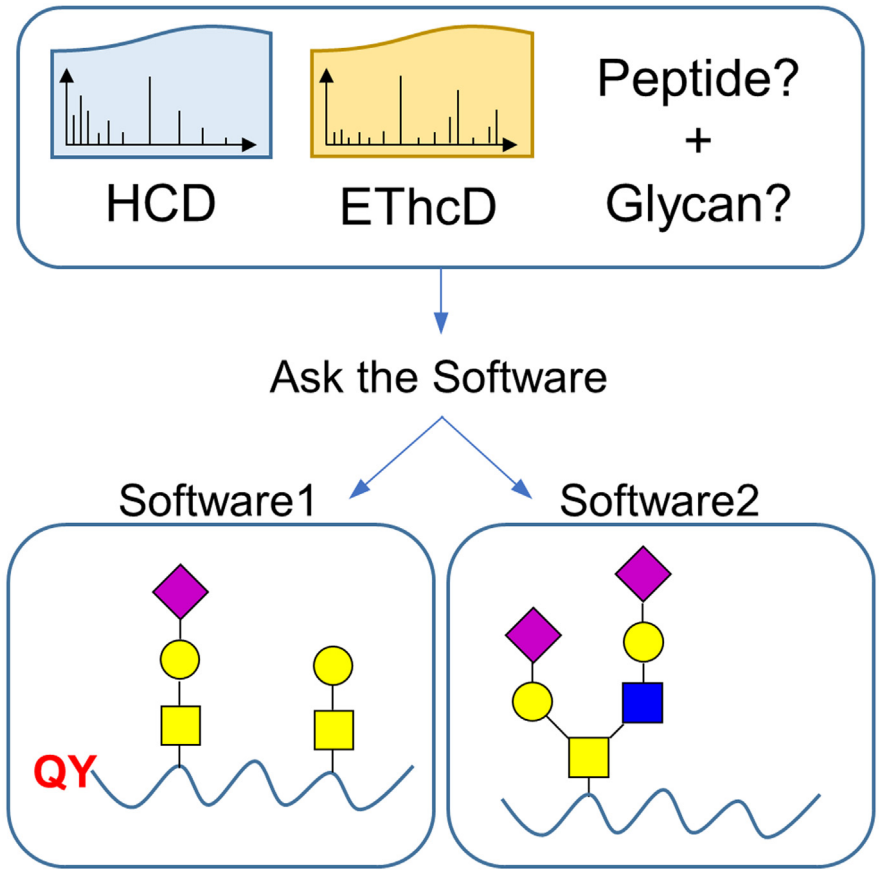
Multiple Layers of Complexity in O-Glycosylation Illustrated With the Urinary Glycoproteome

Authors
Adam Pap, Istvan Elod Kiraly, Katalin F. Medzihradszky, and Zsuzsanna Darula

Correspondence
medzihradszky.katalin@brc.hu;
darula.zsuzsanna@brc.hu

In Brief
In this study, mass spectrometric data acquired on urinary O-glycopeptides were analyzed by four software packages. The results were compared, and the rate of misidentification was assessed. The major factors leading to data misinterpretation were identified, and software development suggestions aiming more reliable automated data interpretation were made.

Graphical Abstract



Highlights

- Urinary O-glycopeptides were enriched using lectin affinity chromatography.
- HCD and EThcD data were analyzed by four proteomic software packages.
- High misidentification rate in spite of strict probability-based acceptance criteria.
- Software development recommendations for more reliable O-glycopeptide analysis.

2022, Mol Cell Proteomics *21*(*12*), 100439

https://doi.org/10.1016/j.mcpro.2022.100439

Multiple Layers of Complexity in O-Glycosylation Illustrated With the Urinary Glycoproteome

Adam Pap[1], Istvan Elod Kiraly[2], Katalin F. Medzihradszky[1,*], and Zsuzsanna Darula[1,3,*]

While N-glycopeptides are relatively easy to characterize, O-glycosylation analysis is more complex. In this article, we illustrate the multiple layers of O-glycopeptide characterization that make this task so challenging. We believe our carefully curated dataset represents perhaps the largest intact human glycopeptide mixture derived from individuals, not from cell lines. The samples were collected from healthy individuals, patients with superficial or advanced bladder cancer (three of each group), and a single bladder inflammation patient. The data were scrutinized manually and interpreted using three different search engines: Byonic, Protein Prospector, and O-Pair, and the tool MS-Filter. Despite all the recent advances, reliable automatic O-glycopeptide assignment has not been solved yet. Our data reveal such diversity of site-specific O-glycosylation that has not been presented before. In addition to the potential biological implications, this dataset should be a valuable resource for software developers in the same way as some of our previously released data has been used in the development of O-Pair and O-Glycoproteome Analyzer. Based on the manual evaluation of the performance of the existing tools with our data, we lined up a series of recommendations that if implemented could significantly improve the reliability of glycopeptide assignments.

Mass spectrometry (MS) has been used for the characterization of glycoproteins ever since the advent of soft ionization techniques and played a significant role in the discovery of novel O-linked modifications (1–3). With the growing importance of recombinant therapeutic proteins, glycosylation analysis has become essential for the pharmacological industry (4, 5). At the same time, an improved tool set—including new enrichment methods, fast mass spectrometers with high mass accuracy and detection sensitivity, and the development of electron-transfer dissociation and electron-transfer/higher-energy collision dissociation (EThcD)—has permitted the characterization of labile post-translational modifications (PTMs), and thus, enabled N- and recently O-glycosylation analysis of wild-type samples aimed at the better understanding of the biological roles of these diverse PTMs and also in the search of biomarkers (6–8).

For over a decade, our research group has been engaged in method development for the enrichment and mass spectrometric characterization of mucin-type O-glycopeptides first from bovine and human serum (9–11) and more recently from human urine. This latter study started out as a quest for elusive biomarkers. Urine was collected from healthy individuals, bladder cancer patients, and one outpatient with bladder inflammation. We used affinity chromatography with wheat germ agglutinin (WGA) to enrich glycopeptides, followed by LC/MS analysis with higher-energy collision dissociation (HCD)–triggered EThcD activation for data acquisition. Thus, we obtained MS/MS spectra by two mechanistically different fragmentation methods that are essential for successful glycopeptide characterization because these techniques deliver complementary information on the amino acid sequence and the glycan (12). HCD activation is still popular even in high-throughput glycosylation analysis of complex mixtures (13–16) because it affords efficient peptide identification. However, relying solely on HCD data prevents localization of glycosylation sites; moreover, only the total composition of the modifying glycans can be determined (12, 17). EThcD offers an additional yet untapped advantage beside site assignment providing information on the size and composition of the modifying glycans and on the connectivity of the different units. EThcD spectra acquired at minimal normalized collision energy (15% NCE) permitted us to distinguish isomeric glycoforms, revealing structural differences in the modifying glycans (18, 19). Manual data interpretation was essential in the identification and characterization of the 36 glycan structures, 29 of them were never reported in a site-specific manner. Evidently, the identity of the sugar units and the stereochemistry and exact positions of the linkages cannot be determined from these MS data. Thus, the structures were assigned based on known glycan biosynthesis pathways and on glycan structures described in mucin glycan studies (20–22). Our results indicated that the

From the [1]Laboratory of Proteomics Research, Biological Research Centre, Eotvos Lorand Research Network (ELKH) Szeged, Hungary; [2]Department of Urology, University of Szeged, Szeged, Hungary; [3]Single Cell Omics Advanced Core Facility, Hungarian Centre of Excellence for Molecular Medicine Szeged, Hungary

*For correspondence: Zsuzsanna Darula, darula.zsuzsanna@brc.hu; Katalin F. Medzihradszky, medzihradszky.katalin@brc.hu.

https://doi.org/10.1016/j.mcpro.2022.100439

urinary O-glycosylation landscape is more complex than expected (18, 19).

In this study, we have evaluated the performance of different software tools for automated interpretation of O-glycopeptide data. We used three different search engines: Byonic (23), Protein Prospector (PP) (24), and O-Pair (25), and the MS-Filter program in PP (26). Byonic and PP have both been adapted for glycopeptide analysis. From our perspective, their major difference is how the glycan fragmentation is handled. Fragments formed *via* glycosidic bond cleavages are searched for and scored by Byonic in both HCD and EThcD spectra, whereas PP can annotate these fragments when prompted to do so, but only activation-dependent peptide fragments contribute to the score. Byonic considers the glycan(s) linked to the peptide even upon collisional activation although it also permits gas-phase deglycosylation, whereas PP considers O-glycans as neutral losses by default. Byonic assigns the glycosylation site(s) even in HCD, and its Delta Mod score signals the reliability of the glycan placement. The built-in site localization (SLIP) (27) score of PP applies to glycosylation only in ET(hc)D, it clearly indicates the site of modification, or signals the lack of sufficient information. A rather limiting factor for both software is that data acquired on the same precursor with different activation methods, for example, HCD and EThcD, are not considered in concert; therefore, the complementary nature of these data is not exploited. O-Pair (25) is the only search engine that analyzes the two datasets combined, starting with the interpretation of HCD data to identify the peptide sequence and the additive mass of the glycan and using the electron-transfer dissociation data to determine the modification sites and further confirm the assignment. Site localization assessment also has been included in the output. Level 1 refers to complete confidence in both the glycan compositions and site localizations, level 2 indicates confidence about at least one glycan in multiply glycosylated peptides, whereas level 3 identifications deliver glycan composition assignment only. MS-Filter is the simplest approach developed for the identification of new glycoforms of glycopeptides confidently identified in preceding database searches. HCD spectra are searched for specific Y fragments (Y_0 and Y_1) (for nomenclature, see Ref. (28)) of the input peptide list, and the glycan composition is assigned based on the mass difference between the peptide and the precursor mass. Peptide backbone fragments, if there are any, are scored (26).

O-glycosylation analysis is a much more complex task than regular protein identification and even simple PTM analysis. A wide variety of glycoforms have to be considered during automated data interpretation because of the lack of consensus sites, frequent occurrence of Ser/Thr residues, and the potential macroheterogeneity and microheterogeneity at each site. Therefore, we tested all search engines with our data in an iterative manner. First glycoproteins modified with the three most common mucin-type glycans, the monosialylated and disialylated core-1 O-glycans, and the disialylated core-2 O-glycan ($HexNAcHexNeuAc_{1-2}$ and $HexNAc_2Hex_2NeuAc_2$) were identified. The follow-up searches were performed with an extended glycan list incorporating all glycans reported in our pilot studies (18, 19) using a protein database restricted to glycoproteins identified in the first round. Our reasoning was that glycoproteins bearing any mucin-type glycans certainly will feature the most common structures and at a higher level than the others. We also identified nonmodified sequences and N-glycopeptides performing Byonic searches. The results were summarized for each donor.

