Precise, Fast and Comprehensive Analysis of Intact Glycopeptides and Monosaccharide-Modifications with pGlyco3

Wen-Feng Zeng[1,3,7,#,*]; Wei-Qian Cao[2,4,6,#]; Ming-Qi Liu[2,4,6]; Si-Min He[1,3]; Peng-Yuan Yang[2,4,5,6,*]

1. Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, China

2. The Fifth People's Hospital of Fudan University and Institutes of Biomedical Sciences, Fudan University, Shanghai, China

3. University of Chinese Academy of Sciences, Beijing, China

4. NHC Key Laboratory of Glycoconjugates Research, Fudan University, Shanghai, China

5. Department of Chemistry, Fudan University, Shanghai, China

6. The Shanghai Key Laboratory of Medical Epigenetics and the International Co-laboratory of Medical Epigenetics and Metabolism, Ministry of Science and Technology, Fudan University, Shanghai, China

7. Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany

# These authors contributed equally to this work

* Correspondence should be addressed to W.-F.Z. and P.-Y.Y.

Email: wzeng@biochem.mpg.de, pyyang@fudan.edu.cn

Abstract

We presented a glycan-first glycopeptide search engine, pGlyco3, to comprehensively analyze intact N- and O-glycopeptides, including glycopeptides with monosaccharide-modifications. We developed an algorithm, termed pGlycoSite, to localize the glycosylation sites and estimate the localization probabilities. We designed a number of experiments to validate the accuracy of pGlyco3 as well as other frequently used or recently developed software tools. These experiments showed that pGlyco3 outperformed the other tools on both N- and O-glycopeptide identification accuracy especially at the glycan level, without loss of the sensitivity. pGlyco3 also achieved a superior performance in terms of search speed. As pGlyco3 was shown to be accurate and flexible for glycopeptide search with monosaccharide-modifications, we then discovered a monosaccharide-modification of Hex (or an uncommon monosaccharide) "Hex+17.027 Da" on both O-mannose and N-glycopeptides in yeast samples, and confirmed this monosaccharide based on released N-glycans and isotopic labelling data. pGlyco3 is freely available on https://github.com/pFindStudio/pGlyco3/releases.

Introduction

Protein glycosylation is a fundamental post-translational modification (PTM) that is involved in many biological functions[1-3]. In recent years, tandem mass spectrometry (MS/MS) has been shown to be a promising technique to analyze site-specific glycans on proteins[4, 5]. Modern MS instruments have integrated different fragmentation techniques for glycopeptide analysis, such as higher-energy collisional dissociation (HCD), electron-transfer dissociation (ETD), electron-transfer/higher-energy collision dissociation (EThcD), and electron-transfer/collision-induced dissociation (ETciD)[6, 7]. HCD, especially the stepped collision energy HCD (sceHCD), could provide abundant glycopeptide Y ions (glycan Y ions with intact peptide attached) and quite a few b/y ions of naked peptides to identify the glycan parts and peptide parts, respectively[8]. The information of Y ions allows us to directly identify the entire glycan composition. But the b/y ions do not provide enough information to determine multiple glycosylation sites on a given sceHCD spectrum. ETxxD (ETD, EThcD, and ETciD) could generate glycan-attached c/z ions to not only identify peptides but also deduce the glycosylation sites[9-12].

Based on modern MS techniques, many glycopeptide search engines have been developed over the last decade. There are three major search strategies for intact glycopeptide identification: peptide-first, glycan-removal, and glycan-first searches. The peptide-first search is arguably the most widely used strategy which is adopted by Byonic[13], gpFinder[14], GPQuest[15], pMatchGlyco[16], GPSeeker[17], and two recently developed tools MSFragger (MSFragger-Glyco[18]) and MetaMorpheus (MetaMorpheus O-Pair[19]). This method first searches the peptide part and then deduces the glycan part as a large variable modification by considering some B/Y ions. MSFragger and MetaMorpheus use the peptide ion-indexing technique[20] to accelerate the peptide-first search. Glycan-removal search deduces the pseudo peptide masses from N-glycopeptide spectra by using potential reducing-end Y ions, and then it modifies the spectral precursor masses as the pseudo peptide masses to identify peptides using conventional peptide search engines[21-23]. The recently developed O-search further extended the glycan-removal strategy for O-glycopeptide identification[24]. These tools, especially Byonic, MSFragger and MetaMorpheus, have increased the identification sensitivity for glycoproteomics[25]; however, they do not pay much attention to control the glycan-level error rates. Glycan-first search is mainly used in pGlyco software series[8, 26], which first searches the glycan parts to remove unreliable glycans and then search the peptide parts. pGlyco 2.0 is the first search engine that can perform the glycan-, peptide- and glycopeptide-level quality control for glycopeptides. And it was extended for peptide identification and tandem mass tag (TMT)-quantification with MS3 by SugarQuant[27]. But pGlyco 2.0 only supports the search for normal N-glycans of mammals in GlycomeDB[28] with sceHCD spectra, and hence users can hardly apply it for analyzing customized glycans or monosaccharide-modifications (e.g., phospho-Hex or mannose-6-phosphate). Furthermore, modern glycopeptide search engines should consider site localization (SL) for site-specific O-glycopeptides, as ETxxD techniques have been widely used in glycoproteomics. Graph-based SL algorithms have been developed in recent years, such as GlycoMID for hydroxylysine O-glycosylation[29] and MetaMorpheus for common O-glycosylation[19], but accurate SL and its validation are still open problems.

Here, we proposed pGlyco3, a novel glycopeptide search engine that enables analysis of monosaccharide-modifications and SL. 1) pGlyco3 applies the glycan-first search strategy to accurately identify glycopeptides. It uses the canonicalization-based glycan database to support glycan-level modification analysis and a glycan ion-indexing technique to speed up the glycan search; 2) For glycosylation SL, we developed a dynamic programming algorithm termed pGlycoSite to efficiently localize the glycosylation sites using ETxxD spectra; 3) We emphasized validation for glycopeptide identification and SL. To validate the accuracy of pGlyco3, we designed several experiments to show that pGlyco3 outperformed the other tools in terms of identification accuracies for both N- and O-glycopeptides, especially at the glycan level. We also designed two methods to validate the SL of pGlycoSite. Compared with MetaMorpheus, we demonstrated that pGlycoSite does not overestimate the localization probabilities; 4) we used pGlyco3 to discover a monosaccharide-modification of Hex (or an uncommon monosaccharide), “Hex+17.027 Da” (simplified as “Hex+17” or “aH”) on N-glycopeptides and O-mannose (O-Man) glycopeptides in yeast samples. We indicated the existence of Hex+17 on yeast with N-glycome data and $^{15}N/^{13}C$-labeled glycopeptide data. This discovery further demonstrated the reliability and flexibility of pGlyco3 for intact glycopeptide and monosaccharide-modification identification.

Results

Workflow of pGlyco3

pGlyco3 uses sceHCD and ETxxD spectra to identify glycopeptides, analyze monosaccharide-modifications, estimate glycan/peptide false discovery rates (FDRs), and localize the glycosylation sites (Fig. 1a). The whole workflow is described in Online Methods. For glycan part identification, pGlyco3 provides several build-in N- and O-glycan databases, and it can also generate new glycan databases from GlycoWorkbench[30] with expert knowledge. Benefitting from the flexible canonicalization-based glycan representation, pGlyco3 enables convenient analysis of modified or labeled glycans of glycopeptides (Online Methods, Supplementary Note 1).

In contrast to Byonic, MetaMorpheus and MSFragger, pGlyco3 applies the glycan-first strategy which firstly searches the glycan parts and filters out unreliable glycans. For fast glycan-first search, we designed a glycan ion-indexing technique to score all glycans as well as their core Y ions by matching the query spectrum only once within the linear search time (i.e., O(#peaks), Supplementary Note 2). The ion-indexing of glycans in pGlyco3 indexes Y-complementary ions instead of Y ions, enabling glycan-first search before peptides are identified (Online Methods, Supplementary Note 2). The schema of the glycan-first search is shown in Fig. 1b. It applies a few verification steps to ensure the reliability of remained glycans, including the number of matched core Y ions, the existence of glyco-specific diagnostic ions, and the rank of glycan scores. After glycan filtration, pGlyco3 performs peptide search, glycan/peptide fine-scoring, post-search processing, and glycopeptide FDR estimation (Fig. 1a). If the ETxxD spectra are provided, pGlyco3 uses the pGlycoSite algorithm to localize the glycosylation sites.

