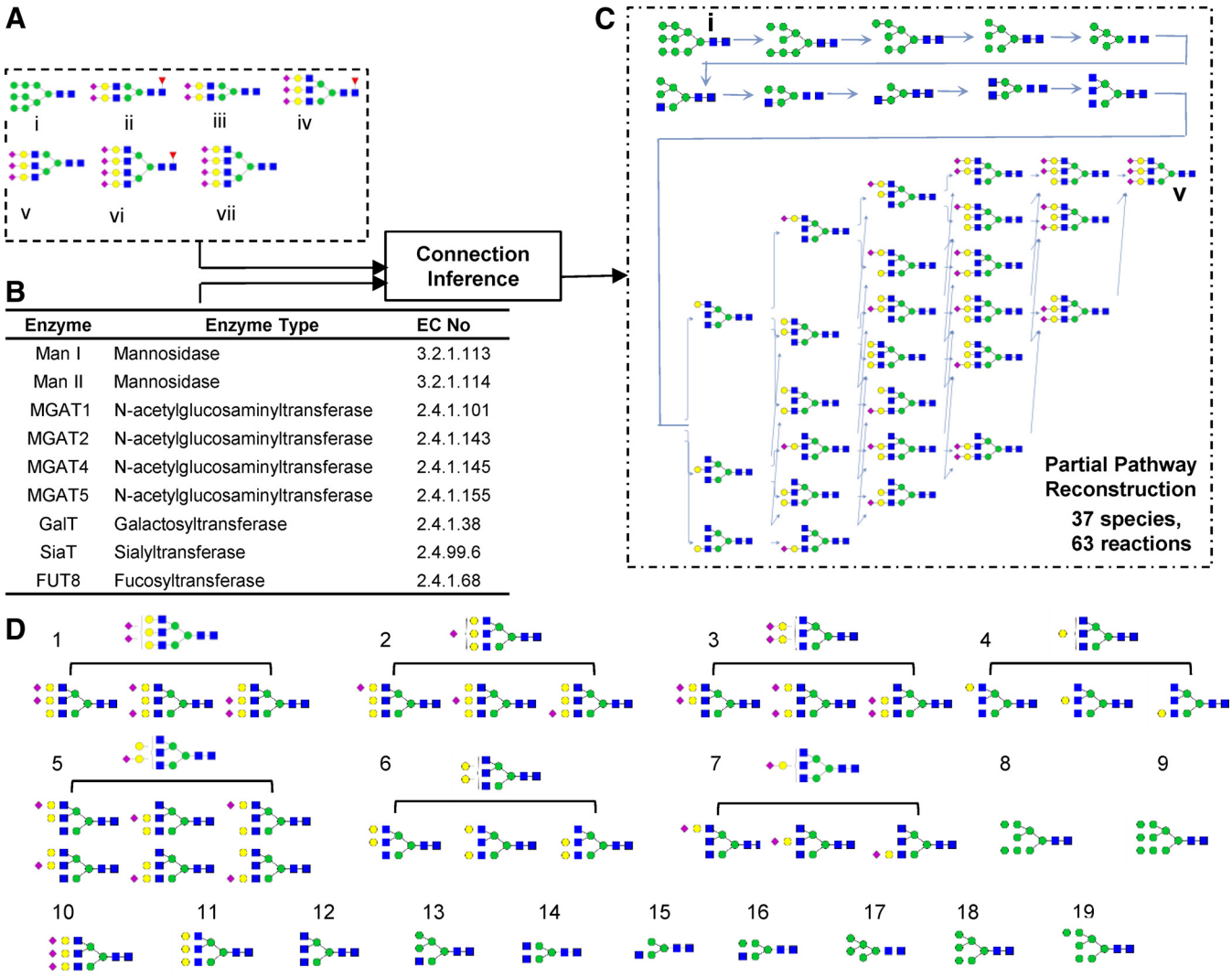


FIG. 1. **SmallGlyPep (SGP1.0) nomenclature and overall data analysis scheme.** *A*, Symbol Nomenclature for Glycans (SNFG) is used to represent a glycopeptide containing an O-glycan, N-glycan and nonglycan PTM (top-left). Individual monosaccharides are transformed to lowercase letters in top-right schematic, with single-letter glycan nomenclatures shown in legends. This glycopeptide is linearized into *SmallGlyPep* (SGP 1.0) nomenclature by introducing curly brackets to enclose glycan PTMs (middle). Here, the open bracket denotes a glycosidic linkage and closed bracket represents the end of the corresponding glycan antenna. Nonglycan PTMs are enclosed by angle brackets (*e.g.* <o> for oxidation). Fragmentation at a particular site, represented by scissors at an open bracket, results in the release of glycans within paired curly brackets (shown using dashed red boxes in all figures). Products formed by simultaneous fragmentation of both the O- and N-glycan appear at the bottom. *B*, Overall program schematic shows 6 inputs enclosed in small circles. Inputs 1 and 2 feed into a "*connection inference*" routine to yield the glycan search library (GlyDB), a list of potential O- and N-glycan PTMs to search for. The theoretical protein(s) (input 3) is decorated with these glycans and additional fixed/variable nonglycan PTMs (input 4). Digestion of this glycoprotein(s) by protease(s) (input 5) results in the "theoretical glycopeptide database (GlyPepDB)." MS^1 precursor mass comparison along with MS^2 fragmentation spectra matching yields various scores: Xcorr, % ion match, *p* value and Top10 peaks. The ensemble score (*ES*) is an overall weighted average of these individual scores. False discovery rate (FDR) calculations are used to identify the minimum acceptable *ES* or $ES_{cut\text{-}off}$ value. *C*, The main Graphical User Interface (GUI) for GlycoPAT. Clicking individual buttons brings up additional GUIs shown in supplemental Fig. S3. GlycoPAT functions can also be executed using MATLAB command line operations and scripts as described in the software manual.

For illustration, a glycopeptide with one N-glycan, one core-2 O-glycan and one nonglycan PTM is shown in Fig. 1*A*. To convert this molecule from the conventional Symbolic Nomenclature for Glycans (39) to SGP1.0, the monosaccharides are represented by single letters with hexose, N-acetylhexosamine, N-Acetylneuraminic acid (sialic acid) and fucose being annotated by "h," "n," "s" and "f," respectively (top of Fig. 1*A*). Curly bracket pairs, color coded in Fig. 1*A*, are then introduced for linearization with the open and closed brackets bracing the carbohydrate arm containing the monosaccharide(s). Thus, the number of curly bracket pairs equals the number of monosaccharides. Fragmentation of a glycosidic bond results in the release of the glycan fragment enclosed within paired curly brackets (shown using red dashed boxes, Fig. 1*A*). Thus, the SGP1.0 nomenclature enables streamlined design of algorithms for *in silico* glycopeptide fragmentation at multiple sites. This is necessary for MS^n data analysis.

GlycoPAT Workflow and Graphical-user-interface (GUI)—Using SGP1.0 as the foundation, a suite of functions was written in MATLAB for tandem-MS glycoproteomics data analysis (workflow in Fig. 1*B*), including GUIs for MS/MS experiments (Fig. 1*C*). The full program includes ~13,000 lines of code and additional libraries.

In this workflow, first, the glycan search database (or "GlyDB") is designed by either manually listing the glycans in text input files or generating them automatically using the "*connection inference*" algorithm described previously (23, 40). Fig. 2 illustrates the latter case, using an input set of 7 seed glycans and 9 enzymes to generate the "standard N-glycan GlyDB." Here, the seed glycans include one high-mannose structure that initiates glycan biosynthesis, and terminal bi-, tri- and tetra-antennary sialylated structures both with and without core fucose (Fig. 2*A*). Among the enzymes (Fig. 2*B*), the mannosidases (Man I and Man II) trim high-mannose structures, N-acetylglucosaminyltransferases (GnT-I, -II, -IV, -V) enable N-glycan branching, and the remaining enzymes either extend or decorate the glycan terminus. Enzyme specificity rules are presented based on previous class structures (supplemental Table S2, (23, 40)). Using this algorithm, the network linking the starting high mannose glycan (species i) to the tri-antennary glycan (species v) contains 37 automatically generated glycans and 63 reactions. Grouping isomeric glycans with identical monosaccharide compositions and similar fragmentation patterns reduced the glycan number in GlyDB from 37 (Fig. 2*C*) to 19 prototypic structures (Fig. 2*D*). Similarly, the full glycosylation network connecting