Careful and multifaceted investigation of all the O-glycopeptide assignments revealed that in spite of carefully chosen probability-based acceptance criteria, the false identification rate is higher than expected. Thus, our focus shifted, and in this article, we will provide insights about the difficulties the research community faces when analyzing wild-type O-glycopeptides. Based on this experience, we have drawn up a list about the data interpretation changes desirable for more reliable assignments.

In addition, as a result of the scrutiny invested in these data, we are also able to show examples about the diversity such studies could reveal that, as far as we know, has not been presented yet by other large-scale O-glycopeptide analyses.

We hope that sharing and consequently scrutinizing this information will lead to better understanding of the layers of complexity one has to tackle in intact O-glycopeptide analysis and will definitely help to develop better tools for data interpretation. In fact, some of our earlier urinary LC–MS/MS data have already been used for this purpose (25, 29).

EXPERIMENTAL PROCEDURES

Experimental Design and Statistical Rationale

Urine samples were collected from 10 donors (supplemental Table S1). The studies in this work abide by the Declaration of Helsinki principles. Consent forms were approved by the Hungarian Scientific and Research Ethics Committee (approval number: 1011/16). This was a discovery phase study only; no statistically significant quantitative results were obtained. The whole experimental workflow is illustrated in Figure 1.

Sample Preparation and MS

Ten random midstream urine samples were collected and stored at 4 °C before processing. Blood contamination was not observed for any of the samples. The previously published sample preparation protocol was followed (18). Briefly, the samples were centrifuged (5000*g*, 4 °C), and then the supernatants (50 ml per patient) were concentrated on 10 kDa molecular weight cutoff cellulose filters to 250 μl (5000*g*, 4 °C). Subsequently, proteins were reduced, alkylated, digested with trypsin and then subjected to a two-round glycopeptide enrichment using a WGA affinity column collecting three glycopeptide fractions, the end of the flow-through peak, its shoulder and

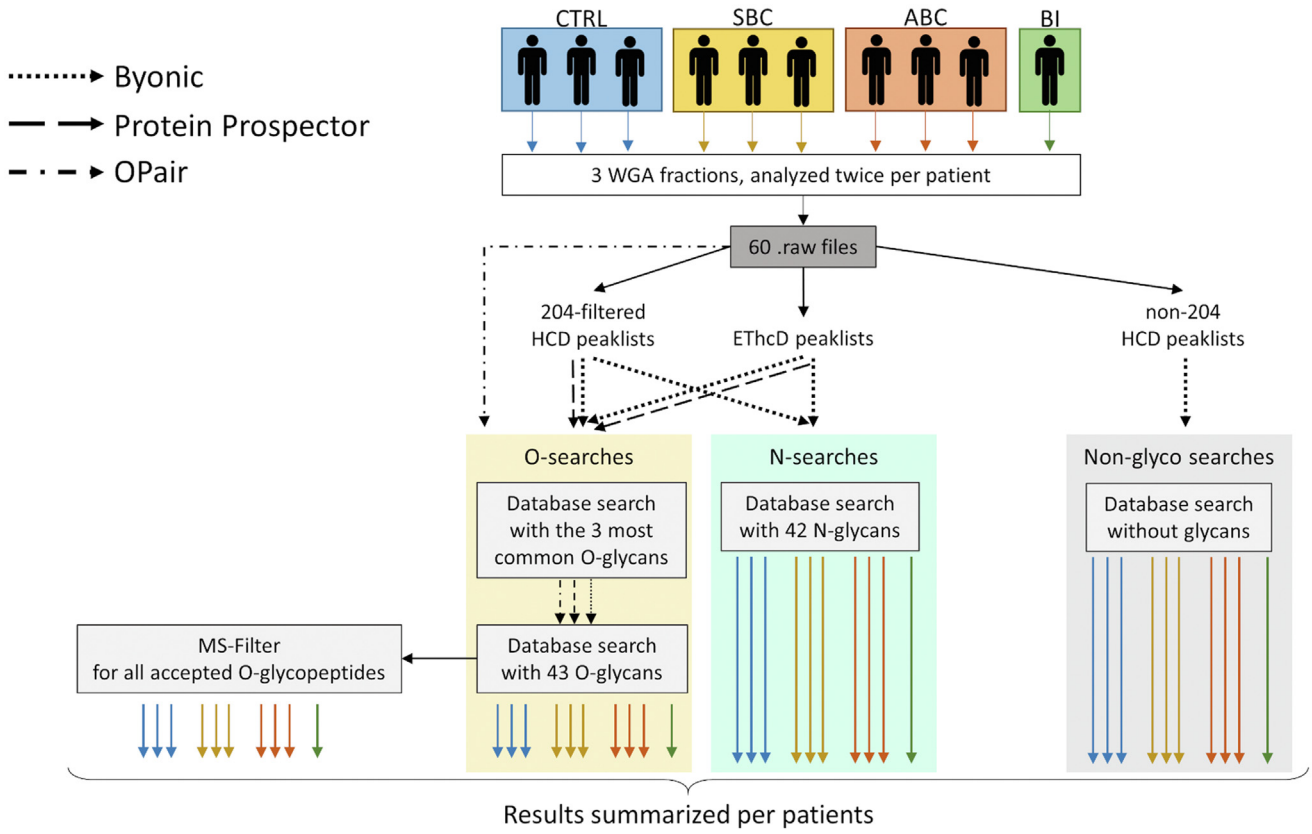


FIG. 1. **Workflow for the MS characterization of the urinary proteome.** HCD data derived from glycopeptides or from nonglycosylated sequences were separated based on the detection of the diagnostic HexNAc oxonium ion (*m/z* 204.0867 ± 10 ppm). O-glycopeptides were assigned in a two-round database search. Glycoproteins identified in the initial search served as a protein database for the second, restricted search. O-glycosylated sequences identified in the second search were used as the input list for the MS-Filter software. "CTRL" indicates the control (healthy) group, "SBC", "ABC," and "BI" stand for superficial bladder cancer, advanced bladder cancer, and bladder inflammation, respectively. HCD, higher-energy collision dissociation; MS, mass spectrometry.

a fraction eluted by GlcNAc (supplemental Fig. S1). Fractions were analyzed separately by LC–MS/MS using a Waters M-Class nano-UPLC online coupled to an Orbitrap Fusion Lumos Tribrid (Thermo Scientific) mass spectrometer. Samples were desalted on a trap column (Waters Acquity UPLC MClass Symmetry C18 180 μm × 20 mm column, particle size of 5 μm, pore size of 100 Å; and flow rate of 10 μl/min) and fractionated by a linear gradient of 10 to 30% B in 60 min (Waters Acquity UPLC M-Class BEH C18 75 μm × 250 mm column, particle size of 1.7 μm, pore size of 130 Å; solvent A: 0.1% formic acid/water; solvent B: 0.1% formic acid/acetonitrile; flow rate: 300 nl/min; and separating column temperature: 45 °C). MS/MS data were acquired using HCD product ion–dependent EThcD data acquisition; the presence of the diagnostic HexNAc-specific oxonium ion, *m/z* 204.0867 among the 20 most abundant HCD fragments, triggered EThcD acquisition. HCD spectra were acquired at 28% NCE, whereas supplemental activation in EThcD was set to 15% NCE. About 40% of the collected WGA fractions were injected, and each fraction was analyzed twice, selecting precursors of different charge states in the consecutive LC–MS/MS experiments (z = 3–5 or z = 2). MS/MS intensity threshold was set to 10^6 in a total cycle time of 3 s. All measurements were performed in the Orbitrap analyzer with a resolution of 60,000 and 15,000 for MS1 and MS/MS, respectively. Dynamic exclusion was set to 30 s.

Data Interpretation

Separate HCD and EThcD peak lists were generated from the .raw files (supplemental Table S1) using Proteome Discoverer (Thermo Scientific), version 2.4. Spectra with minimum 40 peaks were retained. The HCD data were further divided into two peak lists based on the presence of the HexNAc oxonium ion, *m/z* 204.087 ± 10 ppm among the 20 most abundant peaks using MS-Filter. The resulting EThcD and the HCD peak lists were searched separately with Byonic (version 3.7.4) and PP (version 6.2.1). For the O-Pair (version 0.0.308) searches, the original raw files were used. For MS-Filter, the unfiltered HCD peak lists were used as input. Search parameters and acceptance criteria are detailed in supplemental Tables S2–S9 and supplemental Fig. S2. All database searches used the human subset of Swiss-Prot protein database (release date: December 14, 2020).

Nonglycopeptide and N-Glycopeptide Searches Using Byonic—HCD data lacking the abundant HexNAc oxonium ion were searched for nonglycosylated peptides only (search parameters are listed in supplemental Table S2). supplemental Table S3 provides information on the N-glycopeptide searches from EThcD and 204-filtered HCD peak lists, whereas supplemental Table S4 contains the N-glycan database used that was created from Byonic's built-in N-glycan database (N-glycan 57 human plasma.txt) by removing 15 entries representing sodium adducts and truncated N-glycan structures.

O-Glycopeptide Searches Using Byonic, PP, and O-Pair—A two-round database search was implemented. The initial search was performed permitting only the three most common mucin-type structures ($HexNAcHexNeuAc_{1-2}$ and $HexNAc_2Hex_2NeuAc_2$) (supplemental Tables S5 and S7) with each search engine. The O-glycoproteins identified in the initial searches meeting the search engine–specific cutoff criteria formed the restricted database for the second round (supplemental Table S6) using an expanded O-glycan database with 42 additional glycans (supplemental Table S8) containing oligosaccharides identified in our previous investigations (18, 19).

HCD Peak List Processed With MS-Filter—Peptide sequences generated from confident O-glycopeptide identifications (meeting the acceptance criteria) were used as input (supplemental Table S10) for the MS-Filter (supplemental Fig. S2); only those HCD spectra were retained that featured the HexNAc oxonium ion (*m/z* 204.087) and a Y_0, Y_1 pair in a matching charge state among the top 15 peaks and within 10 ppm mass accuracy. Assigned HCD spectra had to feature an additive mass corresponding to a listed glycan in the selected glycan database (supplemental Table S9), and the precursor ion of this glycoform had to be measured within 10 ppm of the calculated value.

RESULTS AND DISCUSSION

The results presented here were obtained from urine samples collected from 10 male donors classified into four categories: healthy controls (3), superficial bladder cancer patients (3), advanced bladder cancer patients (3), and one patient with bladder inflammation. Sample preparation and mass spectrometric analysis were carried out identically for all samples. From each donor, 50 ml urine was concentrated and then digested with trypsin. LC–MS/MS data were generated from glycopeptide mixtures enriched by affinity chromatography using WGA, and HCD-fragment ion-triggered EThcD analysis ensured the acquisition of two MS/MS spectra produced by mechanistically different fragmentation processes for each glycopeptide. The resulting dataset (60 files altogether) was then processed with three different search engines (Byonic, PP, and O-Pair) and with MS-Filter software of PP in order to improve the success rate of O-glycopeptide assignments (Fig. 1).

O-glycopeptides were identified applying a two-round database search using Byonic, PP, and O-Pair.