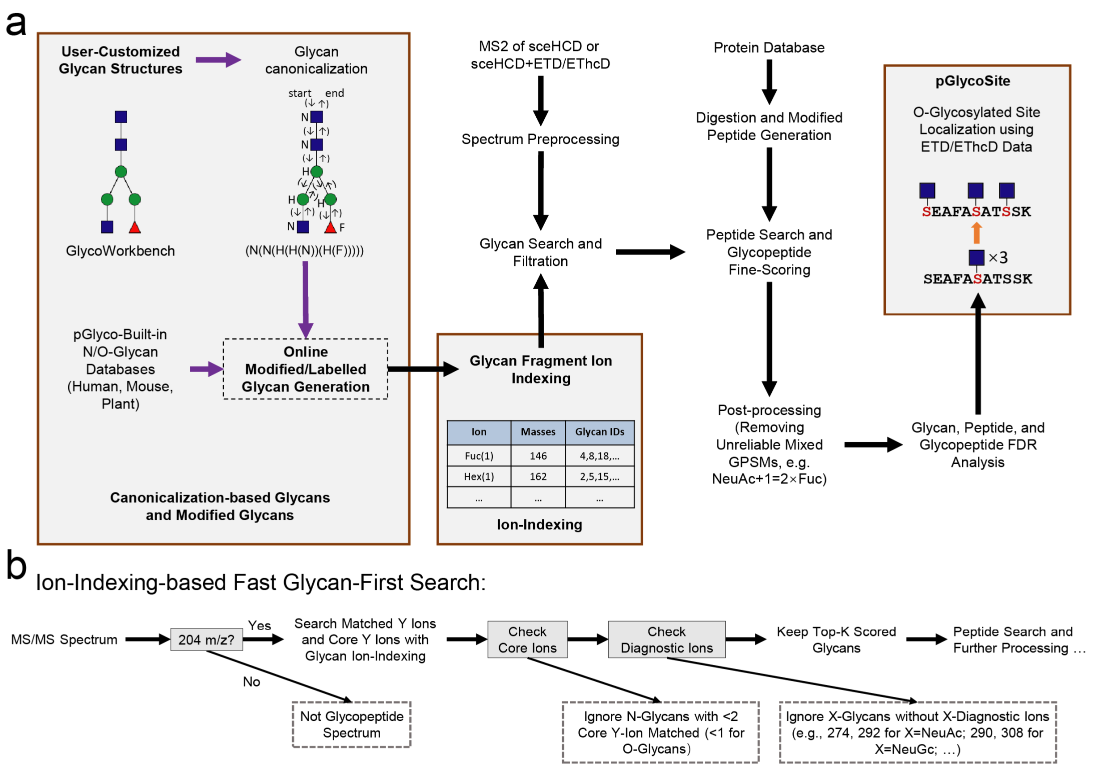

a
User-Customized Glycan Structures
Glycan canonicalization
start end
GlycoWorkbench
(N(N(H(H(N))(H(F)))))
pGlyco-Built-in N/O-Glycan Databases (Human, Mouse, Plant)
Online Modified/Labelled Glycan Generation
Canonicalization-based Glycans and Modified Glycans
MS2 of sceHCD or sceHCD+ETD/EThcD
Spectrum Preprocessing
Glycan Search and Filtration
Glycan Fragment Ion Indexing
Ion | Masses | Glycan IDs
Fuc(1) | 146 | 4,8,18,...
Hex(1) | 162 | 2,5,15,...
Ion-Indexing
Protein Database
Digestion and Modified Peptide Generation
Peptide Search and Glycopeptide Fine-Scoring
Post-processing (Removing Unreliable Mixed GPSMs, e.g. NeuAc+1=2×Fuc)
Glycan, Peptide, and Glycopeptide FDR Analysis
pGlycoSite
O-Glycosylated Site Localization using ETD/EThcD Data
SEAFASATSSK
×3
SEAFASATSSK
b
Ion-Indexing-based Fast Glycan-First Search:
MS/MS Spectrum
204 m/z?
Yes
No
Search Matched Y Ions and Core Y Ions with Glycan Ion-Indexing
Not Glycopeptide Spectrum
Check Core Ions
Check Diagnostic Ions
Keep Top-K Scored Glycans
Peptide Search and Further Processing ...
Ignore N-Glycans with <2 Core Y-Ion Matched (<1 for O-Glycans)
Ignore X-Glycans without X-Diagnostic Ions (e.g., 274, 292 for X=NeuAc; 290, 308 for X=NeuGc; ...)

Figure 1. Overview of pGlyco3. (a) Software schema (Online Methods). (b) Glycan-first search pipeline of pGlyco3. pGlyco3 only keeps reliable glycans for further processing by considering possible glycan-spectral information. It filters out unreliable glycans by checking the matched core Y ions and diagnostic ions.

Comparisons of pGlyco3 with other software tools for N-glycopeptide identification

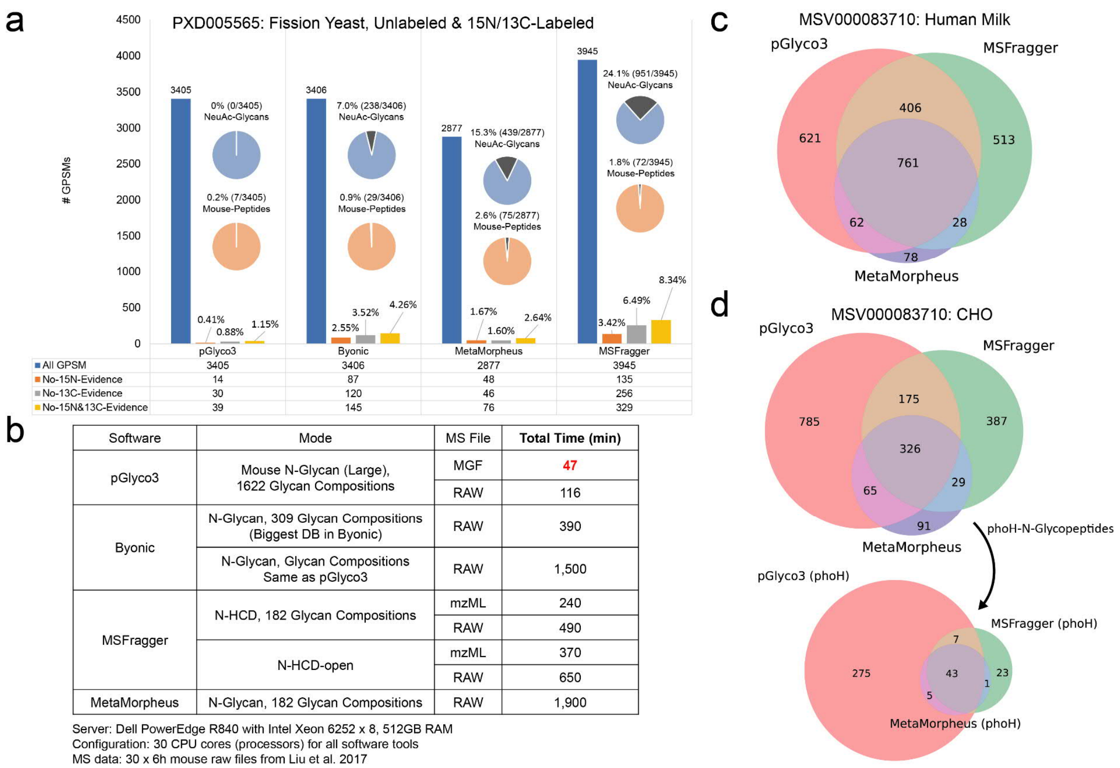
To demonstrate the performance of pGlyco3, we comprehensively compared pGlyco3 with Byonic, MetaMorpheus, and MSFragger on three N-glycopeptide datasets.