Enzyme	Enzyme Type	EC No
Man I	Mannosidase	3.2.1.113
Man II	Mannosidase	3.2.1.114
MGAT1	N-acetylglucosaminyltransferase	2.4.1.101
MGAT2	N-acetylglucosaminyltransferase	2.4.1.143
MGAT4	N-acetylglucosaminyltransferase	2.4.1.145
MGAT5	N-acetylglucosaminyltransferase	2.4.1.155
GalT	Galactosyltransferase	2.4.1.38
SiaT	Sialyltransferase	2.4.99.6
FUT8	Fucosyltransferase	2.4.1.68

FIG. 2. **Generation of candidate glycan list for the standard N-glycan GlyDB library.** *A*, Seven seed glycans were provided as inputs for the *"connection inference"* program. *B*, Glycosylating enzymes used for pathway synthesis. Detailed enzyme rules and properties are defined in supplemental Table S2. *C*, Network diagram generated using one pair of input seed glycans (glycans i and v in panel *A*). *D*, Categorization of all glycans in panel c pathway into 19 groups using structure-based isomer classification. One representative glycan from each group is a member of the glycan database, GlyDB. Full network using all seed glycans and enzymes are shown in supplemental Fig. S1 (.png file). This includes 596 reactions and 250 *N*-linked glycans.

all 7 seed glycans (Fig. 2*A*) has 250 unique glycans and 596 reactions (supplemental Fig. S1). Clustering these glycans reduced the "standard N-glycan GlyDB" to 75 members. Similar to this, a "standard O-glycan GlyDB" was generated with 15 species using seed glycans and enzymes in supplemental Table S3. The full reaction network is illustrated in supplemental Fig. S2. The GlyDB library generated in Fig. 2 can be readily expanded by including additional glycosylation related enzymes, as illustrated later in the study of prostate cancer cells. Alternatively, it may also be generated using glycomics-based MS profiling studies or using curated organism-specific glycan databases.

The "theoretical glycopeptide database" or GlyPepDB, which contains the list of potential glycopeptides in the system, was generated in SGP1.0 format using: (1) GlyDB from the previous step, (2) list of protein inputs provided in FASTA format (input 3, Fig. 1*B*), (3) fixed and variable nonglycan PTMs (input 4, Fig. 1*B*), and (4) protease(s) used for digestion (input 5, Fig. 1*B*). To limit the size of GlyPepDB, GlycoPAT has facilities to limit the maximum and minimum number of variable PTMs on any peptide, and stipulate specific protein backbone locations (*e.g.* the 55th and 68th amino acid) where variable modifications may occur. Such facilities are important to limit the combinatorial expansion of GlyPepDB and focus the specific search.

To determine which of these glycopeptides from GlyPepDB are present in the sample, GlycoPAT first matches each experimental MS^1 mass to the precursor mass of the GlyPepDB

members. Once a "candidate glycopeptide" is identified, a score is generated to relate the corresponding experimental MS^n spectrum with the theoretical spectrum generated by *in silico* fragmentation of the candidate GlyPepDB member. This ensemble score (*ES*) weighs various statistical parameters: (1) X_{corr}: "Cross correlation" score; (2) % ion match: The percentage of ions that are matched between the theoretical and observed spectrum; (3) Top 10 peaks: Number of 10 most intense experimental peaks that matched the theoretical ions; and (4) *p* value: The probability based on the generation of a set of glycopeptide decoys.

The GlycoPAT software is freely available as an open-source, platform-independent toolbox. GUIs are currently available to implement core functions (Fig. 1*C*, supplemental Fig. S3, and Movie SB): (1) Creation of a theoretical glycopeptide database, (2) Searching and scoring experimental MS/MS data against the theoretical database, (3) Browsing of final results including spectrum annotation, (4) Reanalysis of a single MS^n spectrum to determine how changing the fragmentation parameters impacts the peaks matched and *ES* score, and (5) A calculator for the theoretical fragmentation of glycopeptides. Additional command line functions for more complex operations can also be implemented in the software, as explained in the GlycoPAT manual.

Analysis of Single Glycoproteins—Fig. 3 presents data confirming the ability of GlycoPAT to identify standard proteins (fetuin, asialofetuin and RNaseB) in different fragmentation modes. These spectra were identified using the "standard O- or N-glycan GlyDB" described above. Here, Fig. 3*A*–3*C* compares the MS/MS fragmentation patterns of *N*-linked glycans from fetuin in HCD, CID, and ETD modes. Consistent with previous work (8), the HCD spectrum contains abundant low *m/z* peaks corresponding to Hex (*m/z* = 204), Neu5Ac (*m/z* = 292), HexHexNAc (*m/z* = 366), and cross-ring fragments of monosaccharides and water loss (*m/z* = 138, 167, 185) (Fig. 3*A*). At high *m/z*, Y-ions corresponding to the peptide backbone cleavage with short glycan stubs are also evident. CID lead to less extensive fragmentation of sialylated tetra-antennary glycopeptides compared with HCD. Thus, the oxonium B-ions were less intense compared with HCD, and high molecular mass peaks with a loss of either Neu5Ac or Neu5AcHex were abundant (Fig. 3*B*). Whereas the B-ions had *z* = 1, the Y-ions included both neutral loss and loss of charge peaks. The presence of core-fucosylation was evident based on the diagnostic ion at *m/z* = 1381. Fragmentation in ETD mode led to *c/z*-ions though the efficiency of fragmentation was low with a large precursor ion peak remaining in the MS/MS spectra (removed from Fig. 3*C*). A few *b*- and *y*- ions were also noted, likely because of the use of supplemental activation (a low energy HCD) to dissociate the charge-reduced species. These data confirm the ability of GlycoPAT to analyze glycopeptide fragmentation data in three fragmentation modes.

In addition to the above, GlycoPAT also identified other glycan-types including asialoglycopeptides from asialofetuin (Fig. 3*D*), high mannose glycoconjugates from RNAseB (Fig. 3*E*) and O-linked glycopeptides from fetuin (Fig. 3*F*), all using CID fragmentation. Among these, the nonsialylated tri-antennary glycan in Fig. 3*D* displayed a range of B- (*m/z* = 366, 528) and Y-ions. Fragmentation of the high-mannose (Man 8) glycan located at Asn80 of RNaseB resulted in a ladder pattern because of successive loss of one to seven mannose residues in products with *z* = 3–5 (Fig. 3*E*). The CID mode fragmentation of O-linked glycopeptides resulted in a pattern similar to the *N*-linked glycopeptide with B-ions (*m/z* = 366, 657) and Y-ions because of the loss of Neu5Ac, Neu5AcHex and Neu5AcHexHexNAc (Fig. 3*F*). Additionally, a small portion of the peptide backbone was also fragmented, resulting in selected *b*- and *y*- ions.

Ensemble Score and Decoy-based Strategy for Controlling False-positives—It is necessary to determine the minimal *ES* or $ES_{cut\text{-}off}$ value, which identifies high-quality spectrum matches with minimal false-positive hits. Although GlycoPAT has facilities to implement both global and local glycopeptide FDR calculations to determine this $ES_{cut\text{-}off}$, the global approach is illustrated in Fig. 4. Here, one decoy glycopeptide is first generated for each "candidate glycopeptide" that had an MS^1 match. This is generated by scrambling the base peptide sequence and randomly adding or subtracting a mass of up to 50 Da to each monosaccharide while keeping the total mass of each glycan constant (Fig. 4*A*). Alternate methods are also possible in GlycoPAT though they were not used in this manuscript. These methods for generating peptide decoys include swapping the first 1–2 amino acids, or reversing the amino acid sequence (Fig. 4*A*). Overall, the approach used for creating decoy glycopeptides is more comprehensive compared with prior work that either only scrambled the peptide or added a fixed mass to the glycan (41).

The GlycoPAT function names and overall strategy to create decoy glycopeptides are presented in Fig. 4*B*, along with one example in Fig. 4*C*. Here, the decoy monosaccharides are represented using numbers corresponding to the mass of the decoy, because monosaccharides in GlycoPAT can be defined using either predefined single letter nomenclature or molecular mass. In this manner, *ES* is calculated for each "candidate glycopeptide" and its corresponding decoy (Fig. 4*D*). The global FDR at any *ES* is then defined based on the ratio of the number of decoy glycopeptides having scores > *ES* compared with that for the candidate glycopeptides.