Separate HCD and EThcD peak lists were generated for Byonic and PP searches. Peak lists representing the same donors were merged, and HCD data were further filtered for the presence of the HexNAc-related ion *m/z* 204.0867. O-Pair uses the raw data, and each file was searched separately. In the first search, only the most common glycans ($HexNAcHexNeuAc_{1-2}$ and $HexNAc_2Hex_2NeuAc_2$) were permitted, whereas in the second round, only O-glycoproteins identified reliably in the first round by the respective search engine either from HCD or EThcD data were searched using a larger glycan database with 40 additional glycan structures present on urinary proteins (18, 19) (supplemental Tables S11–S20, sheets D–H). Finally, an input list was assembled from the confidently identified O-glycopeptides identified by any of the aforementioned search engines in the second round to find additional confirmation in the form of Y_0 and Y_1 ions using MS-Filter (supplemental Tables S11–S20, sheets I and J). For further details, see the Experimental Procedures section.

N-glycopeptides were identified from EThcD or 204-filtered HCD data using Byonic (supplemental Tables S11–S20, sheet C). The lists of unmodified sequences were compiled from three searches: Byonic peptide spectrum matches (PSMs) from HCD spectra not featuring the HexNAc oxonium ion *m/z* 204 (supplemental Tables S11–S20, sheet N), Byonic PSMs from N-glycopeptide searches (supplemental Tables S11–S20, sheet O), and O-Pair results (supplemental Tables S11–S20, sheet L).

Approximately 115,000 MS/MS spectra featuring more than 40 peaks were recorded per patient (supplemental Table S21). The average identification rate was about 19%. Non-glycosylated sequences represent ~81% of the assigned MS/MS data, and only ~14% and ~5% of the identifications belong to O- and N-glycopeptides, respectively. In contrast, based on the presence of HexNAc oxonium ion (*m/z* 204.087), ~40% of all spectra might represent glycopeptides. The assignments meeting the reported acceptance criteria are summed up for each patient (supplemental Tables S11–S22) listing identifications by search methods separately as well as a compilation of all assignments. For some MS/MS data, multiple assignments are reported. In certain instances, O-Pair identified two or three components from the same spectrum, a unique feature of this search engine. Although it is not obvious from its online manual, it may reassign the precursor ion *m/z* or identify multiple precursors from the raw data. Not surprisingly, the majority of these "extra" identifications represented nonglycosylated peptides, since these produce more backbone fragments and are easier to assign than glycopeptides. We estimate that these multiple assignments represent less than 10% of the overall O-Pair-related PSMs, although the phenomenon of mixture spectra must be a lot more widespread.

O-Glycopeptides

From the identification rates, it is already obvious that automated glycopeptide identification is still a challenging task. We used different methods to extract more information from this large dataset than can be currently achieved with a single tool. We identified O-glycopeptides from HCD spectra (Byonic, PP, and MS-Filter), EThcD data (Byonic and PP), and the combination thereof (O-Pair). The acceptance criteria were set to minimize the decoy hits in Byonic and PP searches. For O-Pair results, we followed the developers' recommendation. For MS-Filter, only the weaker scoring double assignments were eliminated. This acceptance strategy was chosen based on our prior experience with these software. Still from our results (see later), it seems any cutoff threshold is currently a compromise. O-glycopeptides identified are listed individually according to the interpretation methods (supplemental

Tables S11–S20, sheets D–J) and summed up (supplemental Tables S11–S20, sheet B). Our efforts resulted in 43,151 O-glycopeptide PSMs (compiled in supplemental Table S22, sheet A). With all the duplicates removed, we can claim that data derived from 26,205 precursor selections were assigned. Strangely enough, that yielded 27,065 unique identifications (supplemental Table S22, sheet B), since 716 scans were assigned differently mostly by two different methods, but five HCD spectra were interpreted very differently by three tools, whereas O-Pair reported two glycopeptides from the same scans 15 times (respective "Scan#" highlighted in *yellow* in supplemental Table S22, sheet B).

The majority of the O-glycopeptides (56%) were assigned by only one of the software. Among the remaining identifications, 22.8, 13.1, 6, 1.9, and 0.2% were identified by two, three, four, five, or all six methods, respectively (Fig. 2 and supplemental Table S23, sheet A). By overlapping assignments, we mean identical peptide sequences and overall glycan composition—since the precise number and composition determination of the modifying glycans and site localization represent additional layers of complexity, and as we will present later, these issues are rather far from being tackled automatically.

Comparison of Search Engines Regarding Intact O-Glycopeptide Identification—Approximately 82% of the assigned PSMs represented glycopeptides with a molecular mass between 1500 and 4500 Da, and the vast majority of these are peptides of 8 to 25 amino acid residues (supplemental Table S23, sheet B). The average peptide length was 18 to 19 for all the methods, except MS-Filter with 15 amino acids only. At the same time, Byonic and O-Pair on average tend to assign components of ~10% higher molecular masses, that is, higher glycan/peptide ratios, than MS-Filter and PP, and the difference was more marked for the EThcD assignments (supplemental Table S23, sheet C). Figure 3 displays the shared and unique assignments in a precursor ion, charge state, and glycan size–dependent manner. The HCD-based approaches delivered significantly more results than EThcD alone (supplemental Figs. S2, S3 and supplemental Table S23, sheet A). Byonic-HCD searches were on the top, followed by O-Pair, MS-Filter, and PP-HCD (~71, 60, and 47% of the Byonic's numbers, respectively). The EThcD assignments were trailing with ~25% and ~13% (PP and Byonic, respectively). The HCD-based analyses yielded the most unique assignments. MS-Filter and Byonic produce similar numbers, O-Pair with some contribution from EThcD data

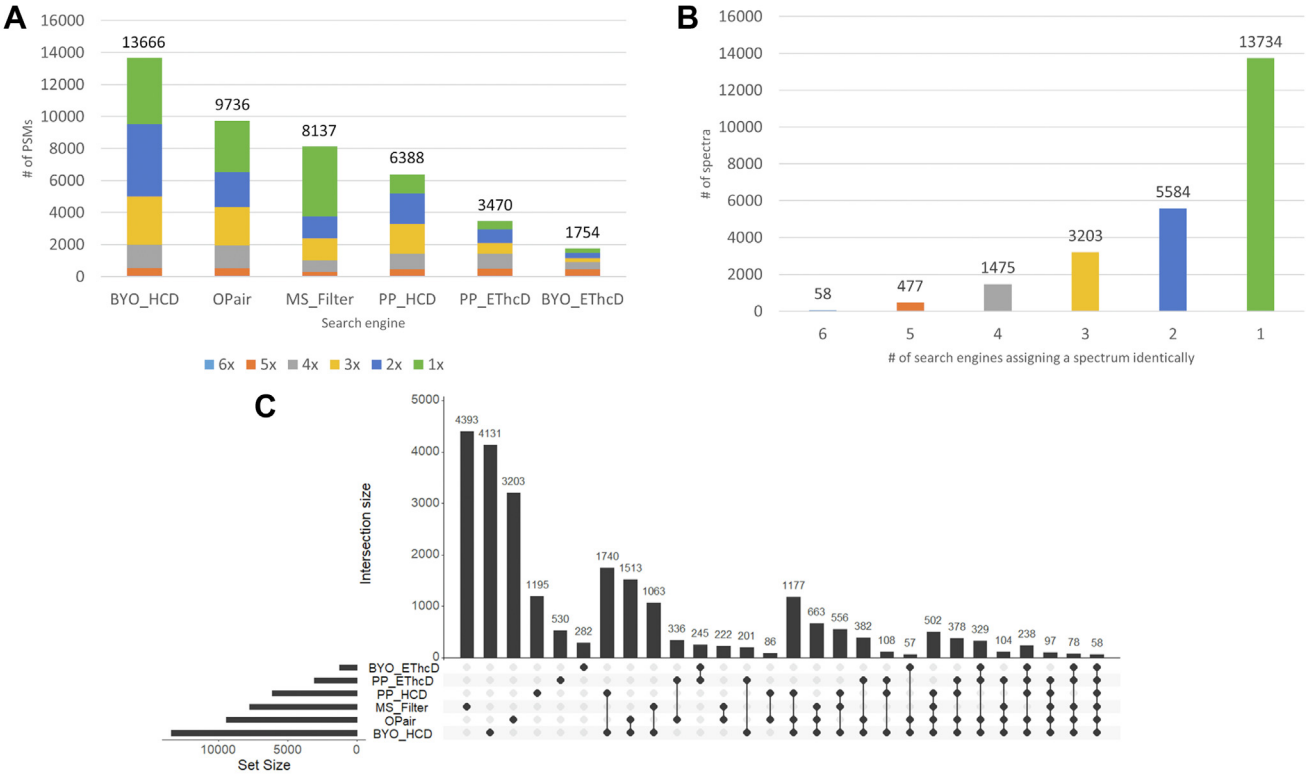


FIG. 2. **Overlap of O-glycopeptide assignments.** *A*, shows the number of O-glycopeptide PSMs delivered by each search engine and the MS-Filter program. The *colors* represent the number of different methods identifying the same peptide sequence and overall glycan composition. *B* and *C*, illustrate the degree of PSM assignment overlap globally and in an UpSet plot (42), respectively. *C*, BYO stands for Byonic. The plot shows only the overlap groups with at least 50 elements. The full plot can be seen in supplemental Table S23, sheet A. PSM, peptide spectrum match.

FIG. 3. **Distribution of overlapping and individual O-glycopeptide assignments.** Interpretation methods (listed on the *right*) and charge states (listed on the *top*) are depicted separately, *green and orange spots* indicate the individual (unique, *i.e.*, identified only by the specified approach) and shared (*i.e.*, the same peptide + glycan composition was identified by at least one other method) assignments, respectively. Glycan masses (in Dalton) are listed on the *x*-axes, whereas precursor *m/z* values are listed on the *y*-axes.

yields 25% less unique IDs, whereas the number of unique hits delivered by PP searches corresponds to ~30% of the Byonic results. The number of unique identifications derived from EThcD is even lower, ~13 and 7%, by PP and Byonic, respectively. When the same results were delivered by two different methods, only 685 assignments (~12% of all hits shared by two methods) are supported by independent HCD (here, O-Pair is considered as such) and EThcD assignments, and in ~8% of such pairs, only MS-Filter could "interpret" the HCD data. Similarly, among the 2833 (~51%) *exclusively* HCD data–supported (*i.e.*, O-Pair not included) "shared by two" assignments, MS-Filter delivered the support in ~40% of the cases (1093 of 2833). From the HCD- and EThcD-based O-Pair hits, 1821 (~33%) were supported by another HCD-based method, and MS-Filter delivered 222 of these. Last but not least, Byonic and PP EThcD searches arrived at the same conclusions for 245 PSMs (~4%). Among assignments shared by three methods, 556 identifications (17%) are still based solely on HCD spectra, and the majority, 1856 IDs (58%), were O-Pair hits, also supported exclusively by other HCD search results. Approximately one-fourth of the identifications is supported by data from both fragmentation methods individually. Obviously each assignment quartet is supported by data derived from both activation methods, but the biggest group, 502 assignments (34%), still consists of O-Pair hits supported by all the other HCD-based results (supplemental Table S23, sheet A).