To compare the precision of identified glycopeptides, we used our previously published data of unlabeled, ^{15}N-labeled, and ^{13}C-labeled fission yeast mixture samples (PXD005565[8]) to test these software tools. The searched protein database was concatenated proteome database of fission yeast and mouse, and the searched N-glycan database contained NeuAc-glycans which should not be identified in yeast samples. The search details are listed in Online Methods and Supplementary Table 3. We analyzed three levels of identification errors: 1) The element-level error. The unlabeled glycopeptide-spectrum matches (GPSMs) may be identified with incorrect numbers of N or C elements if the GPSMs could not be verified by ^{15}N- or ^{13}C-labeled MS1 precursors; 2) The glycan-level error. The glycan parts tend to be incorrectly identified if the GPSMs contained NeuAc-glycans; 3) The peptide-level error. The peptide parts are false positives if they are from the mouse protein database. The testing results are shown in Fig. 2a. All tools showed low peptide-level error rates, implying that peptide- and glycan-first searches are both accurate at the peptide part identification for glycopeptides. However, the peptide-first-based tools showed high glycan-level error rates for yeast glycopeptides. MSFragger achieved the most identified GPSMs, but 24.1% of them contained NeuAc-glycans. The percentages of NeuAc-glycans for MetaMorpheus and Byonic were 7.0% and 15.3%, respectively. The real error rates of identified glycopeptides did not seem to be at the claimed 1% FDR. On the other hand, benefit from the essential glycan ion analyses and glycan FDR estimation, pGlyco3 showed good performance in controlling element-, glycan-, and peptide- error rates even with a larger glycan database (more interferences), without loss of the number of identified GPSMs. This comparison did not imply that the peptide-first strategy is not accurate. Instead, it suggested that glycan verification and glycan-level FDR estimation are also necessary for the peptide-first search to control glycan-level error rates.

Next, we compared the runtime of pGlyco3 with other software tools on large-scale mouse N-glycopeptide data from our previous work[8] (Supplementary Table 3). All these tools including pGlyco3, use multi-processors to accelerate the searches. We used 30 processes for all the tools to search the data on a Dell workstation with 64 CPU cores and 512 GB physical memories. The time comparisons are shown in Fig. 2b. Ignoring the RAW file parsing time (searching from MGF files), pGlyco3 took only 47 minutes to finish the search (~1.6 minutes per file, 1622 glycan compositions in the database). Whereas the second-fastest tool, MSFragger, took 240 minutes (~8 minutes per file) with a ~7 times smaller glycan database (182 glycan compositions). The running time of pGlyco3 starting from RAW files was ~3.9 minutes per file, which was also faster than other tools. The runtime comparisons provided strong evidence showing that the glycan-first search with glycan ion-indexing is very fast for glycopeptide identification.

Finally, we compared pGlyco3 with MSFragger and MetaMorpheus on sceHCD-product-dependent (pd)-EThcD N-glycopeptide datasets from MSV00083710[7]. MSV000083710 contains two datasets, human milk and phospho-enriched Chinese hamster ovary cell (CHO). The comparison results of MSV000083710 are shown in Fig. 2c and Fig. 2d (search parameters are listed in Supplementary Table 3). pGlyco3 showed good identification on unique N-glycopeptides with sceHCD-pd-EThcD data (Fig. 2c). Furthermore, as pGlyco3 is specifically designed for the modified glycan search, it could identify more phospho-Hex (simplified as phoH in this manuscript) N-glycopeptides than the other two tools (Fig. 2d). It is worth mentioning that all identified phoH-contained N-glycopeptides by pGlyco3 were supported by the phoH-diagnostic ion (242.019 m/z), hence they would be quite reliable.

Overall, based-on the glycan-first search and glycan-level quality control, pGlyco3 outperformed the other three software tools in terms of accuracy, search speed, and monosaccharide-modification identification.



	pGlyco3	Byonic	MetaMorpheus	MSFragger
■ All GPSM	3405	3406	2877	3945
■ No-15N-Evidence	14	87	48	135
■ No-13C-Evidence	30	120	46	256
■ No-15N&13C-Evidence	39	145	76	329

Software	Mode	MS File	**Total Time (min)**
pGlyco3	Mouse N-Glycan (Large), 1622 Glycan Compositions	MGF	47
		RAW	116
Byonic	N-Glycan, 309 Glycan Compositions (Biggest DB in Byonic)	RAW	390
	N-Glycan, Glycan Compositions Same as pGlyco3	RAW	1,500
MSFragger	N-HCD, 182 Glycan Compositions	mzML	240
		RAW	490
	N-HCD-open	mzML	370
		RAW	650
MetaMorpheus	N-Glycan, 182 Glycan Compositions	RAW	1,900

Server: Dell PowerEdge R840 with Intel Xeon 6252 x 8, 512GB RAM
Configuration: 30 CPU cores (processors) for all software tools
MS data: 30 x 6h mouse raw files from Liu et al. 2017

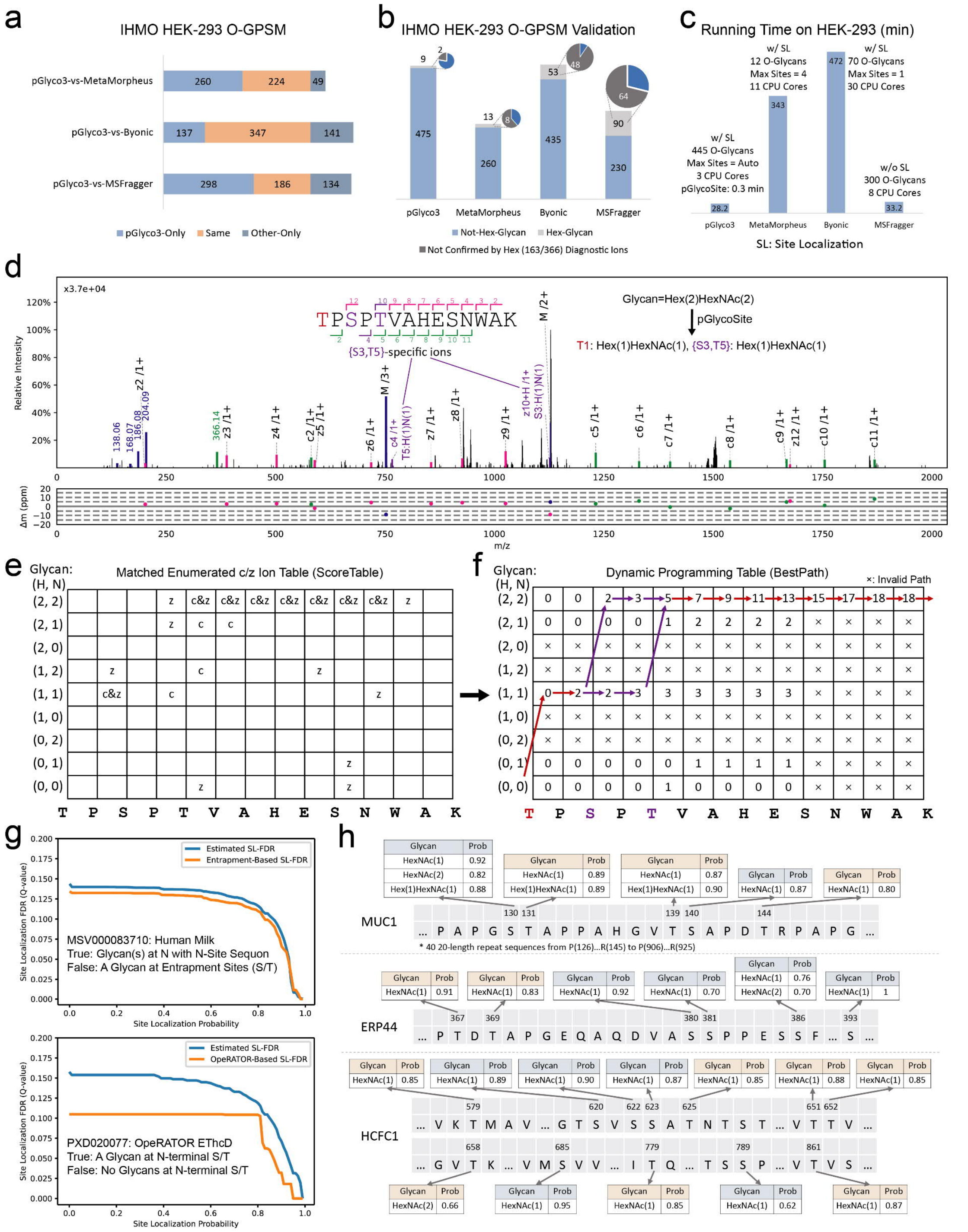
Figure 2. Comparison of pGlyco3 with three other software tools on N-glycopeptide data. (a) Accuracy comparison using ^{15}N/^{13}C-labeled fission yeast data (PXD005565). The element-level error rate (incorrect number of N or C elements) of identified glycopeptides are tested by the ^{15}N/^{13}C-labeled precursor signals. The potential glycan-level error rate is tested by the percentage of NeuAc-containing GPSMs, and the potential peptide-level error rate is tested via GPSMs with mouse peptides. pGlyco3 shows the best accuracies at all three levels. (b) Search speed comparison under N-glycopeptide data of five mouse tissues (Liu et al. 2017, PXD005411, PXD005412, PXD005413, PXD005553, and PXD005555, 6h gradient for each RAW file, 30 RAW files in total). The algorithm part (searching from MGF files) of pGlyco3 is at least 5 times faster than other tools, even when using larger glycan databases. (c-d) Comparison of identified unique N-glycopeptides with sceHCD-pd-EThcD data for (c) human milk and (d) phospho-enriched CHO data. phoH is short for phospho-Hex. All identified phoH-GPSMs by pGlyco3 are supported by the phoH-diagnostic ion (242.019 m/z).