Fig. 4*E*–4*G* presents an example of glycopeptide FDR calculation for trypsin-digested bovine fetuin in CID mode. In Fig. 4*E*, a cumulative *ES* score plot is shown for ~850 candidate glycopeptides and their corresponding decoys. As expected, the candidate glycopeptides have a higher *ES* score. Here, at *ES* = 0.2, global glycopeptide FDR equals ~18.8% (= 150/800 × 100). Fig. 4*F* presents the same data following calculation of glycopeptide FDR at each *ES* value. As seen, glyco-

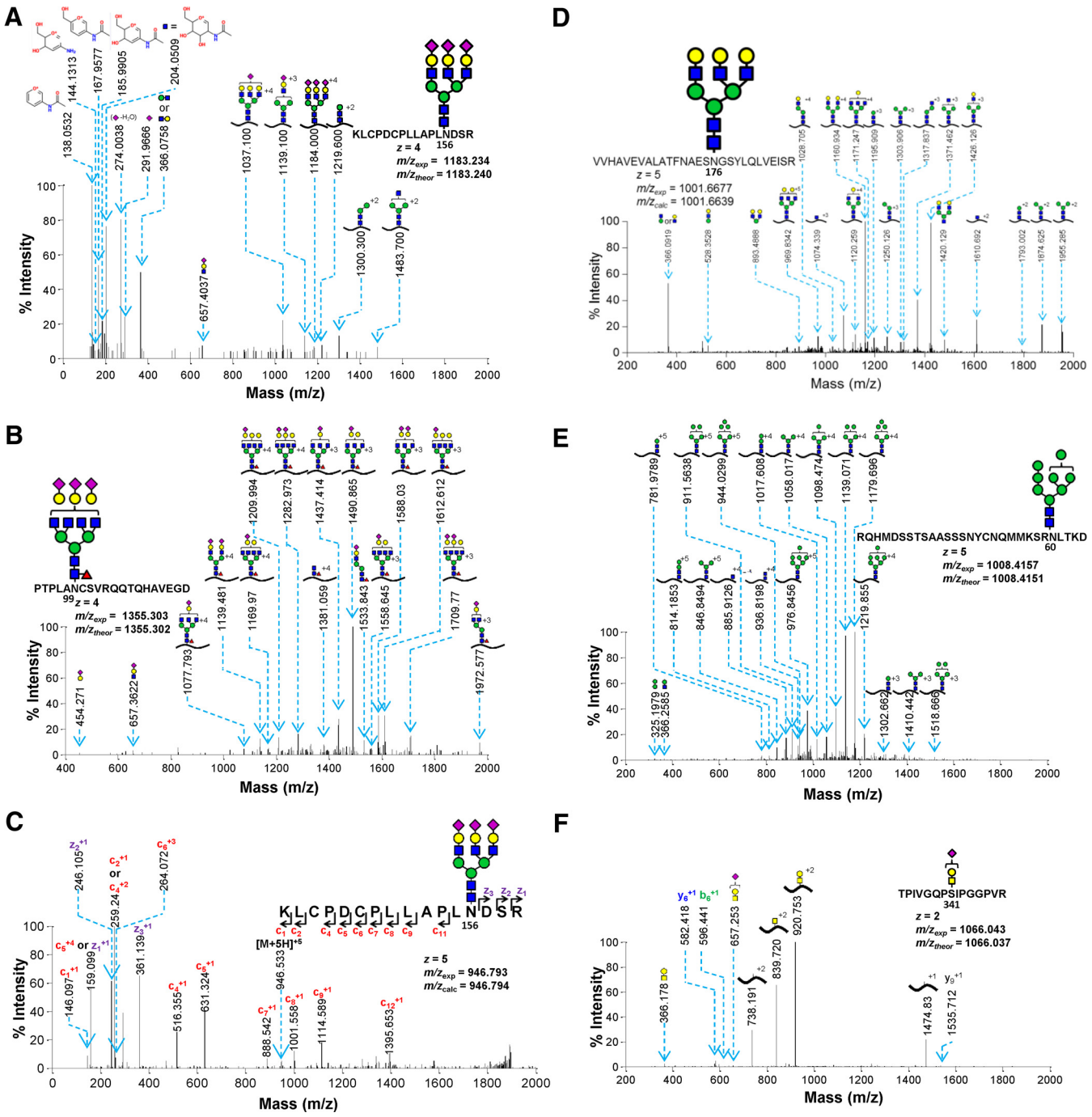


FIG. 3. **Annotated examples of MS/MS spectra with different type of glycopeptide assignments or under different fragmentation modes.** *A*, HCD mode fragmentation of fetuin glycopeptide; *B*, CID mode example of core-fucosylated *N*-linked glycopeptide from fetuin; *C*, ETD mode N-glycopeptide of fetuin; *D*, Glycopeptide from asialofetuin; *E*, High mannose structure from RNase B; *F*, sialyated T-antigen glycopeptide of fetuin. Theoretical and experimental MS^1 *m/z* values are provided in figure. Trypsin was used for digestion in all panels, except panels *B* and *E* which used Glu-C.

peptide FDR increases upon relaxing the *ES* value (Fig. 4*F*). To set a conservative selection criteria for minimizing false-positives, the current manuscript uses a 1% FDR. In this example, this corresponds to an $ES_{cut\text{-}off}$ value of 0.53 (see inset). Fig. 4*G* presents the relation between glycopeptide FDR cut-off values and number of candidate glycopeptide spectra identified as true hits. Here, relaxing FDR increases the number of accepted spectra (Fig. 4*G*). In the fetuin example, 530 of the 850 candidate spectra had $ES_{cut\text{-}off}>0.53$ and FDR<1% (see inset). Because many of these spectra corresponded to the same glycopeptide, the actual number of fetuin glycopeptides identified is smaller.

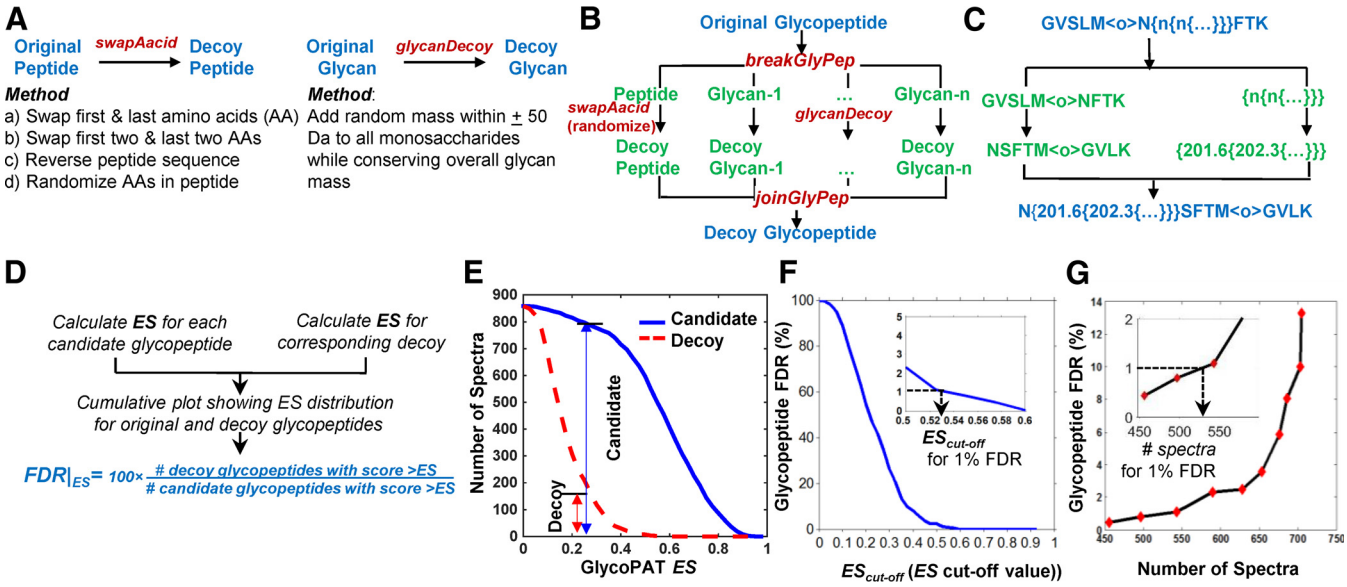


FIG. 4. **Decoy-based strategy for glycopeptide assignment.** *A*, Methods for the generation of decoy peptide and glycan. *B–C*, Procedure for decoy glycopeptide construction (panel *B*) along with an example (*C*). Original glycopeptide is in SGP1.0 format. Monosaccharides in the decoy (±50Da of the original mass) are represented using numbers, instead of single monosaccharide letters. The overall mass of the full glycan is conserved. *D*, Flowchart for glycopeptide FDR calculation. *E–G*. Examples of cumulative plots are shown for *ES* distribution for original and decoy glycopeptides for the case of fetuin MS data (panel *E*), the relationship between glycopeptide FDR and GlycoPAT $ES_{cut\text{-}off}$ value (*F*), and the relationship between glycopeptide FDR and the number of peptide (glycopeptide) spectrum matches (*G*). Inset in panels *F* and *G* expand the region around 1% FDR.