From these observations, it seems evident that HCD data must be used in glycopeptide analysis. However, the discrepancy in spectrum interpretations by the different search engines also signals the need for significant improvements. Furthermore, we have to emphasize that only ET(hc)D enables the site localization and the differentiation between single and multiple glycosylation.

After compiling all O-glycopeptide assignments, we attempted to assess whether there are any significant differences in the urinary glycoproteome that indicate health or disease. Glycan- and peptide-level comparisons (supplemental Table S22, sheets C and D, respectively) tabulating the PSM numbers indicated some potential differences between the different donor groups. However, the manual data evaluation revealed several misassignments. The seemingly increased number of Tn antigens and T-antigens is frequently linked to N-glycosylation consensus motif–containing peptides (supplemental Table S22, sheet D), where either just the GlcNAc is linked to the Asn or a fucosylated GlcNAc and a methionine sulfoxide is hiding behind the HexNAcHex assignments. Since separate N-glycopeptide searches were also performed, comparison of the results indicated that double assignments do occur (supplemental Fig. S3 and supplemental Tables S11–S20, sheet M). Similarly, a uniquely high number of HexNAc$_2$HexNeuAc modifications was linked to one of the bladder cancer patients, the majority of it to Protein AMBP peptide, GPVPTPPDNIQVQENFNISR (supplemental Table S22, sheet D). Careful investigation of both HCD and EThcD data revealed that the glycan composition is a combination of a truncated N-glycan (GlcNAc) and a monosialo core 1 type O-glycan (GalNAcGalNeuAc) at Asn-36 and Thr-24, respectively (supplemental Fig. S4). This finding illustrates that potential N-glycopeptides also qualify as candidates for O-glycosylation. These examples also demonstrate the often ignored presence of truncated N-glycans that cannot be removed from the

peptides readily using PNGase F (30) and do interfere with the O-glycopeptide characterization. We also tried to compare the presence of O-acetyl sialic acids in the different samples and consistent detection of blood-type antigens. Scrutinizing those data, we encountered more complex discrepancies.

Suggestions for Improvement of Automated Data Interpretation of O-Glycopeptides

After realizing that the false identification rate is most likely much higher than suggested by the strict probability-based cutoff values, our focus shifted to a more careful evaluation of the search results, and as an outcome of this process, we attempted to establish rules that ideally should be followed in the future.

Most search engines were developed for the identification of tryptic peptides and gradually grew into more complex packages, permitting and evaluating nonspecific cleavages as well as a wide variety of covalent modifications. Glycosylation is among the most difficult PTMs to characterize, since it requires two different activation methods to achieve the basic structural assignment of a glycopeptide, that is, to decipher the sequence of the peptide (beam-type collision-induced dissociation [CID] [HCD]) and to establish the composition of the individual glycans and their attachment sites (ET(hc)D).

Moreover, the search space to be negotiated by the data interpretation software is unusually large when it comes to in-depth characterization of native glycopeptides in body fluids because of multiple reasons. First, proteolytic activity is rampant in urine; therefore, nonspecific enzymatic cleavages have to be considered: among the most reliable assignments, delivered by all, five, or just four methods, semitryptic sequences were identified in approximately 27, 33, and 55%, respectively. Second, a multientry glycan database has to be considered as variable modification. Specifically for urine, we have shown that in addition to the most common mucin type core-1 and core-2 glycans, at least 40 other structures may also decorate the peptides (18, 19). Third, because of the frequent occurrence of Thr and Ser, the majority of proteolytic peptides feature multiple potential glycosylation sites; therefore, multiple glycans may modify a single peptide in all kinds of combinations. Finally, both N-glycans and O-glycans might be attached to peptides, and telling apart these molecules is not as straightforward as might seem. The glycan databases for these modifications partly overlap, and distinct fragmentation properties are not considered currently by the search engines. Removing N-glycans by PNGase F prior to O-glycosylation analyses only partly addresses this issue as the enzyme is not efficient in the removal of small truncated N-glycans (30), and there is no similar universal endoglycosidase for the removal of O-glycans (*e.g.*, O-glycosylation needs also to be considered in N-glycosylation studies).

In the current study, we used software that was traditionally developed, and thus, the first step is finding the peptide that is glycosylated. This can be achieved successfully from either HCD or ET(hc)D spectra. We strongly feel that these data should be used in concert. For O-glycopeptide assignments, the presence of the gas-phase deglycosylated peptide ion, that is, Y_0 in the HCD spectra, should be required or at least highly valued, even if the glycopeptide is assigned from ET(hc)D data. Examples for the necessary presence of Y_0 are presented in supplemental Fig. S5 and in the Evaluation/Comments column of supplemental Table S22. The majority of Y_0 fragments were observed as singly or doubly charged rather abundant ions. However, long and low charge–density peptide sequences might produce Y_0 ions out of the monitored mass range, whereas highly glycosylated shorter peptides may yield less abundant Y_0 fragments. Statistical assessment of the presence and the intensity of the Y_0 ion as a function of peptide length or composition is out of the scope of the present study; however, we encourage future software development in this direction.

As the next step, we have to agree on a minimum number of peptide fragments to accept an assignment as reliable. Furthermore, these fragment ions have to represent both ends of the peptide, and b–y pairs formed by the preferential N-terminal cleavage at Pro residues should not count as independent proof, since these do not convey additional information about the sequence. Covering both ends of the sequence is especially important for long peptides that frequently do not yield abundant Y_0 fragments within the monitored mass range. The importance of this rule is illustrated with an example where a C-terminal sequence tag as well as Y_0 and Y_1 were detected in the HCD spectrum (Fig. 4, *upper panel*), and O-Pair identified the glycopeptide as LTLSGLSK modified with $HexNAc_2Hex_2NeuAc_2$ (supplemental Table S22, sheet A, file 18041707, scans 5632 and 5854), even when the EThcD interpretation by PP pointed to VATTVISK modified with two trisaccharides (Fig. 4, *lower panel*). This misidentification was not a one-hit-wonder; HCD data most likely derived from VATTVISK were assigned to a series of different sequences: LTLSGLSK, SILSALSK, GLTVTLSK, VLTTGLSK, each featuring the same accurate mass: 818.498 and the same 3 C-terminal residues. HCD data with a more comprehensive fragment ion series enabled the unambiguous assignment of VATTVISK, as illustrated on the later eluting isomeric glycoform bearing the core-2 hexasaccharide (supplemental Fig. S6).

This example also illustrates that information from the spectrum obtained by the other activation method (*e.g.*, confirmatory or contradictory information in the HCD and ET(hc)D data of the same precursor) could make or break the original assignment. However, presently O-Pair performs the primary identification from HCD spectra, and the corresponding EThcD data are considered for further support only and for glycan localization assignment(s). We believe that an alternative, a reverse O-Pair workflow, is also desirable, since we encountered several instances where the contribution of

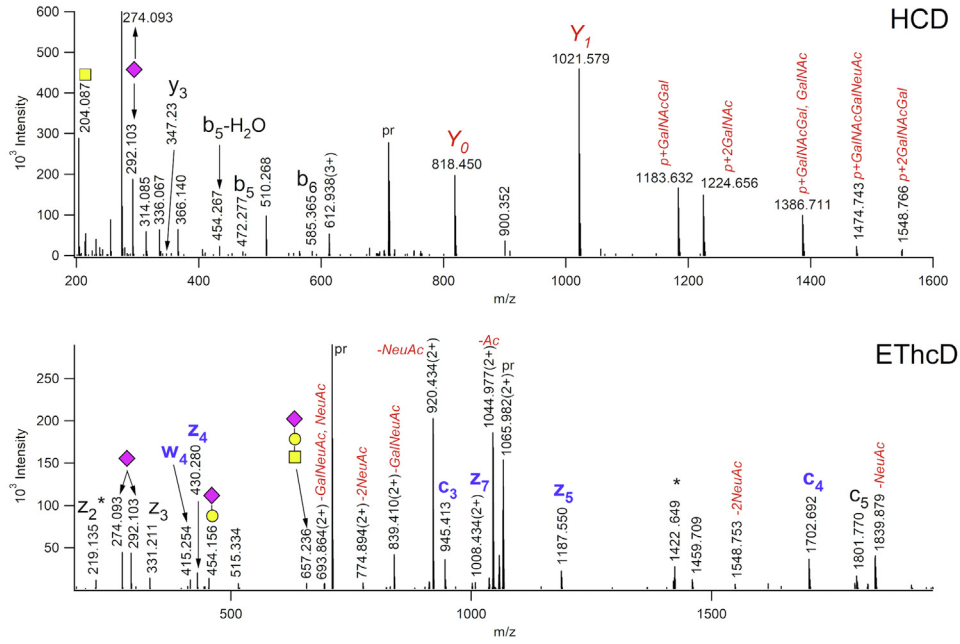


FIG. 4. **MS/MS data of VAT(NeuAcGalGalNAc)T(NeuAcGalGalNAc)VISK misidentified as LTLS(HexNAc$_2$Hex$_2$NeuAc$_2$)GLSK.** HCD data fit both sequences, whereas ions supporting only the first glycopeptide, and also providing site assignments, are highlighted in *blue*. Ions related to glycan fragmentation (depicted in *red*) fit both glycopeptides. "pr" labels the precursor ion, *m/z* 710.989(3+) and its charge-reduced form. The *asterisk* (*) indicates the singly charged form of a coeluting 2+ ion. "p" as well as Y_0 stands for the peptide. Sugar units still attached to it are listed (in *red*), oxonium ions are labeled with SNFG symbols. HCD, higher-energy collision dissociation; SNFG, symbol nomenclature for glycans.

HCD was very limited, whereas EThcD provided sequence coverage as well as glycan localization information. As an example, from the HCD-EThcD (supplemental Table S22, file 18041713, scans 5552 and 5553) spectra shown in Figure 5, LNAT(HexNAc$_2$Hex$_2$NeuAc$_2$Ac)LR was assigned by O-Pair. However, the HCD data do not feature the corresponding Y_0 ion (*m/z* 687.415), and the peptide sequence ions on which the identification is based are of very weak intensity. Since the characteristic oxonium ions (316, 334) of *N*,*O*-diacetyl neuraminic acid are also absent, the glycan composition has to be incorrect. Starting with the EThcD data could have prevented this misidentification, as PP identified IPTNAR bearing the blood-type A antigen (HexNAc$_3$Hex$_2$FucNeuAc) showing nearly complete sequence coverage, and the Y_0 ion detected in HCD provided further support.