pGlyco3 for O-glycopeptide identification and pGlycoSite for glycosylation site localization

As the glycan-first search of pGlyco3 has been shown to be accurate and fast for N-glycopeptide identification, we then demonstrated the performance of pGlyco3 for O-glycopeptide identification. The workflow of O-glycopeptide is similar to that of N-glycopeptide, except that S/T are the candidate glycosylation sites and the core Y ions are changed (Supplementary Table 2). We compared O-glycopeptide identification with pGlyco3 and other tools on inhibitor-initiated homogenous mucin-type O-glycosylation (IHMO) cell line datasets (Fig. 3a-c, Supplementary Note 4). IHMO cells were almost inhibited to be with only truncated HexNAc(1) or HexNAc(1)NeuAc(1) O-glycans on the peptides (Supplementary Note 4). The identified O-GPSMs on the IHMO HEK-293 dataset (Fig. 3a) were then evaluated by the Hex-containing GPSMs, which were further confirmed by checking the Hex-diagnostic ions (163.060 and 366.139 m/z, Fig. 3b, Online Methods). pGlyco3 obtained only 1.9% (9 out of 484) Hex-GPSMs, and only 2 out of 9 Hex-GPSMs could not be validated by the Hex-diagnostic ions, showing a very low Hex-suggested glycan-level error rate (2/484≈0.4%) for the IHMO HEK-293 data. The Hex-suggested glycan-level error rates of the three other software tools on the same dataset were: ~2.9% (8/273) for MetaMorpheus, ~10.0% (48/488) for Byonic, and ~20.0% (64/320) for MSFragger. These results further proved the accuracy of pGlyco3 in glycopeptide identification. pGlyco3 also showed a higher search speed than others on the same IHMO dataset.

pGlyco3 integrated a novel algorithm termed pGlycoSite to localize the glycosylation sites and estimate the localization probabilities using c/z ions in ETxxD spectra. Fig. 3d-f showed an example of how pGlycoSite works. For the given ETxxD spectrum, the glycan composition and peptide could be identified as "TPSPTVAHESNWAK + Hex(2)HexNAc(2)", and their enumerated c/z ions were then generated and matched against the spectrum. All matched c/z ions were recorded in a table ($ScoreTable$, Fig. 3e). A dynamic programming algorithm was designed to find the best-scored path from the bottom left to top right ($BestPath$, Fig. 3f) based on the $ScoreTable$. If multiple paths reached the same best score, the amino acids from the branching position to the merging position were regarded as a "site-group". In Fig. 3f, T1 with Hex(1)HexNAc(1) was uniquely localized, and {S3:T5} was regarded as a "site-group" because either S3 or T5 had a supported site-specific c/z ion (Fig. 3d), resulting in the same score. The details of the pGlycoSite algorithm, including the calculation of $ScoreTable$, $BestPath$ and localization probability estimation were illustrated in Online Methods and Supplementary Note 3. After the SL probabilities were estimated, the SL-FDR could be deduced and used to validate the accuracies of estimated SL probabilities (Online Methods). We designed two kinds of experiments to validate the estimated SL-FDR: entrapment-based and OpeRATOR-based method. The entrapment-based method applies pGlycoSite on N-GPSMs by regarding J/S/T (J is N with the N-glycosylation sequon) as the candidate sites, and the SL would be false in an N-GPSM if a site is not localized at J. The OpeRATOR-based method suggests that the SL would be false for a GPSM if no sites are localized at the N-terminal S/T since the OpeRATOR recognizes O-glycans and cleaves O-glycopeptides at the N-termini of O-glycan-occupied S or T[31]. The entrapment-based SL-FDR and OpeRATOR-based SL-FDR could be deduced (Online Methods). SL-FDRs estimated by pGlycoSite were validated by entrapment-based SL-FDR and OpeRATOR-based SL-FDR, and it suggested that the SL-FDR was conservatively estimated by pGlycoSite, as shown in Fig. 3g and Supplementary Fig. 2. Maybe neither of these two validation methods is perfect, but the combination of them can verify the accuracy of pGlycoSite to a certain extent.

We then performed IHMO on multiple cell lines, including HEK-293, Jurkat T, B3, MCF, and HeLa. The data were identified by pGlyco3 and sites were localized by pGlycoSite. The localized sites with probabilities for proteins Mucin 1 (MUC1), endoplasmic reticulum protein 44 (ERP44), and host cell factor C1 (HCFC1) are listed in Fig. 3h. All these localized O-glycosylation sites could be confirmed on www.uniprot.org, www.oglyp.org[32], or https://www.oglcnac.mcw.edu[33], showing the reliability of these localized sites.



e

Glycan: (H, N)													
(2, 2)				z	c&z	c&z	c&z	c&z	c&z	c&z	c&z	z	
(2, 1)				z	c	c							
(2, 0)													
(1, 2)		z			c				z				
(1, 1)		c&z		c							z		
(1, 0)													
(0, 2)													
(0, 1)										z			
(0, 0)					z					z			

T P S P T V A H E S N W A K

f

Glycan: (H, N)													
(2, 2)	0	0	2	3	5	7	9	11	13	15	17	18	18
(2, 1)	0	0	0	0	1	2	2	2	2	×	×	×	×
(2, 0)	×	×	×	×	×	×	×	×	×	×	×	×	×
(1, 2)	×	×	×	×	×	×	×	×	×	×	×	×	×
(1, 1)	0	2	2	3	3	3	3	3	3	×	×	×	×
(1, 0)	×	×	×	×	×	×	×	×	×	×	×	×	×
(0, 2)	×	×	×	×	×	×	×	×	×	×	×	×	×
(0, 1)	0	0	0	0	0	1	1	1	1	×	×	×	×
(0, 0)	0	0	0	0	1	0	0	0	0	×	×	×	×

T P S P T V A H E S N W A K

Figure 3. O-glycopeptide identification and SL of pGlyco3. (a-c) Software comparison of O-glycopeptide search on IHMO HEK-293 cell line data. (b) The Identified IHMO O-GPSMs are validated by Hex-containing results and further validated by Hex-diagnostic ions. (c) The runtime is also compared. Byonic could not finish searching for “Max Sites=2” after weeks. (d) An example of an EThcD-GPSM (“TPSPTVAHESNWAK + H(2)N(2)”, H=Hex, N=HexNAc) with localized sites using the pGlycoSite algorithm. (e-f) Illustration of the pGlycoSite algorithm for the GPSM in (d). (e) show all possible matched c/z ions against the EThcD spectrum ($ScoreTable$). (f) shows the dynamic programming table from the bottom left to top right ($BestPath$). The arrow lines indicate the best-scored paths, and the purple arrow lines show that two paths are sharing the same score from S3 to T5. T1 is localized with Hex(1)HexNAc(1), and {S3:T5} is localized as a “site-group” with Hex(1)HexNAc(1), as shown in (d). The details of the calculation of

$ScoreTable$ and $BestPath$ are illustrated in Online Methods and Supplementary Note 3. (g) Validation of the SL-FDR of pGlycoSite using entrapment-based SL-FDR and OpeRATOR-based SL-FDR (Online Methods). (h) Localized sites and their O-glycans of MUC1, ERP44, and HCFC1 proteins using all the IHMO cell line data.

Analyses of "Hex+17" in yeast samples

In the previous sections, pGlyco3 was proved to be reliable for glycopeptide identification, we then applied pGlyco3 to analyze "Hex+17"-glycopeptides in yeast samples. In this manuscript, "Hex+17" refers to a monosaccharide-modification of Hex (or an uncommon monosaccharide) with Hex plus 17.027 Da (Supplementary Note 8), and it is abbreviated as "aH".