Comparing GlycoPAT Scoring with SEQUEST and Byonic—The scoring results using GlycoPAT were compared with two popular commercial software, SEQUEST (32) and Byonic (17) (Fig. 5). Whereas SEQUEST is dedicated to MS based proteomics data analysis, Byonic extends the classical proteomics methods for glycopeptide analysis. The first two panels compare the proteomics spectra of these programs by comparing the GlycoPAT *ES* for a single fetuin MS run with equivalent metrics in SEQUEST (Fig. 5*A*) and Byonic (Fig. 5*B*). As seen, most of the assignments with $ES > ES_{cut\text{-}off}$ (0.54) have both high SEQUEST *Sf* score (>0.6, Fig. 5*A*) and high Byonic scores (>200, Fig. *B*. Additionally, some assignments with *ES*< 0.54 also have high Byonics and/or SEQUEST *Sf* scores, consistent with the conservative practice of GlycoPAT *ES* to limit the number of true-positive spectra based on low glycopeptide FDRs.

Upon comparing the glycoproteomics score using GlycoPAT with Byonic, differences were evident because the scoring criteria are not similar (Fig 5*C*, 5*D*). This is because GlycoPAT considers the extensive fragmentation of glycans during scoring, whereas Byonic (in HCD and CID modes) primarily only considers the oxonium ions, nonglycosylated peptide, peptide plus core HexNac (+ core fucose, if present) and loss of sialic acid (42). In this regard, there is reasonable agreement between both programs when scoring HCD MS/MS spectra because this is dominated by the peptide Y_0-ion and small glycopeptide stubs (Fig. 5*C*). Spectra in quadrant-I score better in Byonic (score>175) because it considers water-losses and some larger glycopeptides that are not considered by GlycoPAT. These peaks, that are unique to Byonic, are indicated by red arrows in Fig. 5*E*. GlycoPAT scores are higher in quadrant-IV because the SGP1.0 nomenclature enables the simultaneous fragmentation of both the glycan and peptide backbone. In Fig. 5*F*, several such peaks with simultaneous glycan and peptide breaks are evident (green arrows in Fig. 5*F*). Another example spectrum from quadrant-IV is also shown in supplemental Fig. S4, with the raw output window from Byonic and GlycoPAT, contrasted with manual spectrum annotation.

The importance of considering extensive glycan fragmentation is very clear when considering CID data analysis, where the breakage of glycosidic bonds dominants the spectrum (Fig. 5*D*). Here, several spectra in quadrant-II were identified to be good hits by both programs with *ES*>0.47 and Byonic score>175 (example in Fig. 5*G*). The glycopeptides in quadrant IV had high GlycoPAT but low Byonic scores. As seen in the representative spectrum in Fig. 5*H*, this is because GlycoPAT comprehensively identifies both the glycan *B*-ions (*m/z* = 366.2, 657.4) and *Y*-ions corresponding to progressive monosaccharide losses.

Although some representative spectra are shown in Fig. 5, the conclusions drawn here were generally true for at least three different runs performed in HCD mode on a Thermo Q Exactive instrument, and 8–10 runs performed in CID mode on Thermo LTQ-Fusion. GlycoPAT annotated spectra for additional quadrants are provided in supplemental Fig. S5 for HCD data, and supplemental Fig. S6 for CID. Overall, these results highlight the importance of considering both full glycan

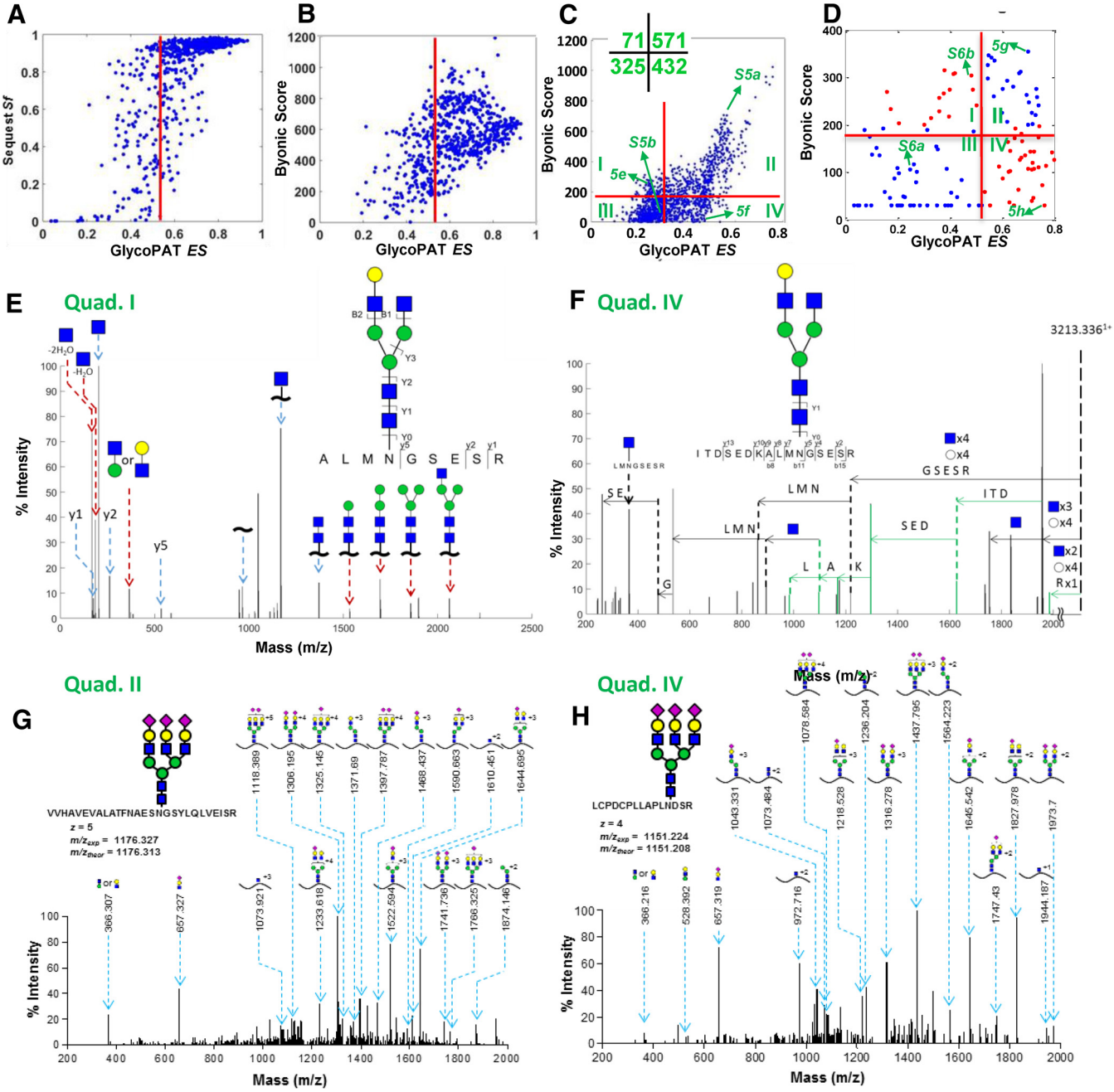


FIG. 5. **Comparison of GlycoPAT, Byonic and SEQUEST for proteomics and glycoproteomics analysis.** *A*, Plot compares the SEQUEST final score (*Sf*) and GlycoPAT ensemble score (*ES*) for trypsin digested peptides of fetuin. Each point represents a file analyzed using both GlycoPAT and SEQUEST. There is good agreement between both software for proteomics studies. *B*, Byonic score compared with GlycoPAT *ES* for trypsin digested fetuin peptides (*i.e.* pure proteomics analysis). Good agreement is seen between the two programs. *C*, Comparison of Byonic *versus* GlycoPAT scores for HCD-mode MS/MS data obtained from trypsin-digested Basigin/CD147 glycopeptides. Spectrum in quad.-I received higher Byonic scores whereas those in quad.-IV had higher GlycoPAT *ES*. Number of spectra in each quadrant is enumerated in inset. *D*, Byonic score *versus* GlycoPAT *ES* for CID mode MS/MS spectra obtained using trypsin cleaved glycopeptides of fetuin. Scores are in agreement in quad.-II and -III only. *E* and *F*, Annotated HCD-mode MS/MS spectrum for selected files from Quad-1 and -IV of panel *C*. Peaks uniquely identified by Byonic are marked using red arrows (panel *E*, Quad-I). Peaks unique to GlycoPAT because of simultaneous glycan and peptide backbone fragmentation are in green (panel *F*, Quad-IV). *G* and *H*, Annotated CID-mode MS/MS spectrum using GlycoPAT. Good agreement is seen for both Quad-II (high *ES*, Byonic score, panel *G*) and Quad-IV (high *ES*, low Byonic score, panel *H*). GlycoPAT analyzes the ladder pattern of products formed following serial glycan fragmentation. Byonic, however, does not score glycan fragmentation.