Obviously, once the peptide is assigned, the mass of the glycan composition is also revealed, usually unambiguously. The Y_0 requirement eliminates interferences between the potential peptide modifications and glycan composition changes. For example, without knowing the peptide mass accurately, a 16 Da mass difference may be translated into a Met oxidation or a Fuc-Hex or a NeuAc-NeuGc difference in the oligosaccharide as discussed previously. Thus, the glycan composition calculated from the mass difference of Y_0 and the measured mass of the molecule should be translated into individual glycans. Obviously, the higher the level of glycosylation the harder this job becomes. To make this process a bit easier, we could and should rely on diagnostic oxonium ions to ascertain the presence of certain building units in the modifying glycans. The HexNAc-related *m/z* 204.087 ion already became the hallmark of most glycosylation studies, enabling the most efficient HCD-product ion–dependent ET(hc)D data acquisition approach (31) that was successfully applied in both N-glycosylation and O-glycosylation analyses. Sialic acids also produce abundant diagnostic ions that could also be exploited during automated data interpretation. Accordingly, when the diagnostic 292 and 274 ions are not present in the HCD spectra, Neu5Ac containing glycans should not be permitted in the assignments. Similarly, glycolyl, O-acetyl, or *O*,*O*-diacetyl sialic acids produce diagnostic fragments (18, 32). High mass accuracy measurement affords unambiguous detection of all these ions. We attempted to identify data derived from O-acetyl sialic acid–containing glycopeptides by filtering for the presence of *m/z* 316.1027, formed *via* water loss from the Neu5,9Ac$_2$ oxonium ion. About ~4% of the O-glycopeptide identifications featured this ion among the top 20 most abundant fragment ions in HCD. However, manual evaluation of a subset of spectra indicated that some correct identifications did not make the intensity cutoff (supplemental Table S22, sheet A, evaluations/comments). At the same time, in 47 assignments, the identified glycan did not feature *O*-acetyl sialic acid, despite the strong presence of its diagnostic ions (supplemental Table S22, sheet A). Obviously, in a complex mixture, precursor ion

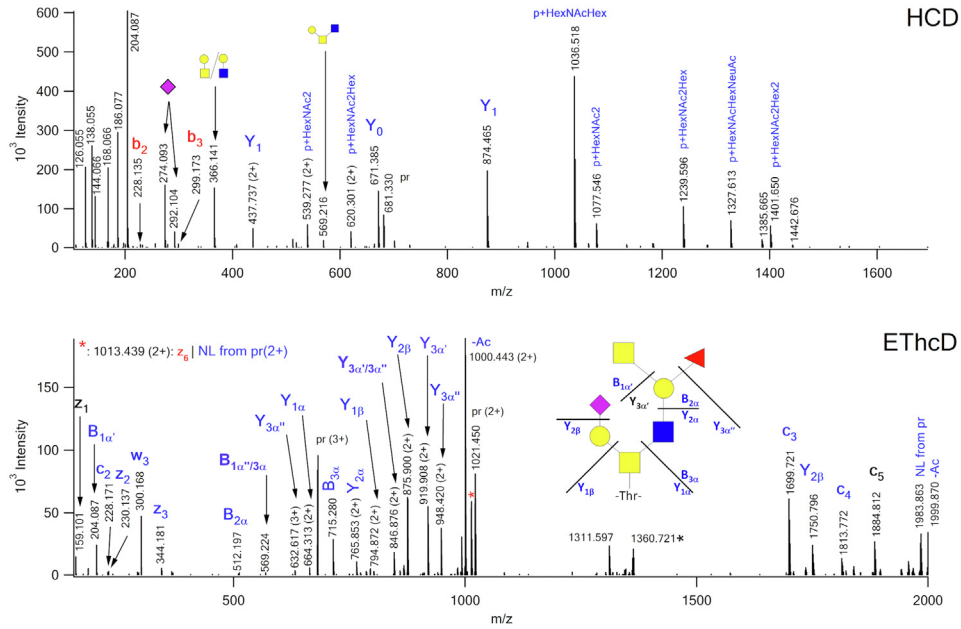


FIG. 5. **HCD and EThcD data of *m/z* 681.300 (3+) identified as LNAT(HexNAc$_2$Hex$_2$NeuAcNeuAcAc)LR by O-Pair (score: 15.1, Q: 0.0025).** A reverse O-Pair workflow, that is, starting from the EThcD spectrum could have delivered the correct assignment as IPT(HexNAc$_3$Hex$_2$FucNeuAc)NAR (from PP search results). The size of the peptide is verified by the Y_0 ion in the HCD. Fragments supporting the correct assignment are printed in *blue*, shared ions are in *black*, unique fragments for the original (incorrect) assignment are in *red*. In the EThcD spectrum, the ion marked with * represents the charge-reduced form of a coeluting 2+ precursor. EThcD, electron-transfer/higher-energy collision dissociation; HCD, higher-energy collision dissociation.

interference might be also blamed for the detection of such "reporter ions," but we believe it is worthwhile to have a second look at the original assignments. At the same time, the lack of the aforementioned diagnostic fragments in the HCD spectra is a good indication that such units are not part of the modifying glycans.

Additional glycan fragment ions, though might not be specific for any glycan structure, used in combination may further enhance the reliability of glycopeptide identifications. It is again beneficial to use HCD and EThcD data in concert. Lower *m/z* glycan ions tend to be more abundant in HCD, whereas in EThcD, single-bond cleavages dominate, and even larger glycans up to seven monosaccharide units might survive the mild activation (15% NCE) (19). The fragment ions *m/z* 407 and 569 indicate the presence of a glycan with HexNAc$_2$ and HexNAc$_2$Hex connectivity, respectively and may help to distinguish a core-2 glycan from two core-1 type modifications (29). These ions tend to show up in HCD, whereas they are typically missing from EThcD as multiple glycosidic bond cleavages are necessary to generate them. These observations are illustrated with the earlier mentioned VATTVISK glycoforms: the peptide carrying two core-1 trisaccharides does not display the aforementioned ions (Fig. 4), whereas the peptide decorated with the core-2 hexasaccharide, albeit weakly, does produce these ions in HCD (supplemental Fig. S6). Fucose-containing glycans also may produce characteristic ions. While the HexNAcHexFuc oxonium ion at *m/z* 512 can be observed both in HCD and EThcD, larger Fuc-containing B ions characteristic to the A and B antigens (*m/z* 715 for HexNAc$_2$HexFuc and 674 for HexNAcHex$_2$Fuc, respectively) can frequently be detected in EThcD (Figs. 5 and S7). Similarly, glycans with disialic acid units might yield oxonium ions at *m/z* 583 for Neu5Ac2 and at *m/z* 625 for Neu5AcNeu5,9Ac$_2$ in EThcD (18), whereas their respective water loss ions might be detected in HCD. Similarly, the Y ions especially in EThcD, such as Fuc, HexNAc, and Hex losses from the precursor ion, may reveal the identity of terminal structures (Fig. 5, scheme). While all these bits of information may help to correctly assign the glycan composition and even the individual glycans, our efforts might be undermined by the fact that Na- and K-adduct formation may occur, and/or monoisotopic precursor masses cannot always be determined unambiguously, that is, the mass difference between the peptide and the precursor ion may be misleading (adduct formation) or may not be accurate (faulty peak-picking). We noticed that several abundant glycan fragments retained the metal ion in Na-adduct spectra, their usefulness in the assignments has to be evaluated further. Faulty peak-picking is harder to correct; the most typical example for this is the recurring Fuc2 *versus* NeuAc question. As mentioned earlier, Fuc loss indicating Y-fragments may hold the answer.

Preferential single-bond cleavages in EThcD may yield larger B ions that can confirm the identity of the glycans. For example, *m/z* 1313 confirms that a core-2 hexasaccharide decorates the peptide (supplemental Fig. S6). We did detect such B-

fragments: core-1 tetrasaccharides and pentasaccharides (*m/z* 948, 990, 1032 for GalNAc(NeuAc)GalNeuAc with 0 to 2 Neu5,9Ac_2 and 1239, 1281, 1323 for GalNAc($NeuAc_2$)GalNeuAc with 0 to 2 Neu5,9Ac_2) and core-2 hexasaccharides (*m/z* 1313, 1355, 1397 for GalNAc(GlcNAcGalNeuAc)GalNeuAc with 0 to 2 Neu5,9Ac_2 and 1168 for GalNAc(GlcNAcGalFuc)GalNeuAc representing a H antigen capping unit), even a core-2 heptasaccharide decorated with the A antigen (*m/z* 1371 for GalNAc(GlcNAcGalNeuAc)Gal(Fuc)GlcNAc) could sporadically be detected (18, 19, 33). These ions are typically of low intensity; hence, it is unlikely that precursor ion interference would be responsible for their presence. Thus, their detection could be rewarded during automated data interpretation.

Preferential single-bond cleavages in EThcD also can provide further insights into the glycan structure. Sialic acid–related B ions showed that the O-acetyl sialic acid in core-1 type glycans was Gal-linked in the tetrasaccharide but decorated the GalNAc when present as a terminal unit in disialic acid in the pentasaccharide (18). Furthermore, the single-bond cleavages also enabled the assignment of the O-acetyl sialic acid position in chromatographically resolved glycoforms bearing core-2 hexasaccharides (18) and characterization of isomeric oligosaccharides displaying the blood-type antigen A on different arms of the core-2 glycan (19). Since in EThcD, the Y-ion formation is also controlled by preferential single-bond cleavages, the terminal positions of single sugar units as well as multiunit assemblies can be verified from these fragments. We demonstrated that based on the fragmentation pattern, isomeric core-2 glycans can be distinguished, for example, it can be determined whether the A blood-type determinant is located on the core GalNAc or on the GlcNAc linked to it (18, 19).

In summary, HCD and EThcD data should be used in concert. We recommend the following for more reliablePeptide identification

1. Y_0 rule—fine-tuned, based on further statistical analysis (HCD).
2. Minimum five independent backbone fragments, covering both termini—data from both activation methods should be considered.
3. "Reverse O-Pair workflow," that is, starting with the EThcD data.

Glycan composition assignment

1. Y_0 rule—this is the only way to exclude fortuitous peptide modifications (HCD).
2. Diagnostic fragment ion requirement for the different sialic acids (HCD and EThcD).
3. Reward for diagnostic building unit losses, such as HexNAc, Hex, and Fuc loss from the precursor (EThcD).

Individual glycan assignment

1. Reward for the detection of some characteristic oligosaccharide fragments in HCD and EThcD—should be fine-tuned by statistical analysis (29).
2. EThcD Y and B fragment evaluation, considering the single-bond cleavage rule (at NCE 15%).

Finally, we have some recommendations for telling apart data acquired on N-linked or O-linked glycopeptides. While larger glycans can frequently be rendered to the appropriate site by knowing the glycan biosynthetic pathways, some smaller structures can easily be misinterpreted. For example, the truncated GlcNAc1-2 glycan can be misinterpreted as one to two Tn antigens (GalNAc), or the paucimannose, GlcNAc2Man2 can be overlooked as two T antigens (GalNAcGal) or an asialo core-2 O glycan (Gal(GalGlcNAc)GalNAc).