We first searched N-glycopeptides and O-Man glycopeptides on the $^{15}N/^{13}C$ labeled fission yeast dataset (PXD005565[8]) with Hex "modified" by aH (Online Methods). Then, the identified aH-glycopeptides were confirmed at the element level by ^{15}N- and ^{13}C-labeled precursors. As shown in Fig. 4a, we found 579 unique aH-N-glycopeptides with one aH and 164 aH-N-glycopeptides with two aH. In total, 571 out of 579 and 155 out of 164 aH-glycopeptides could be confirmed by the ^{15}N&^{13}C MS1 evidence for aH×1 and aH×2, respectively. To prove that aH occurred on glycans instead of peptides (as a 17 Da modification on the peptides), we released N-glycans from fission yeast using peptide:N-glycosidase F (PNGase F) and analyzed the N-glycome with MS/MS (Online Methods). Among the 50 identified aH-N-glycans (27 with aH×1 and 23 with aH×2) from N-glycopeptides, 86% (22 + 21 out of 50) aH-N-glycans could be confirmed by yeast N-glycome MS/MS data, as shown in Fig. 4a. aH was also found in O-Man glycopeptides in PXD005565 and confirmed by ^{15}N&^{13}C MS1 evidence (Fig. 4a, right).

Fig. 4b show the MS/MS, ^{15}N&^{13}C MS1, and N-glycome evidence of the aH-N-glycopeptide "VQASJ(N)WTGTR + Hex(9)HexNAc(1)aH(1)". This glycopeptide was identified at MS/MS scan 17948 and was validated by its unlabeled, ^{15}N-labeled, and ^{13}C-labeled MS1 precursors at scan 17943. MS/MS spectral annotation showed that the aH-glycan was confidently identified with many continuous Y ions matched (Fig. 4b, top). MS1 precursor evidence showed high Pearson correlation coefficients (R) and low matching mass errors for all the unlabeled, ^{15}N-labeled and ^{13}C labeled precursors (Fig. 4b middle). And the N-glycan "Hex(9)HexNAc(1)aH(1)" was eventually proven by its N-glycome MS/MS spectrum, of which almost all peaks were interpreted by the fragments of this N-glycan (Fig. 4b, bottom). In this N-glycome spectrum, we could also find three aH-supporting ions, although these ions had quite low intensities.

Fig. 4a and 4b strongly suggested the existence of the aH monosaccharide in fission yeast samples, showing that pGlyco3 can confidently identify the aH-glycopeptides. We then generated large-scale fission yeast and budding yeast datasets to further analyze the aH-glycopeptides (Supplementary Note 6, the search parameters were shown in Supplementary Table 3), the results are shown in Fig. 4c. We found that there were large proportions of aH-glycopeptides in both fission yeast and budding yeast samples. In total, 55.8% (1059 out of 1898) of aH-N-glycopeptides and 29.4% (64 out of 218) of aH-O-Man glycopeptides were identified in fission yeast. And the percentages were 58.0% and 40% for budding yeast. pGlyco3 also identified several phoH-N-glycopeptides and phoH-O-Man glycopeptides on the budding yeast dataset. We analyzed site-specific N-glycans and O-Man glycans of protein O-mannosyltransferase (*ogm1*) in fission yeast, and the O-Man sites were localized by pGlycoSite on EThcD data, as displayed in Supplementary Figure 5.

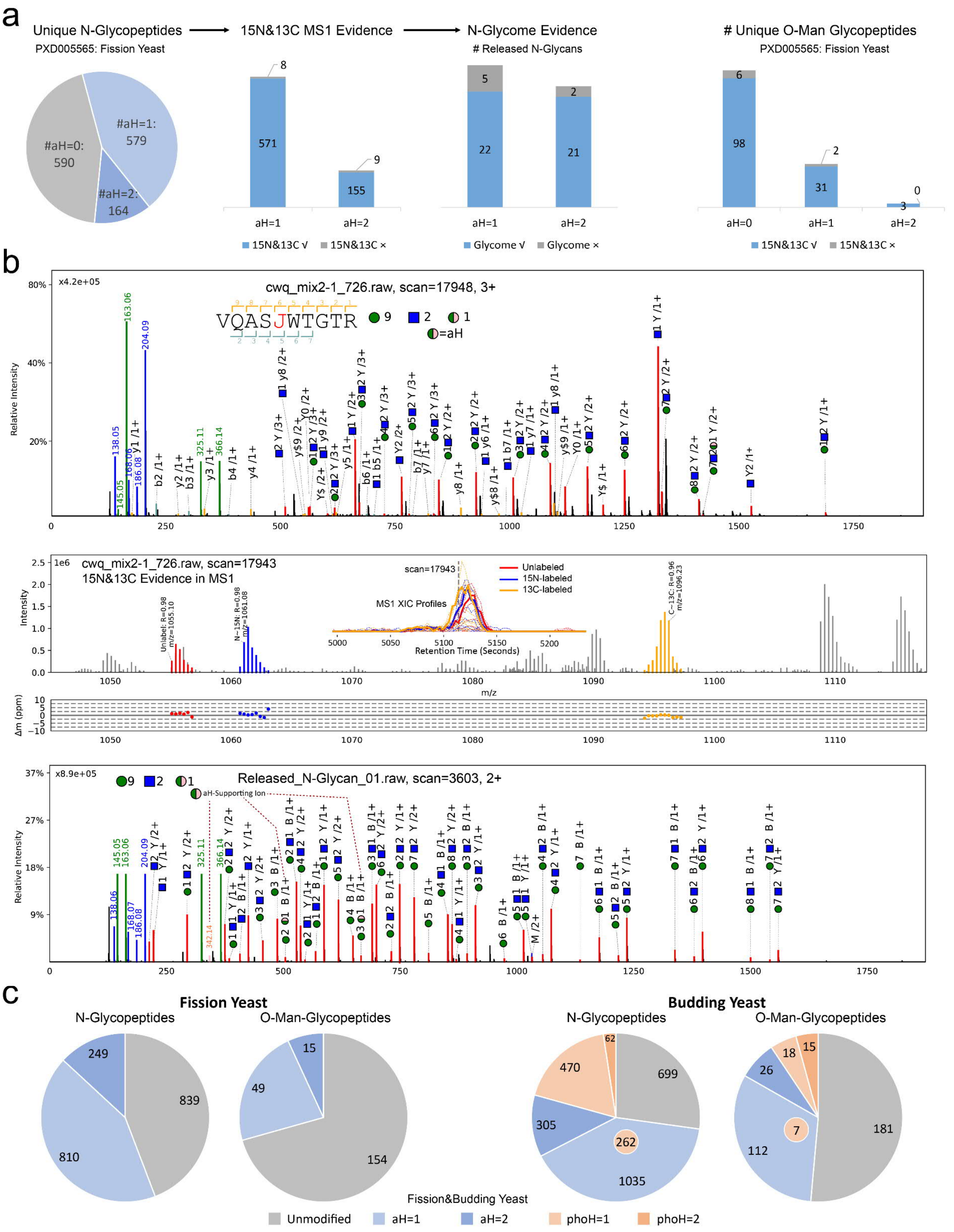


Figure 4. Analysis and verification of the “Hex+17 (aH)”. (a) Unlabeled aH-N-glycopeptides are identified on fission yeast data (PXD005565) and are verified by the ^{15}N/^{13}C-labeled MS1 precursors and the N-glycome MS/MS data. Unlabeled aH-O-Man glycopeptides are also identified and verified by ^{15}N/^{13}C-labeled MS1 evidence. (b) MS/MS

spectral annotation (top), ^{15}N&^{13}C MS1 evidence (middle), and N-glycome MS/MS evidence (bottom) of the aH-N-glycopeptide "VQASJ(N)WTGTR + Hex(9)HexNAc(2)aH(1)". In the MS1 annotation, "R" is the Pearson correlation coefficient between the theoretical and experimental isotope distributions. In the N-glycome spectrum annotation, the B ions are calculated as the sum of the neutral masses of monosaccharides, and the Y ions are calculated as the sum of the neutral masses of monosaccharides plus a water group (18.0105646863 Da), all possible B/Y ions are generated for this glycan composition. The blue square and green circle refer to HexNAc and Hex, respectively. (c) Deep analysis of aH-N-glycopeptides and aH-O-Man glycopeptides in fission yeast and budding yeast data. Phospho-Hex (phoH) glycopeptides are also found in budding yeast samples. 262 N-glycopeptides and 7 O-Man glycopeptides in budding yeast contain both aH and phoH.