TABLE I
Glycoproteomics analysis summary

Sample type	Protein name	N-glycosylation sites	Number of unique glycan structures	Number of unique glycopeptides identified
Single proteins	Asialofetuin	3	4	8
	Fetuin	3	23	41
	RNase B	1	6	6
Four-protein mixture	Fetuin	3	8	8
	AGP-1	3	14	17
	Fibronectin	9	21	42
	RNase B	1	5	5
Plasma cryoprecipitate	A2MG	3	4	4
	FIBA	3	3	3
	FIBB	1	2	2
	FIBG	1	7	7
	FINC	7	10	12
	PLMN	1	1	1
	VWF	9	17	17
Prostate cancer cells: LNCap& PC-3	Human proteome	230	52	960

fragmentation, and simultaneous glycan and peptide fragmentation during glycoproteomics scoring.

Analysis of Single Proteins, Simple Mixtures and Human Blood Plasma Cryoprecipitate—Table I summarizes all the N-glycans in the "standard N-glycan GlyDB" that were identified for three single standard proteins (fetuin, asialofetuin, RNase B) and also mixtures of standard proteins (fetuin, AGP-1, fibronectin plus RNase B). These runs were performed following digestion using either trypsin or Glu-C in CID fragmentation mode. They did not apply either chromatography methods to enrich for glycopeptides or HCD product-dependent strategy to select them (43). A 1% glycopeptide FDR cut-off criterion was used for the initial screen followed by manual validation of each spectrum. During such validation, multiple glycans were grouped using curly brackets when the observed glycan fragmentation pattern was consistent with more than one member of GlyPepDB. Unique structural assignments were also possible in some cases. supplemental Material provides the detailed structures identified for the single proteins (supplemental Table S4) and mixtures (supplemental Table S5), along with individual annotated spectra in jpeg and MATLAB .fig formats.

In the single protein study (supplemental Table S4), N-glycans were identified at the single N-glycosylation site of RNase B (Asn60) and all three putative sites of fetuin and asialofetuin (Asn 99, 156, and 176). Consistent with previous reports (44–49), the glycans of fetuin included sialylated bi-, tri-, and tetra-antennary carbohydrates both with and without fucose. Fucosylated N-glycans are annotated in supplemental Fig. S7*A*. As seen, the bi-antennary glycans are preferred at Asn99 and Asn156 but not Asn176. supplemental Fig. S7*B* presents the retention time profiles of various fetuin glycopeptides generated by trypsin digestion. Here, the sialylated bi-antennary N-glycan on Asn156 eluted first at 47 min, followed by sialylated bi- and tri-antennary structures at Asn156 and finally the larger tri- and tetra-antennary glycans of Asn99 and Asn176. Glycans identified on asialofetuin (supplemental Table S4) are similar to normal fetuin, only they lack sialic acids. High mannose structures dominate RNaseB as previously reported (50), with one additional hybrid N-glycan (Man3GlcNAc).

In the glycoprotein mixture with four components (supplemental Table S5), 15 N-glycosylation sites were identified including 2 of 3 putative sites on fetuin, 3 of 5 on AGP-1, 9 of 11 on fibronectin and the single site on RNase B. In the case of AGP-1, we detected 14 of the 22 bi-, tri- and tetra-antennary *N*-linked glycans reported in a previous glycomics profiling study (51). All glycans were sialylated, with glycan structural diversity being greatest at amino acids 56 and 103. The Man5 and Man6 high mannose structures, which are most abundant in RNAseB (52), were measured in the mixed sample. Additionally, complex structures were observed, which is consistent with an earlier multiple-laboratory collaborative investigation (52). The study of fibronectin also revealed the presence of many of the bi- and tri-antennary carbohydrates reported earlier (53), along with additional tetra-antennary structures that are thought to be elevated following oncogenic transformation (54). The current study reports N-glycans at all 7 fibronectin-sites reported previously (53), along with additional glycosylation at amino acids 1236 and 1417. Altogether, 72 unique glycopeptides were determined in this mixture including 42 on fibronectin. Several of the site-specific glycosylation data identified here were not reported previously, especially for AGP-1 (52, 55) and fibronectin (53).

In a last example, the ability of GlycoPAT to profile human *N*-linked glycans in a complex mixture was assessed by analyzing plasma cryoprecipitate prepared from 5 ml blood drawn from an O-type blood donor. Here, spectra potentially corresponding to glycopeptides were identified using product-dependent HCD-mode fragmentation. These putative gly-

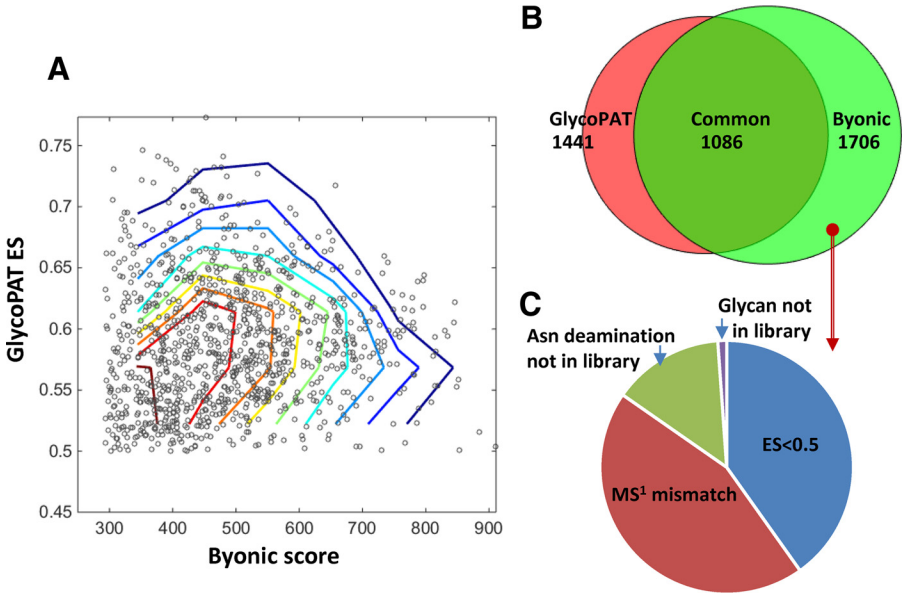


FIG. 6. **Glycoproteome of prostate cancer cells.** Glycoproteomics analysis using GlycoPAT was performed for 22 tandem-MS runs that analyzed cell lysates from a mixture of two prostate cancer cell lines, LNCaP and PC-3. Results were compared with Byonic results reported by Shah *et al.* (2015) (36). *A*, Dot plot showing summary scores for GlycoPAT (*ES*) and Byonic, along with contour lines. Each dot represents the same MS/MS spectrum-glycopeptide identification by both GlycoPAT and Byonic. High GlycoPAT scores generally agreed with Byonic assignments, though some scatter in the data is evident. *B*, Venn diagram showing overlap in the number of spectra identified using GlycoPAT and Byonic. Most of the identifications were common, though some hits were unique to GlycoPAT because of consideration of simultaneous fragmentation events on the glycan and peptide backbone. *C*, 620 files uniquely identified by Byonic were not identified by GlycoPAT. This was because of differences in precursor monoisotopic mass assignment, lower than acceptable ES score or library used for the search.