The different relative intensities of Y ions observed in HCD could be exploited here. In HCD, the Y_1 ion is always more abundant than Y_0 for N-linked glycopeptides, and respective peptide sequence ions typically carry the innermost GlcNAc. On the other hand, O-linked structures with up to five to six monosaccharide units tend to produce more abundant Y_0 ions, and peptide sequence ions are predominantly present fully deglycosylated. Therefore, glycopeptides with only N-linked structures or carrying both N- and O-glycans could be identified based on the Y_1/Y_0 ratio, and peptide sequence ions observed in HCD and EThcD can resolve the finer structural details. Furthermore, the oxonium ion intensity profile of GlcNAc and GalNAc is different (34) and can be exploited to strengthen the identifications especially if only either GlcNAc or GalNAc is present in the glycan.

Even without further software development, the presence of an N-glycosylation consensus motif and the ion intensity ratio of *m/z* 138 and 144 should be included in the results output of the software as already performed by O-Pair.

O-Glycosylation Landscape of Selected Proteins

The rules drawn previously are based on our observations during the manual interpretation of hundreds of MS/MS spectra. Evidently, manual evaluation of all identifications is unattainable; therefore, we decided to investigate a few proteins in detail. Partly, to further scrutinize the reliability of the assignments and also to validate modification sites and characterize microheterogeneity. Probably, the most exciting subset of proteins are those with blood group antigens as currently our knowledge on the occurrence of these structures on individual proteins and sites is quite limited.

Altogether, 683 PSMs (derived from 405 precursor selections), 97 sequences from 49 proteins, represented glycopeptides carrying ABO blood group antigens on core-2 O-glycans ($HexNAc_2Hex_2Fuc_1NeuAc_1$, $HexNAc_3Hex_2Fuc_1NeuAc_1$, or $HexNAc_2Hex_3Fuc_1NeuAc_1$ for the H, A, or B antigen, respectively). Strangely, over half of the PSMs (347 of 683) signaled the presence of B antigens, although seven of the 10 donors were of blood groups A or O (the blood group

of two donors was unknown; supplemental Table S1). We also observed that the B antigen carrying glycoforms of ITIH4 peptides coeluted with glycoforms bearing $HexNAc_2Hex_2NeuAc_2$. Glycoforms featuring more sialic acids usually elute later than less acidic ones when formic acid is used in the mobile phase. Thus, we believe that the +1329 Da glycoform here represents an ammonium adduct of that hexasaccharide (additive mass: 1312.455 + 17.027 = 1329.482) instead of the assigned $HexNAc_2Hex_3Fuc_1NeuAc_1$ structure (additive mass: 1329.471). Incomplete removal of the ammonium salt used during the affinity chromatography makes such adduct formation feasible. We wanted to identify a reliable set of site-specific O-linked blood-type identifications, and since we have already encountered several dubious assignments associated with longer sequences, we have decided to remove peptides longer than 16 residues from the list. Furthermore, a candidate that featured the N-glycosylation sequon was also discarded along with those assignments that were not corroborated by both HCD and EThcD data. The remaining subset consists of 10 peptides representing seven proteins (fractalkine [UniProt ID: P78423], insulin-like growth factor II [UniProt ID: P01344], macrophage colony-stimulating factor 1 [UniProt ID: P09603], protein HEG homolog 1 [UniProt Q9ULI3], protein YIPF3 [UniProt ID: Q9GZM5], SPARC-like protein 1 [UniProt ID: Q14515], and transforming growth factor [TGF]-beta receptor type 2 [UniProt ID: P37173]) (supplemental Table S22, sheet E). All identifications representing donors with known ABO blood groups matched the expected glycan structures. PSMs from the two patients of unknown blood types (patients 3 and 10, supplemental Table S21) unequivocally indicated that the blood group of these patients is B. Now if we introduce back all PSMs of the 10 peptides with the blood-type antigens, the list contains 188 IDs, with only one MS-Filter hit indicating a B-antigen structure incorrectly. These results indicate that the corresponding sites in the aforementioned proteins are consistently carrying the ABO blood group epitopes, and these glycoforms are of significant abundance. Thus, these structures also have to be considered during comprehensive characterization of glycoproteins, and blood typing prior to biomarker studies is highly advisable.

Furthermore, we examined all O-glycopeptide PSMs related to five of the aforementioned proteins: SPARC-like protein 1, insulin-like growth factor II, fractalkine, TGF-beta receptor type 2, and protein HEG homolog 1. Our observations are included in the evaluation/comments column of supplemental Table S22 (sheet A). Please note that we considered all available data (primarily both HCD and EThcD spectra and occasionally MS1 data and retention times as well) when deciding whether an assignment was correct.

Earlier, we have already demonstrated the microheterogeneity of the C-terminal peptide of YIPF3 (18, 19), whereas macrophage colony-stimulating factor 1 did not yield unambiguous sequence confirmation for the blood group–related glycopeptides.

We compared our findings to data listed in the UniProt database and in two previous studies that acquired site-specific O-glycosylation data. UniProt entries for the five proteins selected were compiled from three reports (35–37), all applying the same capture-and-release workflow to enrich sialic acid–containing glycoproteins. Using CID and ECD activation, the authors characterized glycosylation in human cerebrospinal fluid (CSF) (35, 37) and urine (36). Recently, Zhao *et al.* (15) also published data on the urinary O-glycoproteome. O-glycopeptides were enriched from tryptic digests using hydrophilic interaction liquid chromatography and identified sialic acid–containing glycoforms separately from glycoforms void of this unit using HCD and EThcD data acquired on PNGase F-treated samples. King *et al.* (38) analyzed O-glycopeptides isolated by lectin affinity chromatography from desialylated proteolytic digests of plasma, platelets, and endothelial cell samples. Although the sample source was different in this study, the authors reported on the modification of ~650 proteins; therefore, we decided to include their results as reference for the modification sites.

In summary, the results from different sources showed limited overlap (supplemental Fig. S8). For the five proteins selected, 49 glycosylation sites were identified, but only two were reported by all studies, and an additional four were found in at least three studies. King *et al.* (38) and the current study contributed the most unique sites, whereas most of the sites listed in UniProt, based on urine (15, 36) or CSF analysis (35, 37), were confirmed by other studies. Figures 6 and 7 show the glycan structures assigned unambiguously to the listed glycosylation sites; sheet F of supplemental Table S22 specifies those additional glycan compositions that were detected by us on certain sequence stretches but could not be resolved from the data.

The O-glycosylation studies quoted previously analyzed plasma, urine, and CSF and followed different protocols. Thus, not entirely surprising that the resulting findings were also different. The sialic acid capture upon release yields all asialo structures. Thus, we lose diversity, but at the same time, this simplification increases the sensitivity by combining the signals of originally different glycoforms, and the glycan database will also shrink accordingly. The sites listed in UniProt were identified using this enrichment method, a less sensitive mass spectrometric methodology (CID-MS3 and ECD), but the data interpretation was performed very carefully (35–37). The study by King *et al.* (38) used a selective enrichment method, sacrificing diversity for more efficient glycosylation site determinations. The closest to our approach was the analysis of urinary proteome (15). Hydrophilic interaction liquid chromatography enrichment of the glycopeptides, followed by HCD and EThcD analysis, could have allowed the identification of most structures we found. The glycan database they built did not feature all the glycan structures we reported and

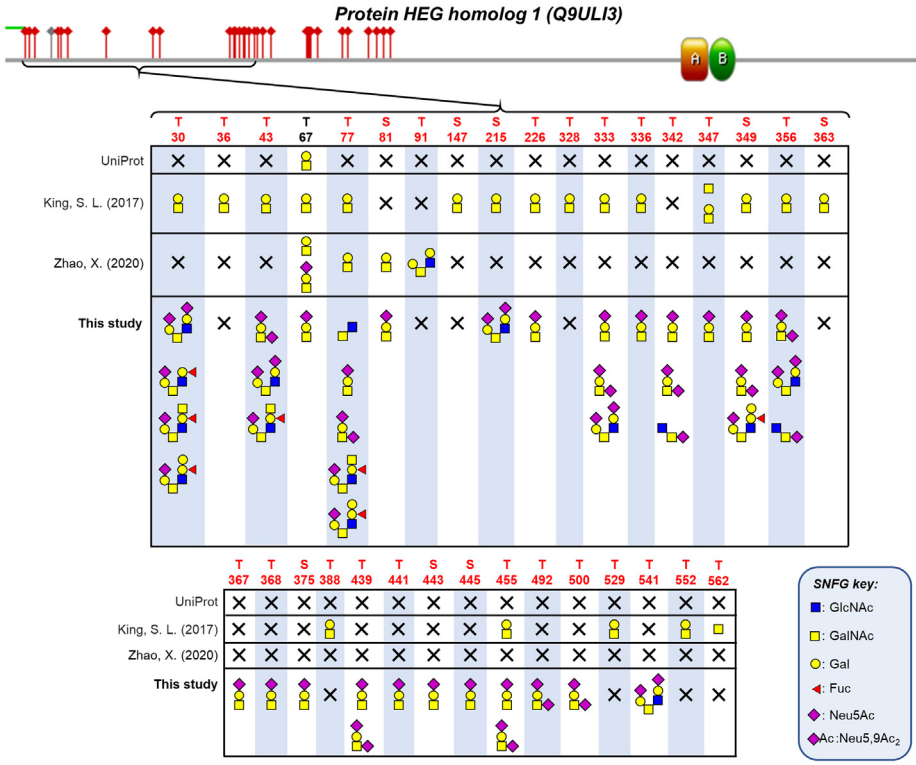


FIG. 6. **Comparison of the O-glycan landscape of Protein HEG homolog 1.** O-glycan structures are illustrated following the SNFG recommendations. Glycosylation sites that are reported in the UniProt database are indicated with *gray markers*. O-glycosylation sites not reported in UniProt are indicated with *red markers*. The *green line* at the N terminus denotes the signal peptide. Domain- and region-specific information of the protein was collected from the UniProt database: A: EGF-like 1 and B: EGF-like 2; calcium binding. EGF, epidermal growth factor; SNFG, symbol nomenclature for glycans.

searched for. In addition, the vast majority of their assignments are based solely on HCD data that usually does not permit accurate individual glycan and localization site assignments. In addition, considering nonspecific cleavages would have been essential for the identification of certain glycosylation sites. For example, the N-terminal peptides of the TGF beta and SPARC-like proteins are not tryptic but are the products of enzymatic processing.

SUMMARY AND CONCLUSIONS

WGA affinity chromatography fractions enriched in both N- and O-glycopeptides from the urine of 10 individuals were analyzed acquiring HCD and HexNAc oxonium ion–triggered EThcD spectra. Using these data, we established some rules about EThcD fragmentation and reported the presence of more than 30 unexpected sialoglycans, among them even some isomer pairs (18, 19). These discoveries have proven that a preliminary “wild card” or “open mass addition” search might reveal novel or unexpected components that would be otherwise overlooked. For example, O-acetylated neuraminic acids reported by us would not survive the oligosaccharide release; thus, their presence would not be suspected even after a glycan-pool analysis.