Discussion

In this work, we emphasized the importance of glycan-level quality control for the development of glycopeptide search engines to ensure the glycan-level accuracy. A glycan is not just a simple PTM, its mass is sometimes so large that traditional PTM searches may obtain different glycan compositions or even different glycan and amino acid combinations[21]. Fortunately, glycans have their own fragment ions and diagnostic ions, allowing search engines to achieve more accurate glycan identification. Strict glycan-level quality control may reduce the sensitivity, but accuracy should be always the first priority for any search engine. Therefore, for peptide-first search engines, we also recommend performing the glycan-level quality control after peptides are identified, as they have obtained superior peptide identification performance.

SL is important, especially for O-glycosylation. MetaMorpheus uses a graph-based algorithm to localize the sites and estimate the site probabilities, but it needs to build different graphs for different combinations of glycans. pGlycoSite provides a one-step algorithm to deduce the sites based on the $ScoreTable$ with a dynamic programming algorithm, making the SL step extremely fast. As heuristic algorithms for glycosylation SL have been developed only in the last few years, there is still plenty of room to improve the scoring schema, SL probability estimation, and result validation.

pGlyco3 provides a convenient way to analyze the glycan-level modifications based on canonicalization-based glycan representations, making the analysis of glycan-level modifications similar to that of peptide-level modifications for users. This new feature enables us to discover the "aH" monosaccharide, which could be confirmed by the N-glycome data and ^{15}N/^{13}C-labeled data, on N-glycopeptides and O-Man glycopeptides of yeast samples. As our results showed that aH-glycopeptides occur quite often in yeast data, we believe aH should have some functions in yeast cells. Because mass spectrometry could only help to deduce the mass of aH, further investigation of the exact chemical composition and linkage information of aH is needed. We also do not know if aH is a monosaccharide-modification of Hex or a new monosaccharide, but pGlyco3 could perform the search once the aH mass is known, and the partial chemical composition could be manually speculated using ^{15}N/^{13}C labeled data.

Online Methods

Spectrum pre-processing. pGlyco3 can process HCD and "HCD+ETxxD" (HCD-pd-ETxxD, or HCD followed by ETxxD) data for N-/O-glycopeptide identification, here ETxxD could be either ETD, EThcD, or ETciD. Note that pGlyco3 is optimized for both glycan and peptide fragment analysis using stepped collision energy HCD (sceHCD), hence sceHCD is always recommended. If HCD and ETxxD spectra are generated for the same precursor, pGlyco3 will automatically merge them into a single spectrum for searching. pGlyco3 then deisotopes and deconvolutes all MS2 spectra, and removes the precursor ions. pGlyco3 uses pParse to determine the precursor mono ions and to export chimera spectra, which has been also used in peptide and cross-linked peptide identification[34, 35]. pGlyco3 filters out non-glycopeptide spectra by checking glycopeptide-diagnostic ions, then searches the glycan parts with the glycan ion-indexing technique. The diagnostic ions can be defined by users, the default ion is 204.087 m/z for both N-glycosylation and O-glycosylation.

Glycan-first search and glycan ion-indexing. In pGlyco3, each glycan structure is represented by a canonical string in the glycan database. pGlyco3 provides quite a few build-in N- and O-glycan databases, it also supports to convert glycan structures of GlycoWorkbench[30] into canonical strings of pGlyco3. For monosaccharide-modifications, pGlyco3 will automatically substitute one or several unmodified monosaccharides in each of canonical strings into modified forms, making it very convenient for monosaccharide-modification analysis (Supplementary Note 1).

For each peak in an MS2 spectrum, pGlyco3 assumes it is a Y ion of a glycopeptide with the peptide attached to search against the databases. But we cannot calculate the theoretical Y ions of glycopeptides if the peptide parts are unknown in the glycan-first search. Therefore, pGlyco3 searches the Y-complementary ions ("precursor mass – Y ion mass") instead of Y ions. To speed up the Y-complementary ion search, pGlyco3 builds the ion-indexing table for all Y-complementary ions of all possible Y ions for the glycan database (Supplementary Note 2). The Y-complementary ion composition is defined as "full glycan composition – Y-ion glycan composition". A table of all possible unique Y-complementary ion compositions is generated, and the list of glycan IDs where the Y-complementary compositions are originated from are also recorded in the table. For glycan-first search, core Y ions (e.g., trimannosyl core Y ions in N-glycans, Supplementary Table 2) are the key ions for N-glycan scoring. Hence pGlyco3 encodes the glycan ID using 31 bits of the 32-bit integer and uses the extra bit to record whether the corresponding Y ion is the core ion of the glycan or not (Supplementary Note 2). The table is then sorted by the masses, and hashed by the masses for fast query (within O(1) query time). As a result, it only takes O(#Peak) time to get the matched ion counting scores as well as the matched core ion counting scores of all glycans for every spectrum, here #Peak refers to the number of peaks in a spectrum. pGlyco3 keeps the glycans with ≥ n core Y ions matched (n = 2 for N-glycan and n = 1 for O-glycan). Glycans are further filtered out if they contain the specific monosaccharide but are not supported by corresponding monosaccharide-diagnostic ions (Supplementary Table 1). At last, pGlyco3 keeps the 100 top-scored (sum of the ion counting score and the core ion counting score) candidate glycan compositions for further peptide search. pGlyco3 also keeps all small glycans (number of monosaccharides ≤ 3) for peptide search because small glycans have too few Y ions to obtain high glycan scores.

Peptide search and glycopeptide false discovery rate estimation. In protein sequence processing, every Asn (N) with the sequon "N-X-S/T/C (X is not P)" in all protein sequences is converted to "J" while keeping the same mass and chemical elements as N. J is then the candidate N-glycosylation site, and S/T are the candidate O-glycosylation sites. Proteins are digested into peptide sequences, then modifications are added into peptide sequences. Modified peptides are indexed by their masses for O(1) time access as well. For the given spectrum and each of the candidate glycans, the peptide mass is deduced by "precursor mass – glycan mass". pGlyco3 then queries the peptide mass from the mass-indexed peptides. For peptide search, pGlyco3 considers b/y ions for the HCD mode, it also considers b/y + HexNAc ions for N-glycopeptides. pGlyco3 further considers c/z ions as well as their hydrogen-rearranged for the HCD+ETxxD mode. The candidate glycan is fine-scored by the matched Y ions as well. The glycan and peptide scoring schemes of pGlyco3 are the same as pGlyco 2.0, but some parameters were tuned in pGlyco3 to obtain better identification performance (Supplementary Figure 6). Only top-ranked glycopeptide is kept as the final result for each spectrum. For potential chimera spectra, pGlyco3 removes unreliable mixed glycopeptides by checking if one's precursor is others' isotopes. For example, if NeuAc(1) and Fuc(2) are simultaneously identified in the same MS2 scan but with different precursors, the Fuc(2)-glycopeptide will be removed because "NeuAc(1) + 1 Da = Fuc(2)". pGlyco3 also uses pGlyco 2.0's method to estimate the false discovery rates (FDR) for all glycopeptide-spectrum-matches (GPSMs) at the glycan, peptide, and glycopeptide level. pGlyco3 skips the glycan FDR estimation step for small glycans.

Fast glycosylation site localization with pGlycoSite. Glycosylation site localization in pGlyco3 is not only to determine the glycosylation sites but also to determine the attached glycan composition to each site. For a given spectrum with the identified peptide and glycan composition, enumerating all possible glycopeptide-forms and generating their c/z ions for site localization is not computationally easy. The worst computation complexity could be $O(L \times \prod_{i=1}^{T} C_{S+G_i-1}^{G_i-1})$, where L is the peptide length, T is the number of monosaccharide types, S is the number of candidate glycosylation sites, and G_i is the number of the ith monosaccharide type (Supplementary Note 3). The enumeration complexity would be exponentially large as S and G_i increase, as illustrated in Supplementary Note 3.

In pGlyco3, the pGlycoSite algorithm is designed to avoid the enumeration. The key observation for pGlycoSite algorithm is that, no matter how many glycopeptide-forms there are for a given peptide and glycan composition, the number of all possible c or z ions is at most $F \times (L-1)$. Here, F is the number of sub-glycan compositions for the identified glycan composition, and the sub-glycan is defined in Supplementary Note 3. pGlyco3 generates a c/z-ion table of which each cell contains the glycan-attached-c/z ions (Supplementary Note 3). After removing all Y and b/y ions from the given spectrum, the ion table with c/z ions is then matched and scored against the spectrum (called $ScoreTable$ table, as illustrated in Fig. 3b and Supplementary Note 3).

pGlycoSite currently uses c/z ion-counting scores for each cell of the table, but other comprehensive scoring schemes could be supported in the table if they could achieve better performance.