copeptides were then fragmented using CID. Here, once a candidate glycopeptide was identified based on MS^1 mass match, CID MS/MS was used to determine glycan structure. The presence of a prominent peak corresponding to the underlying peptide with 0–2 HexNAcs (+0–2Hex or +Fuc in the case of core-fucose) in the HCD MS/MS spectra confirmed the peptide backbone identity. Additionally, the MS^1 spectrum was manually inspected to verify that the fragmented ion was the monoisotopic peak. GlyPepDB library, in this case, included the seven most abundant proteins present in the sample based on sequence coverage. The glycans in this data set included the "standard N-glycan GlyDB" and additional carbohydrate structures corresponding to blood group antigens. This experiment revealed 57 unique glycopeptides, including 22 glycopeptides on von Willebrand Factor (VWF) (supplemental Table S6). Although the glycans identified here were themselves identical to a previous study based on Glycomics analysis (56), several novel site-specific glycosylation events or glycopeptides are reported in supplemental Table S6. Detailed structural analysis performed to distinguish between different glycoforms confirmed the presence of core-fucosylated glycans on several VWF glycopeptides. However, unequivocally confirmation of the presence of blood group antigens on VWF was not possible (56), as this requires higher-levels of MS^n analysis. The single site of glycosylation on plasminogen (PLMN) contained a tetra-antennary glycan. Fibronectin (FINC) had a subset of the glycopeptides identified in the 4-protein mixture studies. The glycopeptides of fibrinogen were mostly bi-antennary sialylated structures particularly on the β (FIBB) and γ (FIBG) chains of the protein, consistent with a previous glycomics investigation (57). Higher level of glycan branching was noted on the α-chain (FIBA). Finally, the study identified complex N-glycans at four of the eight potential sites of alpha-2-macroglobulin (58). Overall, the pilot study illustrates the ability of GlycoPAT to analyze the plasma glycoproteome.

Glycoproteomics Analysis of Prostate Cell Lines—To illustrate the ability of GlycoPAT to identify glycopeptides in complex samples, we analyzed previously published prostate cancer glycoproteomics experiments using GlycoPAT, and compared the findings with Byonic scores (supplemental Table S7 in (36)). The GlyPepDB in this case had 429,841 members. Peptides and glycans used to generate this library are listed in supplemental Tables S7 and S8. Such analysis identified 1441 spectrum including 960 unique glycopeptides (supplemental Table S9). 1086 of these spectrum were common between GlycoPAT and Byonic (Fig. 6*A*), and thus identifications with high GlycoPAT *ES* also typically displayed high Byonic scores. Though the total number of files/spectra identified by both software were similar, 355 identifications were unique to GlycoPAT with $ES>0.5$, and 616 were unique to Byonic with score>300 (Fig. 6*B*). MS/MS spectra uniquely identified using GlycoPAT with $ES>0.5$ are individually annotated as part of supplemental Material. Many of these were

identified because of the unique scoring scheme of GlycoPAT, which is probability based, and which weights simultaneous glycan and peptide fragmentation events. Unique identifications reported by Shah *et al.* (2015) (36) fell below the GlycoPAT acceptance criteria, primarily either because of differences in the monoisotopic peak assignment or low *ES* score (Fig. 6*C*). Overall, the analysis of this complex data set using two independent approaches illustrates the complexity of the glycoproteomics data analysis problem. It suggests that further refinement of the scoring strategy is necessary. Additionally, reliable glycopeptide identifications likely requires comprehensive MS^n scoring in more than one fragmentation mode.

DISCUSSION

This manuscript presents a well-documented, open-source software for glycoproteomics data analysis. The program presents several novel concepts and commonly used functions for glycopeptide digestion, database generation, glycopeptide fragmentation, spectrum scoring and glycopeptide FDR calculations. The scoring scheme developed has been validated for CID, HCD and ETD modes, primarily for N-glycans and O-linked glycans. The focus of the current effort was on software development, rather than the discovery of new biology, because open-source, easy-to-use, modular computational resources are currently lacking in the field of glycoproteomics. This represents a major research bottleneck that hampers the field (7, 10). To this end, the current program comes with systematic class definitions, modular design, comprehensive documentation and online tutorials to facilitate program expansion by various investigators in the field. Using this tool, arbitrary monosaccharides, PTMs and fragmentation rules for additional methods like EThcD (59) can be rapidly introduced to enable tandem-MS data analysis (detailed examples in user manual). Additionally, the program is written in MATLAB (with JAVA libraries) because the presence of a vast library of well-written MATLAB functions will enable the rapid expansion of GlycoPAT capabilities without the need for extensive coding. This includes functions for GUI development, statistical analysis, text manipulation, data visualization and table handling. Using a 5-node, 60 core computing cluster and default GlycoPAT settings, ~90,000 MS spectra for a single plasma cryoprecipitate run can be analyzed in 8h against a ~50,000-member GlyPepDB. Finally, the use of MATLAB facilitates the ready integration of programming modules developed in this package with other programs in the fields of Systems Glycobiology that already use the same platform (7, 23, 60). Together, these developments aim to make the field of glycoproteomics more accessible to the larger biological community.

Unlike proteomics based programs like Byonic, GlycoPAT is glycan-centric, with a greater focus on carbohydrate fragmentation/structure analysis based on CID-mode data. In this regard, low-energy CID yields a pattern of glycan B-/Y- ions that reflect the carbohydrate assembly. The analysis of this fragmentation pattern enables partial validation of the glycan structure, but it misses the underlying peptide. To address this limitation, GlycoPAT also has facilities for high-energy HCD and ETD MS^n spectra analysis, as these provide clues regarding the glycosylation site. Here, HCD enables the identification of glycopeptide fragmentation spectra because of the release of prominent oxonium-ions at low *m/z*, and it also fragments the peptide backbone. Additionally, as shown in Fig. 5, simultaneous glycan and peptide backbone fragmentation is also a common occurrence in HCD and must be considered during scoring. Unlike the collision activated dissociation modes, ETD predominantly leaves the glycan intact, but it results in glycopeptide backbone fragmentation. This mode is however, most useful only for multiply charged glycopeptides with precursor $m/z < 1000$. Because of the above, the ideal experimental workflow and data analysis software should combine complementary information emerging from fragmentation of the same peptide in all three fragmentation modes to arrive at a final identification.

Samples analyzed by GlycoPAT thus far include standard glycoproteins, simple mixtures, the most abundant entities in plasma cryoprecipitate and prostate cancer cell line preparations. The *ES*, measured in the current version, is limited to MS/MS spectrum analysis. It is anticipated that future versions of the program will include more complex *ES* estimates that integrate the individual scores from HCD, ETD, and CID fragmentation modes at multiple MS^n levels. Additionally, as shown recently, there are several challenges with glycopeptide identification in complex mixtures that are not handled well in current software (35, 61). These "challenging assignments" occur because of the following identical or near-identical mass balances that can lead to false-identifications, especially in complex mixtures where the search library size is large: (1) Neu5Ac-Neu5Gc = Hex-Fuc = oxidation mass (35); (2) 2×Fuc-Neu5Ac = 1Da (35); (3) HexNac-Fuc = carboamidomethyl modification (61); (4) HexNAc-Hex = carboamidomethyl -oxidation (61); (5) Asn/Gln deamidation. As suggested by Darula *et al.* (61), additional software development, analysis of data from more than one fragmentation mode, and consideration of LC retention time is necessary to handle these challenging assignments.

Although the current manuscript introduces an extensive computational infrastructure for glycoproteomics analysis, additional functional modules are currently being developed to handle more complex experimental workflows and to reduce computational time. Specifically, the current version uses the SGP1.0 nomenclature as it accommodates arbitrary monosaccharide types and enables easy *in silico* glycopeptide fragmentation at multiple locations. This format is currently being expanded to also accommodate bond linkage information. Additionally, a new module called *DrawGlycan* is being integrated into GlycoPAT to render high-quality glycan drawings, including bond fragmentation data, in the final an-

notated output spectrum (60). Modules are also being added to quantitatively discriminate between isomeric glycan structures that share the same precursor mass. Finally, the current code has been validated using single glycoproteins, glycoprotein mixtures or plasma cryoprecipitate, without the implementation of glycopeptide enrichment strategies. Such enrichment methods using lectins, ion-exchange or other specialized columns may enhance the fidelity of the glycopeptide identifications. Together, these advancements, along with the utilization of quantitative MS methods, are planned as this can reveal new details regarding the heterogeneous glycoproteome that is currently masked by the more abundant nonglycosylated entities.

Acknowledgments—We thank Marshall Bern (Protein Metrics, Inc.) for kindly sharing the Byonic software.

DATA AVAILABILITY

GlycoPAT program source code, compiled GUIs, detailed instructional manuals and videos are available from the software homepage: www.VirtualGlycome.org/glycopat. Raw MS data are at the ProteomeXchange Consortium (https://www.ebi.ac.uk/pride/archive/, PRIDE identifier: PXD006031).