In this article, we present an in-depth analysis of all the data files focusing on O-glycosylation. We share our raw files as well as our data interpretation methods and lists. We are aware that our compilations contain several incorrect assignments as pointed out previously, but these were not discarded because they serve as illustration of the occurrence, frequency, and potential reasons for misinterpretation. Unfortunately, these may occur regularly in any large-scale glycopeptide analysis. A community study on a much simpler and mostly N-glycopeptide mixture already demonstrated that even with the same software very different results can be achieved and offered guidelines for setting up search parameters for different purposes (39), although the big question how to estimate false positive rate for glycopeptides has not been solved yet.

In our study, we have chosen rather limiting parameters for database searches and strict acceptance criteria. PTM analysis in general leads to significant search space expansion. O-glycosylation is even more problematic in this aspect, since it represents multiple different structures, multiple potential sites within the same peptide sequence, no consensus sequences, and distinctive fragmentation interfering with the assignment of the underlying amino acid sequence.

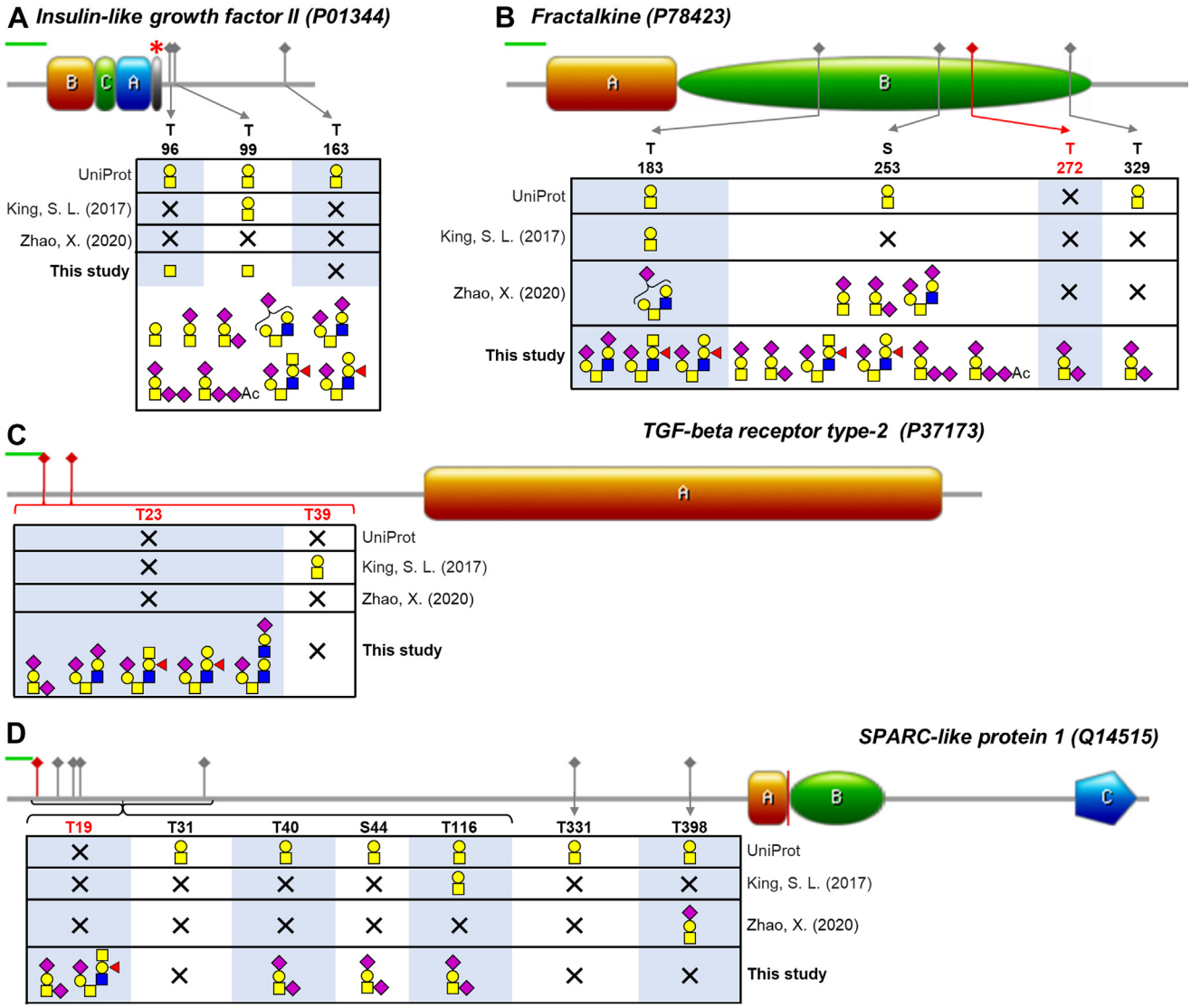


FIG. 7. **Comparison of the O-glycan landscape of four additional proteins identified with O-glycans containing ABO blood group antigens.** O-glycan structures are illustrated following the SNFG recommendations. Glycosylation sites that are reported in the UniProt database are indicated with *gray markers*. O-glycosylation sites not reported in UniProt are indicated with *red markers*. The *green line* at the proteins' N termini denotes the signal peptide. Domain- and region-specific information of the proteins was collected from the UniProt database: (*A*) B, C, A, and D (marked with a *red asterisk*) regions of the insulin-like growth factor II. *B*, regions of fractalkine: A: chemokine and involved in interaction with ITGAV:ITGB3 and ITGA4:ITGB1, B: mucin-like stalk. *C*, domain A of TGF-beta receptor type-2: protein kinase. *D*, domains of SPARC-like protein 1. A: Follistatin-like, B: Kazal-like, C: EF-hand. The A and B domains overlap. The *vertical red line* shows the start of the B domain. SNFG, symbol nomenclature for glycans.

Considering an increased glycan database and permitting each glycan as modifier multiple times increases the search space exponentially, especially when other search space widening factors characterize the samples, such as nonspecific cleavages, N- and O-glycosylation on the same sequence, glycan compositions that may represent single or multiple glycans, glycan combinations that may represent different individual glycans, or positional isomers or glycan isomers or the combinations thereof. Thus, we used our newly gained knowledge about the glycan pool in an iterative fashion. Permitting the extended glycan list on a prefiltered protein database seems reasonable and speeds up the process. Limiting the number of modifications also helps but will eliminate correct identifications as well. We used Byonic for the interpretation of both HCD and EThcD data aiming the identification of unmodified sequences as well as N- and O-glycopeptides as this is a commercially available popular software used by the glycoscience research community. We explored additional tools for the most comprehensive O-glycopeptide characterization. PP was used to assign

O-glycopeptides from both HCD and EThcD data. In addition, we used the MS-Filter program to identify potential glycoforms based on the presence of diagnostic Y_0 and Y_1 fragments. Finally, we tested the recently developed O-Pair that uses the data provided by the two activation methods in concert. We have to emphasize that despite the high number of glycopeptide identifications, HCD by itself is not suitable for O-glycopeptide characterization (12, 17), since the number, size, composition, and site localization of the modifying glycans cannot be determined. This applies even when an endoglycoprotease is used for the generation of O-glycopeptides (40, 41) as all enzymes may miss cleavages especially in densely glycosylated sequence stretches, and certain glycans may prevent the proteolysis. In addition, in certain cases, partial digestion could be beneficial even with these enzymes. For example, to characterize macroheterogeneity in combination with site-specific variations. Unfortunately, EThcD that may deliver much more comprehensive information about glycopeptides is a much less efficient activation method. Thus, we have to use both sets of fragmentation data together. The attempt of O-Pair at this job is a very promising one but far from complete. We have already discussed the advantages and drawbacks of the individual methods. Here, we would like to draw attention to some common issues. The assignment of long glycopeptides is definitely a challenge, and it would require further investigation of what software tools or novel analytical methods could improve this situation. Similarly, the presence of both N- and O-glycosylation in a mixture or even on the same peptide represents an unresolved issue, although some initial steps have been taken to address the latter problem. Last but not least, we think that a scoring approach similar to evaluating cross-linked peptides by PP could be more efficient. The amino acid sequence should be assigned based on the peptide fragments, but the glycan B and Y fragments also have to be evaluated. The assignment would be considered reliable only if both "halves" of the molecule received a convincing score. As pointed out in the discussion, the presence of certain diagnostic glycan fragments should be required for even considering structures containing the corresponding sugar units. The retention times of the different glycoforms also could be used to strengthen or weaken certain assignments, or as shown previously, it can be used to indicate potential noncovalent adduct formation. Obviously, further datasets and detailed investigations are necessary to establish the appropriate rules. We feel that besides innovative computer programs, human intervention is still necessary to assess the reliability of the new data interpretation methods. We hope that our data and our observations will aid the development of such new tools.

DATA AVAILABILITY

LC–MS raw files, peak lists, and the Byonic and O-Pair result files have been uploaded to Massive (https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp). Project identifier: MSV000090536. Download *via* FTP: ftp://massive.ucsd.edu/MSV000090536/.

O-Pair search results (.psmtsv) can be viewed using O-Pair. Byonic search results (.byrslt) can be viewed in Byonic Viewer.

The search keys for the MS-Viewer files showing the results of PP and MS-Filter searches are listed in supplemental Table S26.

Supplemental data—This article contains supplemental data (15, 18, 19, 24, 26, 28, 35–38).

Acknowledgments—This work was supported by the following grants: the Economic Development and Innovation Operative Programmes (grant nos.: GINOP-2.3.2-15-2016-00001 and GINOP-2.3.2-15-2016-00020). We thank the ELKH Cloud for housing our Protein Prospector server. Hungarian Centre of Excellence for Molecular Medicine has received funding from the European Unions's Horizon 2020 research and innovation program under grant agreement no. 739593.

Author contributions—K. F. M. and Z. D. conceptualization; K. F. M. and Z. D. methodology; A. P., K. F. M., and Z. D. formal analysis; A. P., K. F. M., and Z. D. investigation; I. E. K. and Z. D. resources; A. P. and K. F. M. writing–original draft; K. F. M. and Z. D. writing–review & editing; A. P. and K. F. M. visualization; K. F. M. and Z. D. supervision; K. F. M. and Z. D. funding acquisition.

Conflict of interest—The authors declare no competing interests.

Abbreviations—The abbreviations used are: CID, collision-induced dissociation; CSF, cerebrospinal fluid; EThcD, electron-transfer/higher-energy collision dissociation; HCD, higher-energy collision dissociation; MS, mass spectrometry; NCE, normalized collision energy; PP, Protein Prospector; PSM, peptide spectrum match; PTM, post-translational modification; TGF, transforming growth factor; WGA, wheat germ agglutinin.