The best-scored path starting from bottom left $[g_0, 0]$ to top right $[G, L]$ (Fig. 3c and Supplementary Note 3) are then calculated by a dynamic programming algorithm:

$$BestPath[g,p] = \begin{cases} \max\limits_{\forall g_s \le g} BestPath[g_s, p-1] + ScoreTable\,[g,p] & if\ IsValidPath(g_s, g, p) \\ \times & if\ not\ \exists g_s \le g\ \ IsValidPath(g_s, g, p) \end{cases}.$$

where G is the identified full glycan composition, g refers to a sub-glycan composition of G, g_0 refers to the zero-glycan composition, and p refers to pth position of the peptide sequence. Here, all glycan compositions (from g_0 to G) are represented as vectors, hence could be compared with each other. So $g_s \le g$ means that g_s is the sub-glycan of g. $IsValidPath(g_s, g, p)$ is designed to check whether the path starting from $[g_s, p-1]$ to $[g, p]$ is valid path or not (Supplementary Note 3). pGlycoSite sets $BestPath[g, 0] = 0$ $(\forall g\colon g_0 \le g \le G)$, and iteratively calculates the $BestPath$ table for all $g_0 \le g \le G$ and $0 < p \le L$. $BestPath[G, L]$ is then the final best path score which we are going to solve. Finally, pGlycoSite deduces all the paths that can reach $BestPath[G, L]$ score by backtracking the $BestPath$ table from $[g_0, 0]$ to $[G, L]$. If the best-scored path contains the cell $[g_s, p-1]$ and $[g, p]$ with $g_s < g$, then the pth amino acid is localized as a site with glycan $g - g_s$. pGlycoSite introduces the “site-group” if there are multiple paths can achieve the same $BestPath[G, L]$ score (Fig. 3c, Supplementary Note 3). The time complexity of site localization of pGlycoSite including the dynamic programming and the backtracking for a GPSM is only $O(L \times F^2)$ (Supplementary Note 3).

Site localization probability estimation of pGlycoSite. Glycosylation site probability refers to the probability that a site is correctly localized. As the peptide and glycan composition have been identified for a given MS2 spectrum, the incorrect localization would come from the random assignment of randomly selected sub-glycans to random sites for the same peptide and glycan composition. To simulate the incorrect localization for each localized site, pGlycoSite randomly samples 1000 paths from bottom left to top right on the $ScoreTable$. For a given site or site-group to be estimated, the random paths could have overlap with the $BestPath$, except that they must not contain the path that can determine this site or site-group (i.e., path from $[g_s, p_i]$ to $[g, p_j]$ for site p_i $(j = i + 1)$ or site-group $\{p_i, p_{j-1}\}$ $(j > i + 1)$). pGlycoSite then calculates 1000 ion counting scores of these paths, and estimates a Poisson distribution from these random scores. It estimates the p-values based on the Poisson distribution for the $BestPath[G, L]$ and the best random score (denoted as $RandomBest$), then estimates the probability as

$$Prob_{poisson} = \frac{\log\left(pvalue(BestPath[G,L])\right)}{\log(pvalue(BestPath[G,L])) + \log\left(pvalue(RandomBest)\right)}.$$

To ensure the localized glycopeptide-spectrum-matching quality, pGlycoSite adds a regularization factor to the estimated $Prob_{poisson}$, and the final localized probability becomes

$$Prob = Prob_{poisson} * r = Prob_{poisson} * \left(\frac{BestPath[G,L]}{2(L-1)}\right)^{\alpha},$$

where α is set as a small value (0.05) to make it not too much affect the value of $Prob_{poisson}$. But when $BestPath[G, L]$ gets a very small score, r will be closed to zero, hence limiting the final $Prob$ value. $L-1$ is the number of the considered c/z ions.

Benchmark, software versions. pGlyco3 is compared with MetaMorpheus (v0.0.312, downloaded in 2020.10), MSFragger (v3.1.1 with FragPipe v14.0 and philosopher v3.3.11, downloaded in 2020.10), and Byonic (v3.10).

Previously published mass spectrometry data. sceHCD raw files of mixed unlabeled, ^{15}N-labeled, and ^{13}C-labeled fission yeast glycopeptide samples were downloaded from PXD005565[8] on PRIDE. 30×6 h sceHCD raw files of five mouse tissues were also downloaded from PXD005411, PXD005412, PXD005413, PXD005553, and PXD005555[8] on PRIDE. sceHCD-pd-EThcD raw files of human milk and Chinese hamster ovary cell (CHO) samples were obtained from MassIVE (dataset MSV000083710[7]). Raw files of OpeRATOR-processed O-glycopeptide data were obtained from PXD020077[9] on PRIDE. Raw files of StcE-processed O-glycopeptide data were obtained from PXD017646[12] on PRIDE. Detailed searching parameters for all these RAW data were listed in Supplementary Table 3.

Validation of N-glycopeptide search with ^{15}N/^{13}C-labeled fission yeast data. The protein sequence database was the fission yeast protein sequence database (*S. pombe*, Swissprot, 2018.08) concatenated with the mouse protein sequence database (*M. musculus*, Swissprot, 2018.08). Identified GPSMs with mouse peptides would be false identification, and hence mouse-peptide GPSMs could be used to test the peptide-level error rates. The N-glycan database for MetaMorpheus, MSFragger, and Byonic is the 182-glycan database which includes 74 NeuAc-contained N-glycan compositions. The N-glycan database for pGlyco3 is the build-in Mouse N-glycan database, which contains 1234 N-glycan compositions (6662 structures) and has 659 NeuAc-contained compositions. NeuAc-contained N-glycan compositions identified in fission yeast data would be false identifications, which could be used to test the glycan-level error rates. The detailed searching parameters were listed in Supplementary Table 3. For each software tool, all spectra were regarded as the unlabeled spectra while searching, and the identified GPSMs were then validated by using their ^{15}N/^{13}C-labeled precursor signals in the MS1 spectra (Fig. 2a). This validation method was also used in our previous works for peptide, glycopeptide, and cross-linked peptide identification[8, 34, 35]. Peptide-level and glycan-level FDR were also tested by using mouse peptides and NeuAc-contained glycans, respectively (Fig. 2a).

Validation of O-glycopeptide search with IHMO data. In IHMO (Inhibitor-initiated Homogenous Mucin-type O-glycosylation), a kind of O-glycan elongation inhibitor, benzyl-N-acetyl-galactosaminide (GalNAc-O-bn), was applied to truncate the O-glycan elongation pathway during cell culture, generating cells with only truncated HexNAc(1) or HexNAc(1)NeuAc(1) O-glycans. sceHCD-pd-EThcD spectra were generated after O-glycopeptides were enriched by FASP[36], experimental details were shown in Supplementary Note 4. IHMO of HEK-293 cells were then verified by laser confocal microscopy as displayed in Supplementary Note Fig. 4. Spectra were then searched by pGlyco3, MetaMorpheus, MSFragger, and Byonic, the searching parameters were listed in Supplementary Table 3. For all software tools, Hex-contained O-glycopeptide could be still identified due to the inhibitor's imperfect efficiencies. The Hex-contained O-GPSMs were further validated by the summed intensities of the Hex-diagnostic ions (163.060 and 366.139 m/z) in their HCD spectra. The summed intensity threshold was set as 10% to the base peak.

Validation of pGlycoSite algorithm. For the given site-localization (SL) probabilities ($Prob$) of all identified GPSMs, the SL-FDR could be estimated as

$$\widehat{FDR}_{SL}(x) = \frac{\sum_{\forall i:\, 1\le i\le N, Prob_i\ge x}(1-Prob_i)}{\sum_{i=1}^{N} I(Prob_i\ge x)},$$

where $\widehat{FDR}_{SL}(x)$ is the estimated SL-FDR for a given probability threshold x, N is the total number of localized sites, and $I(bool)$ is the indicator function which returns 1 when $bool$ is true otherwise 0. It is not easy to validate the estimated SL-probability for a given site, but we can validate the accuracy of $\widehat{FDR}_{SL}(x)$, enabling SL-probability validation from another perspective. In this work, we designed two methods to validate $\widehat{FDR}_{SL}(x)$, entrapment-based validation and OpeRATOR-based validation.