* This work was supported by the National Institutes of Health grants HL103411 and HL77258, and the Program for Excellence in Glycosciences grant HL107146. CYL was supported by a T32 NIH Ruth L. Kirschstein Postdoctoral Research Training Grant. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors declare that they have no conflicts of interest.

‡‡ To whom correspondence should be addressed: State University of New York – Buffalo, 906 Furnas Hall, Buffalo, NY 14260; Tel.: 716-645-1200; Fax: 716-645-3822; E-mail: neel@buffalo.edu.

S This article contains supplemental material.

** These authors contributed equally to this work.

REFERENCES

1. Neelamegham, S., and Liu, G. (2011) Systems glycobiology: biochemical reaction networks regulating glycan structure and function. *Glycobiology* **21,** 1541–1553
2. Varki, A. (2017) Biological roles of glycans. *Glycobiology* **27,** 3–49
3. Neelamegham, S., and Mahal, L. K. (2016) Multi-level regulation of cellular glycosylation: from genes to transcript to enzyme to structure. *Curr. Opin. Struct. Biol.* **40,** 145–152
4. Alley, W. R., Jr, Mann, B. F., and Novotny, M. V. (2013) High-sensitivity analytical approaches for the structural characterization of glycoproteins. *Chem. Rev.* **113,** 2668–2732
5. Hu, H., Khatri, K., and Zaia, J. (2017) Algorithms and design strategies towards automated glycoproteomics analysis. *Mass Spectrom. Rev.* **36,** 475–498
6. Thaysen-Andersen, M., Packer, N. H., and Schulz, B. L. (2016) Maturing glycoproteomics technologies provide unique structural insights into the N-glycoproteome and its regulation in health and disease. *Mol. Cell. Proteomics* **15,** 1773–1790
7. Liu, G., and Neelamegham, S. (2015) Integration of systems glycobiology with bioinformatics toolboxes, glycoinformatics resources, and glycoproteomics data. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **7,** 163–181
8. Segu, Z. M., and Mechref, Y. (2010) Characterizing protein glycosylation sites through higher-energy C-trap dissociation. *Rapid Commun. Mass Spectrom.* **24,** 1217–1225
9. Woodin, C. L., Maxon, M., and Desaire, H. (2013) Software for automated interpretation of mass spectrometry data from glycans and glycopeptides. *Analyst* **138,** 2793–2803
10. Dallas, D. C., Martin, W. F., Hua, S., and German, J. B. (2013) Automated glycopeptide analysis–review of current state and future directions. *Brief Bioinform* **14,** 361–374
11. Serang, O., Froehlich, J. W., Muntel, J., McDowell, G., Steen, H., Lee, R. S., and Steen, J. A. (2013) SweetSEQer, simple de novo filtering and annotation of glycoconjugate mass spectra. *Mol. Cell. Proteomics* **12,** 1735–1740
12. Liang, S. Y., Wu, S. W., Pu, T. H., Chang, F. Y., and Khoo, K. H. (2014) An adaptive workflow coupled with Random Forest algorithm to identify intact N-glycopeptides detected from mass spectrometry. *Bioinformatics* **30,** 1908–1916
13. Toghi Eshghi, S., Shah, P., Yang, W., Li, X., and Zhang, H. (2015) GPQuest: A spectral library matching algorithm for site-specific assignment of tandem mass spectra to intact N-glycopeptides. *Anal. Chem.* **87,** 5181–5188
14. Joenvaara, S., Ritamo, I., Peltoniemi, H., and Renkonen, R. (2008) N-glycoproteomics - an automated workflow approach. *Glycobiology* **18,** 339–349
15. Ozohanics, O., Krenyacz, J., Ludanyi, K., Pollreisz, F., Vekey, K., and Drahos, L. (2008) GlycoMiner: a new software tool to elucidate glycopeptide composition. *Rapid Commun. Mass Spectrom.* **22,** 3245–3254
16. Pompach, P., Chandler, K. B., Lan, R., Edwards, N., and Goldman, R. (2012) Semi-automated identification of N-Glycopeptides by hydrophilic interaction chromatography, nano-reverse-phase LC-MS/MS, and glycan database search. *J. Proteome Res.* **11,** 1728–1740
17. Bern, M., Kil, Y. J., and Becker, C. (2012) Byonic: advanced peptide and protein identification software. *Curr. Protoc. Bioinformatics* **Chapter 13,** Unit13 20
18. Mayampurath, A., Yu, C. Y., Song, E., Balan, J., Mechref, Y., and Tang, H. (2014) Computational framework for identification of intact glycopeptides in complex samples. *Anal. Chem.* **86,** 453–463
19. He, L., Xin, L., Shan, B., Lajoie, G. A., and Ma, B. (2014) GlycoMaster DB: Software to assist the automated identification of N-linked glycopeptides by tandem mass spectrometry. *J Proteome Res.* **13,** 3881–3895
20. Park, G. W., Kim, J. Y., Hwang, H., Lee, J. Y., Ahn, Y. H., Lee, H. K., Ji, E. S., Kim, K. H., Jeong, H. K., Yun, K. N., Kim, Y. S., Ko, J. H., An, H. J., Kim, J. H., Paik, Y. K., and Yoo, J. S. (2016) Integrated GlycoProteome analyzer (I-GPA) for automated identification and quantitation of site-specific N-glycosylation. *Sci. Rep.* **6,** 21175
21. Chalkley, R. J., and Baker, P. R. (2017) Use of a glycosylation site database to improve glycopeptide identification from complex mixtures. *Anal. Bioanal. Chem.* **409,** 571–577
22. Ashline, D. J., Zhang, H., and Reinhold, V. N. (2017) Isomeric complexity of glycosylation documented by MSn. *Anal. Bioanal. Chem.* **409,** 439–451
23. Liu, G., and Neelamegham, S. (2014) A computational framework for the automated construction of glycosylation reaction networks. *PLoS ONE* **9,** e100939
24. Pedrioli, P. G., Eng, J. K., Hubley, R., Vogelzang, M., Deutsch, E. W., Raught, B., Pratt, B., Nilsson, E., Angeletti, R. H., Apweiler, R., Cheung, K., Costello, C. E., Hermjakob, H., Huang, S., Julian, R. K., Kapp, E., McComb, M. E., Oliver, S. G., Omenn, G., Paton, N. W., Simpson, R., Smith, R., Taylor, C. F., Zhu, W., and Aebersold, R. (2004) A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol.* **22,** 1459–1466
25. Dayananda, K. M., Singh, I., Mondal, N., and Neelamegham, S. (2010) von Willebrand factor self-association on platelet Gplbalpha under hydrodynamic shear: effect on shear-induced platelet activation. *Blood* **116,** 3990–3998
26. An, B., Zhang, M., Johnson, R. W., and Qu, J. (2015) Surfactant-aided precipitation/on-pellet-digestion (SOD) procedure provides robust and rapid sample preparation for reproducible, accurate and sensitive LC/MS quantification of therapeutic protein in plasma and tissues. *Anal. Chem.* **87,** 4023–4029
27. Duan, X., Young, R., Straubinger, R. M., Page, B., Cao, J., Wang, H., Yu, H., Canty, J. M., and Qu, J. (2009) A straightforward and highly efficient precipitation/on-pellet digestion procedure coupled with a long gradient nano-LC separation and Orbitrap mass spectrometry for label-free ex-