Received May 25, 2022, and in revised form, October 17, 2022
Published, MCPRO Papers in Press, November 9, 2022, https://doi.org/10.1016/j.mcpro.2022.100439

REFERENCES

1. Harris, R. J., Leonard, C. K., Guzzetta, A. W., and Spellman, M. W. (1991) Tissue plasminogen activator has an O-linked fucose attached to threonine-61 in the epidermal growth factor domain. *Biochemistry* **30**, 2311–2314
2. Harris, R. J., van Halbeek, H., Glushka, J., Basa, L. J., Ling, V. T., Smith, K. J., *et al*. (1993) Identification and structural analysis of the tetrasaccharide NeuAc alpha(2–>6)Gal beta(1–>4)GlcNAc beta(1–>3)Fuc alpha 1–>O-linked to serine 61 of human factor IX. *Biochemistry* **32**, 6539–6547
3. Hofsteenge, J., Müller, D. R., de Beer, T., Löffler, A., Richter, W. J., and Vliegenthart, J. F. (1994) New type of linkage between a carbohydrate and

a protein: C-Glycosylation of a specific tryptophan residue in human RNase Us. *Biochemistry* **33**, 13524–13530
4. Hashii, N., Suzuki, J., Hanamatsu, H., Furukawa, J. I., and Ishii-Watabe, A. (2019) In-depth site-specific O-Glycosylation analysis of therapeutic Fc-fusion protein by electron-transfer/higher-energy collisional dissociation mass spectrometry. *Biologicals* **58**, 35–43
5. Stavenhagen, K., Gahoual, R., Dominguez Vega, E., Palmese, A., Ederveen, A. L. H., Cutillo, F., *et al*. (2019) Site-specific N- and O-glycosylation analysis of atacicept. *MAbs* **11**, 1053–1063
6. Napoletano, C., Steentoff, C., Battisti, F., Ye, Z., Rahimi, H., Zizzari, I. G., *et al*. (2020) Investigating patterns of immune interaction in ovarian cancer: probing the O-glycoproteome by the macrophage galactose-like C-type lectin (MGL). *Cancers* **12**, E2841
7. Pirro, M., Schoof, E., van Vliet, S. J., Rombouts, Y., Stella, A., de Ru, A., *et al*. (2019) Glycoproteomic analysis of MGL-binding proteins on acute T-cell leukemia cells. *J. Proteome Res.* **18**, 1125–1132
8. Chernykh, A., Kawahara, R., and Thaysen-Andersen, M. (2021) Towards structure-focused glycoproteomics. *Biochem. Soc. Trans.* **49**, 161–186
9. Darula, Z., and Medzihradszky, K. F. (2009) Affinity enrichment and characterization of mucin core-1 type glycopeptides from bovine serum. *Mol. Cell. Proteomics* **8**, 2515–2526
10. Darula, Z., Sherman, J., and Medzihradszky, K. F. (2012) How to dig deeper? Improved enrichment methods for mucin core-1 type glycopeptides. *Mol. Cell. Proteomics* **11**. https://doi.org/10.1074/mcp.O111.016774
11. Darula, Z., Sarnyai, F., and Medzihradszky, K. F. (2016) O-glycosylation sites identified from mucin core-1 type glycopeptides from human serum. *Glycoconj. J.* **33**, 435–445
12. Darula, Z., and Medzihradszky, K. F. (2018) Analysis of mammalian O-Glycopeptides-We have made a good start, but there is a long way to go. *Mol. Cell. Proteomics* **17**, 2–17
13. Kim, J., Ryu, C., Ha, J., Lee, J., Kim, D., Ji, M., *et al*. (2020) Structural and quantitative characterization of mucin-type O-glycans and the identification of O-glycosylation sites in bovine submaxillary mucin. *Biomolecules* **10**, 636
14. Wang, S., Qin, H., Mao, J., Fang, Z., Chen, Y., Zhang, X., *et al*. (2020) Profiling of endogenously intact N-linked and O-linked glycopeptides from human serum using an integrated platform. *J. Proteome Res.* **19**, 1423–1434
15. Zhao, X., Zheng, S., Li, Y., Huang, J., Zhang, W., Xie, Y., *et al*. (2020) An integrated mass spectroscopy data processing strategy for fast identification, in-depth, and reproducible quantification of protein O-glycosylation in a large cohort of human urine samples. *Anal. Chem.* **92**, 690–698
16. Kawahara, R., Ortega, F., Rosa-Fernandes, L., Guimarães, V., Quina, D., Nahas, W., *et al*. (2018) Distinct urinary glycoprotein signatures in prostate cancer patients. *Oncotarget* **9**, 33077–33097
17. Riley, N. M., Malaker, S. A., Driessen, M. D., and Bertozzi, C. R. (2020) Optimal dissociation methods differ for N- and O-glycopeptides. *J. Proteome Res.* **19**, 3286–3301
18. Darula, Z., Pap, Á., and Medzihradszky, K. F. (2019) Extended sialylated O-glycan repertoire of human urinary glycoproteins discovered and characterized using electron-transfer/higher-energy collision dissociation. *J. Proteome Res.* **18**, 280–291
19. Pap, A., Tasnadi, E., Medzihradszky, K. F., and Darula, Z. (2020) Novel O-linked sialoglycan structures in human urinary glycoproteins. *Mol. Omics* **16**, 156–164
20. Rossez, Y., Maes, E., Lefebvre Darroman, T., Gosset, P., Ecobichon, C., Joncquel Chevalier Curt, M., *et al*. (2012) Almost all human gastric mucin O-glycans harbor blood group A, B or H antigens and are potential binding sites for Helicobacter pylori. *Glycobiology* **22**, 1193–1206
21. Jin, C., Kenny, D. T., Skoog, E. C., Padra, M., Adamczyk, B., Vitizeva, V., *et al*. (2017) Structural diversity of human gastric mucin glycans. *Mol. Cell. Proteomics* **16**, 743–758
22. Tanaka-Okamoto, M., Hanzawa, K., Mukai, M., Takahashi, H., Ohue, M., and Miyamoto, Y. (2018) Identification of internally sialylated carbohydrate tumor marker candidates, including Sda/CAD antigens, by focused glycomic analyses utilizing the substrate specificity of neuraminidase. *Glycobiology* **28**, 247–260
23. Bern, M., Kil, Y. J., and Becker, C. (2012) Byonic: advanced peptide and protein identification software. *Curr. Protoc. Bioinform.* **13**. https://doi.org/10.1002/0471250953.bi1320s40
24. Baker, P. R., Medzihradszky, K. F., and Chalkley, R. J. (2010) Improving software performance for peptide electron transfer dissociation data analysis by implementation of charge state- and sequence-dependent scoring. *Mol. Cell. Proteomics* **9**, 1795–1803
25. Lu, L., Riley, N. M., Shortreed, M. R., Bertozzi, C. R., and Smith, L. M. (2020) O-pair search with MetaMorpheus for O-glycopeptide characterization. *Nat. Met.* **17**, 1133–1138
26. Chalkley, R. J., Medzihradszky, K. F., Darula, Z., Pap, A., and Baker, P. R. (2020) The effectiveness of filtering glycopeptide peak list files for Y ions. *Mol. Omics* **16**, 147–155
27. Baker, P. R., Trinidad, J. C., and Chalkley, R. J. (2011) Modification site localization scoring integrated into a search engine. *Mol. Cell. Proteomics* **10**. https://doi.org/10.1074/mcp.M111.008078
28. Domon, B., and Costello, C. E. (1988) A systematic nomenclature for carbohydrate fragmentations in FAB-MS/MS spectra of glycoconjugates. *Glycoconjugate J.* **5**, 397–409
29. Park, G. W., Lee, J. W., Lee, H. K., Shin, J. H., Kim, J. Y., and Yoo, J. S. (2020) Classification of mucin-type O-glycopeptides using higher-energy collisional dissociation in mass spectrometry. *Anal. Chem.* **92**, 9772–9781
30. Chu, F. K. (1986) Requirements of cleavage of high mannose oligosaccharides in glycoproteins by peptide N-glycosidase F. *J. Biol. Chem.* **261**, 172–177
31. 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* **2012**, 560391
32. Medzihradszky, K. F., Kaasik, K., and Chalkley, R. J. (2015) Characterizing sialic acid variants at the glycopeptide level. *Anal. Chem.* **87**, 3064–3071
33. Pap, A., Klement, E., Hunyadi-Gulyas, E., Darula, Z., and Medzihradszky, K. F. (2018) Status report on the high-throughput characterization of complex intact O-glycopeptide mixtures. *J. Am. Soc. Mass Spectrom.* **29**, 1210–1220
34. Halim, A., Westerlind, U., Pett, C., Schorlemer, M., Rüetschi, U., Brinkmalm, G., *et al*. (2014) Assignment of saccharide identities through analysis of oxonium ion fragmentation profiles in LC-MS/MS of glycopeptides. *J. Proteome Res.* **13**, 6024–6032
35. Nilsson, J., Rüetschi, U., Halim, A., Hesse, C., Carlsohn, E., Brinkmalm, G., *et al*. (2009) Enrichment of glycopeptides for glycan structure and attachment site identification. *Nat. Met.* **6**, 809–811
36. Halim, A., Nilsson, J., Rüetschi, U., Hesse, C., and Larson, G. (2012) Human urinary glycoproteomics; attachment site specific analysis of N- and O-linked glycosylations by CID and ECD. *Mol. Cell. Proteomics* **11**. https://doi.org/10.1074/mcp.M111.013649
37. Halim, A., Rüetschi, U., Larson, G., and Nilsson, J. (2013) LC-MS/MS characterization of O-glycosylation sites and glycan structures of human cerebrospinal fluid glycoproteins. *J. Proteome Res.* **12**, 573–584
38. King, S. L., Joshi, H. J., Schjoldager, K. T., Halim, A., Madsen, T. D., Dziegiel, M. H., *et al*. (2017) Characterizing the O-glycosylation landscape of human plasma, platelets, and endothelial cells. *Blood Adv.* **1**, 429–442
39. Kawahara, R., Chernykh, A., Alagesan, K., Bern, M., Cao, W., Chalkley, R. J., *et al*. (2021) Community evaluation of glycoproteomics informatics solutions reveals high-performance search strategies for serum glycopeptide analysis. *Nat. Met.* **18**, 1304–1316
40. Riley, N. M., Malaker, S. A., and Bertozzi, C. R. (2020) Electron-based dissociation is needed for O-glycopeptides derived from OpeRATOR proteolysis. *Anal. Chem.* **92**, 14878–14884
41. Vainauskas, S., ne Guntz, H., McLeod, E., McClung, C., Ruse, C., Shi, X., *et al*. (2022) A broad-specificity O-glycoprotease that enables improved analysis of glycoproteins and glycopeptides containing intact complex O-glycans. *Anal. Chem.* **94**, 1060–1069
42. Conway, J. R., Lex, A., and Gehlenborg, N. (2017) UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics (Oxford, England)* **33**, 2938–2940