pression profiling of the swine heart mitochondrial proteome. *J. Proteome Res.* **8,** 2838–2850
28. Qu, J., Jusko, W. J., and Straubinger, R. M. (2006) Utility of cleavable isotope-coded affinity-tagged reagents for quantification of low-copy proteins induced by methylprednisolone using liquid chromatography/tandem mass spectrometry. *Anal. Chem.* **78,** 4543–4552
29. Tu, C., Li, J., Young, R., Page, B. J., Engler, F., Halfon, M. S., Canty, J. M., Jr, and Qu, J. (2011) Combinatorial peptide ligand library treatment followed by a dual-enzyme, dual-activation approach on a nanoflow liquid chromatography/orbitrap/electron transfer dissociation system for comprehensive analysis of swine plasma proteome. *Anal. Chem.* **83,** 4802–4813
30. Kessner, D., Chambers, M., Burke, R., Agus, D., and Mallick, P. (2008) ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **24,** 2534–2536
31. Lo, C. Y., Antonopoulos, A., Gupta, R., Qu, J., Dell, A., Haslam, S. M., and Neelamegham, S. (2013) Competition between core-2 GlcNAc-transferase and ST6GalNAc-transferase regulates the synthesis of the leukocyte selectin ligand on human P-selectin glycoprotein ligand-1. *J. Biol. Chem.* **288,** 13974–13987
32. Eng, J. K., Mccormack, A. L., and Yates, J. R. (1994) An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. *J. Am. Soc. Mass Spectr.* **5,** 976–989
33. Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., Yang, X. Y., Shi, W. Y., and Bryant, S. H. (2004) Open mass spectrometry search algorithm. *J. Proteome Res.* **3,** 958–964
34. Meng, F. Y., Cargile, B. J., Miller, L. M., Forbes, A. J., Johnson, J. R., and Kelleher, N. L. (2001) Informatics and multiplexing of intact protein identification in bacteria and the archaea. *Nat. Biotechnol.* **19,** 952–957
35. Lee, L. Y., Moh, E. S., Parker, B. L., Bern, M., Packer, N. H., and Thaysen-Andersen, M. (2016) Toward Automated N-Glycopeptide Identification in Glycoproteomics. *J. Proteome Res.* **15,** 3904–3915
36. Shah, P., Wang, X., Yang, W., Toghi Eshghi, S., Sun, S., Hoti, N., Chen, L., Yang, S., Pasay, J., Rubin, A., and Zhang, H. (2015) Integrated proteomic and glycoproteomic analyses of prostate cancer cells reveal glycoprotein alteration in protein abundance and glycosylation. *Mol. Cell. Proteomics* **14,** 2753–2763
37. Senko, M. W., Beu, S. C., and McLaffertycor, F. W. (1995) Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *J. Am. Soc. Mass Spectrom.* **6,** 229–233
38. Kronewitter, S. R., An, H. J., de Leoz, M. L., Lebrilla, C. B., Miyamoto, S., and Leiserowitz, G. S. (2009) The development of retrosynthetic glycan libraries to profile and classify the human serum N-linked glycome. *Proteomics* **9,** 2986–2994
39. Varki, A., Cummings, R. D., Aebi, M., Packer, N. H., Seeberger, P. H., Esko, J. D., Stanley, P., Hart, G., Darvill, A., Kinoshita, T., Prestegard, J. J., Schnaar, R. L., Freeze, H. H., Marth, J. D., Bertozzi, C. R., Etzler, M. E., Frank, M., Vliegenthart, J. F., Lutteke, T., Perez, S., Bolton, E., Rudd, P., Paulson, J., Kanehisa, M., Toukach, P., Aoki-Kinoshita, K. F., Dell, A., Narimatsu, H., York, W., Taniguchi, N., and Kornfeld, S. (2015) Symbol nomenclature for graphical representations of glycans. *Glycobiology* **25,** 1323–1324
40. Liu, G., Puri, A., and Neelamegham, S. (2013) Glycosylation Network Analysis Toolbox: a MATLAB-based environment for systems glycobiology. *Bioinformatics* **29,** 404–406
41. Strum, J. S., Nwosu, C. C., Hua, S., Kronewitter, S. R., Seipert, R. R., Bachelor, R. J., An, H. J., and Lebrilla, C. B. (2013) Automated assignments of N- and O-site specific glycosylation with extensive glycan heterogeneity of glycoprotein mixtures. *Anal. Chem.* **85,** 5666–5675
42. Lee, L. Y., Moh, E. S., Parker, B. L., Bern, M., Packer, N. H., and Thaysen-Andersen, M. (2016) Toward automated N-glycopeptide identification in glycoproteomics. *J. Proteome Res.* **15,** 3904–3915
43. Saba, J., Dutta, S., Hemenway, E., and Viner, R. (2012) Increasing the productivity of glycopeptides analysis by using higher-energy collision dissociation-accurate mass-product-dependent electron transfer dissociation. *Int. J. Proteomics* **560391,** 2012
44. Green, E. D., Adelt, G., Baenziger, J. U., Wilson, S., and Van Halbeek, H. (1988) The asparagine-linked oligosaccharides on bovine fetuin. Structural analysis of N-glycanase-released oligosaccharides by 500-megahertz 1H NMR spectroscopy. *J. Biol. Chem.* **263,** 18253–18268
45. Nwosu, C. C., Seipert, R. R., Strum, J. S., Hua, S. S., An, H. J., Zivkovic, A. M., German, B. J., and Lebrilla, C. B. (2011) Simultaneous and extensive site-specific N- and O-glycosylation analysis in protein mixtures. *J. Proteome Res.* **10,** 2612–2624
46. Rebecchi, K. R., Wenke, J. L., Go, E. P., and Desaire, H. (2009) Label-free quantitation: a new glycoproteomics approach. *J. Am. Soc. Mass Spectrom.* **20,** 1048–1059
47. Takasaki, S., and Kobata, A. (1986) Asparagine-linked sugar chains of fetuin: occurrence of tetrasialyl triantennary sugar chains containing the Gal beta 1—3GlcNAc sequence. *Biochemistry* **25,** 5709–5715
48. Ren, J. M., Rejtar, T., Li, L., and Karger, B. L. (2007) N-Glycan structure annotation of glycopeptides using a linearized glycan structure database (GlyDB). *J. Proteome Res.* **6,** 3162–3173
49. Thaysen-Andersen, M., Mysling, S., and Hojrup, P. (2009) Site-specific glycoprofiling of N-linked glycopeptides using MALDI-TOF MS: strong correlation between signal strength and glycoform quantities. *Anal. Chem.* **81,** 3933–3943
50. Prien, J. M., Ashline, D. J., Lapadula, A. J., Zhang, H., and Reinhold, V. N. (2009) The high mannose glycans from bovine ribonuclease B isomer characterization by ion trap MS. *J. Am. Soc. Mass Spectrom.* **20,** 539–556
51. Imre, T., Kremmer, T., Heberger, K., Molnar-Szollosi, E., Ludanyi, K., Pocsfalvi, G., Malorni, A., Drahos, L., and Vekey, K. (2008) Mass spectrometric and linear discriminant analysis of N-glycans of human serum alpha-1-acid glycoprotein in cancer patients and healthy individuals. *J. Proteomics* **71,** 186–197
52. Thobhani, S., Yuen, C. T., Bailey, M. J., and Jones, C. (2009) Identification and quantification of N-linked oligosaccharides released from glycoproteins: an inter-laboratory study. *Glycobiology* **19,** 201–211
53. Tajiri, M., Yoshida, S., and Wada, Y. (2005) Differential analysis of site-specific glycans on plasma and cellular fibronectins: application of a hydrophilic affinity method for glycopeptide enrichment. *Glycobiology* **15,** 1332–1340
54. Wagner, D. D., Ivatt, R., Destree, A. T., and Hynes, R. O. (1981) Similarities and differences between the fibronectins of normal and transformed hamster cells. *J. Biol. Chem.* **256,** 11708–11715
55. Clerc, F., Reiding, K. R., Jansen, B. C., Kammeijer, G. S., Bondt, A., and Wuhrer, M. (2016) Human plasma protein N-glycosylation. *Glycoconj. J.* **33,** 309–343
56. Canis, K., McKinnon, T. A., Nowak, A., Haslam, S. M., Panico, M., Morris, H. R., Laffan, M. A., and Dell, A. (2012) Mapping the N-glycome of human von Willebrand factor. *Biochem. J.* **447,** 217–228
57. Adamczyk, B., Struwe, W. B., Ercan, A., Nigrovic, P. A., and Rudd, P. M. (2013) Characterization of fibrinogen glycosylation and its importance for serum/plasma N-glycome analysis. *J. Proteome Res.* **12,** 444–454
58. Sottrup-Jensen, L., Stepanik, T. M., Kristensen, T., Wierzbicki, D. M., Jones, C. M., Lonblad, P. B., Magnusson, S., and Petersen, T. E. (1984) Primary structure of human alpha 2-macroglobulin. V. The complete structure. *J. Biol. Chem.* **259,** 8318–8327
59. Frese, C. K., Altelaar, A. F., van den Toorn, H., Nolting, D., Griep-Raming, J., Heck, A. J., and Mohammed, S. (2012) Toward full peptide sequence coverage by dual fragmentation combining electron-transfer and higher-energy collision dissociation tandem mass spectrometry. *Anal. Chem.* **84,** 9668–9673
60. Cheng, K., Zhou, Y., and Neelamegham, S. (2017) DrawGlycan-SNFG: a robust tool to render glycans and glycopeptides with fragmentation information. *Glycobiology* **27,** 200–205
61. Darula, Z., and Medzihradszky, K. F. (2015) Carbamidomethylation side reactions may lead to glycan misassignments in glycopeptide analysis. *Anal. Chem.* **87,** 6297–6302