For entrapment-based validation, after N-glycopeptide data were searched, the sites of GPSMs were localized using pGlycoSite by regarding the candidate sites as "J/S/T", which could be enabled by setting "glycosylation_sites=JST" in the searching parameter file. For a given GPSM, it would be the true positive (TP_{trap}) if J were the only localized sites, else all sites were the false positive (FP_{trap}). The entrapment-based SL-FDR could be calculated as

$$FDR_{trap}(x) = \frac{\#FP_{trap}(Prob\ge x)}{\#FP_{trap}(Prob\ge x)+\#TP_{trap}(Prob\ge x)},$$

for a given probability threshold x. Then SL probabilities of pGlycoSite could be validated by comparing $\widehat{FDR}_{SL}(x)$ with $FDR_{trap}(x)$.

For OpeRATOR-based validation, we used the data digested by OpeRATOR[31] and StcE[37] (StcE has similar cleavage amino acid sites for O-glycosylation as OpeRATOR). The identified peptides not starting with ST at the N-terminal were removed. Then for a given GPSM, we regarded it as the true positive (TP_{OpR}) if localized sites contained a site that was at the N-terminal S/T, else it was the false positive (FP_{OpR}). The OpeRATOR-based SL-FDR ($FDR_{OpR}(x)$) could be calculated from $TP_{OpR}(x)$ and $FP_{OpR}(x)$, which is similar to $FDR_{trap}(x)$.

Comparisons of the $\widehat{FDR}_{SL}(x)$ of pGlycoSite with $FDR_{trap}(x)$ and $FP_{OpR}(x)$ were displayed in Fig. 3g, and Supplementary Fig. 2. We also compared $\widehat{FDR}_{SL}(x)$ of MetaMorpheus with $FDR_{OpR}(x)$, as shown in Supplementary Fig. 2.

Analysis and verification of "Hex+17" of yeast samples. The "Hex+17" (abbreviated as "aH"), is defined as the Hex plus 17.027 Da. Peptides were searched by the yeast protein sequence databases (*S. pombe* for fission yeast and *S. cerevisiae* for budding yeast, Swissprot, 2018.08). N-glycan parts were searched against the high-mannose-only N-glycan database, O-mannosylation (O-Man) glycan parts were searched against the Hex-only glycan database. aH was regarded as a modified Hex for pGlyco3 search, and the maximal number of aH was set as 2 per glycan. For O-Man-glycopeptide search, the glycopeptide-diagnostic ion was set as Hex (163.060 m/z). $^{15}N/^{13}C$-labeled fission yeast (PXD005565) results were also validated by the ^{15}N-labeled and ^{13}C-labeled precursor signals in MS1 spectra. For $^{15}N/^{13}C$ validation of aH, as we did not know the exact chemical elements of the 17 Da of aH, here we assumed that the elements of aH are the same as Hex, with a 17.027-Da mass offset.

To further prove that the aH occurs on glycans instead of peptides (as a 17 Da modification on the peptide), we released N-glycans from fission yeast samples using PNGase F, and the released N-glycans were analyzed by mass spectrometry with sceHCD (Supplementary Note 5). All unique N-glycan compositions including aH-glycans were compiled into an N-glycan list. The N-glycan spectra were searched against the N-glycan list and scored by the glycan B/Y ion-counting score. Only the top-scored glycan was kept for each spectrum, and the results were further filtered by score ≥ 10. The identified aH-glycopeptides were validated by the filtered N-glycan search results.

O-Man glycopeptides of fission yeast were also analyzed by HCD followed by EThcD to investigate the O-mannosylation sites. The data were searched by the “HCD+EThcD” mode of pGlyco3, and the sites were localized by pGlycoSite. See Supplementary Note 7 for details.

Acknowledgements

This work is supported by grants from the National Key Research and Development Program (2018YFC0910302 to W.-F.Z. / M.-Q.L., 2016YFA0501300 to S.-M.H. / M.-Q.L. / W.-Q.C., and 2016YFB0201702 to P.-Y.Y.), the National Natural Science Foundation of China Project (91853102 to W.-Q.C.), and the innovative research team of high-level local university in Shanghai.

Data Availability

Data generated in this work, including yeast glycoproteomic data, yeast N-glycomic data and IHMO O-glycoproteomic data, could be downloaded from MassIVE with identifier MSV000086771 (username: MSV000086771_reviewer, password: pglyco).

Code Availability

pGlyco3 could be downloaded from https://github.com/pFindStudio/pGlyco3/releases. Analysis results and Python Notebooks to reproduce the comparison results could be downloaded from https://figshare.com/projects/Searched_results_and_python_notebooks_for_pGlyco3_manuscript/97592.

Author Contributions

W.-F.Z. conducted this project, developed the software, and analyzed the data. W.-Q.C. performed the MS experiments and analyzed the data. M.-Q.L. analyzed the data. S.-M.H. and P.-Y.Y. supervised this project. All the authors wrote and revised the manuscript.

Ethics declarations

Competing Interests

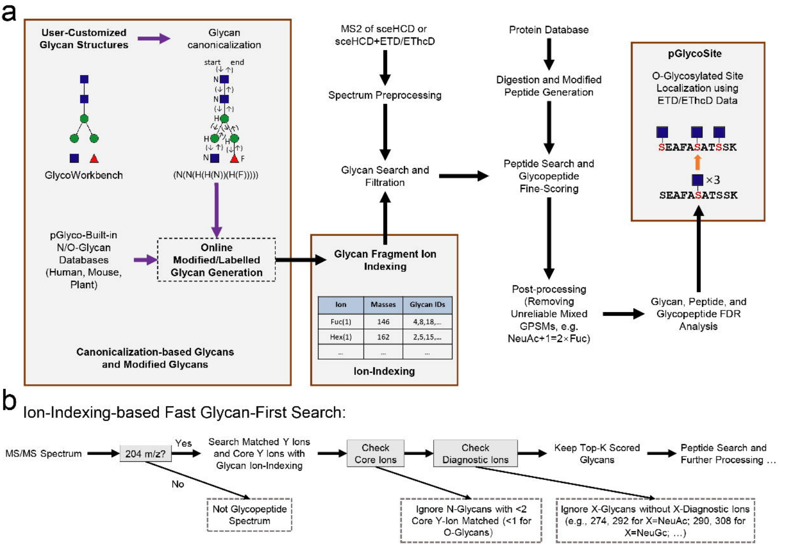
The authors declare no competing interests.

References

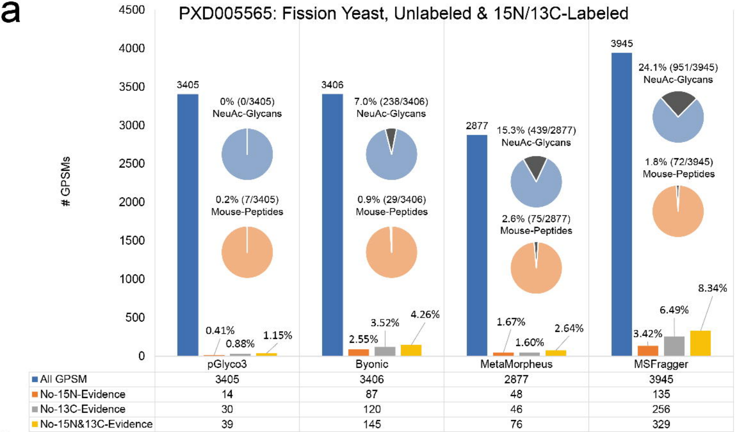
1. Reily, C., Stewart, T.J., Renfrow, M.B. & Novak, J. Glycosylation in health and disease. *Nat Rev Nephrol* 15, 346-366 (2019).
2. Smith, B.A.H. & Bertozzi, C.R. The clinical impact of glycobiology: targeting selectins, Siglecs and mammalian glycans. *Nat Rev Drug Discov* (2021).
3. Schjoldager, K.T., Narimatsu, Y., Joshi, H.J. & Clausen, H. Global view of human protein glycosylation pathways and functions. *Nat Rev Mol Cell Biol* 21, 729-749 (2020).
4. Ruhaak, L.R., Xu, G., Li, Q., Goonatilleke, E. & Lebrilla, C.B. Mass Spectrometry Approaches to Glycomic and Glycoproteomic Analyses. *Chem Rev* 118, 7886-7930 (2018).
5. Chen, Z., Huang, J. & Li, L. Recent advances in mass spectrometry (MS)-based glycoproteomics in complex biological samples. *Trends Analyt Chem* 118, 880-892 (2019).
6. Yu, Q. et al. Electron-Transfer/Higher-Energy Collision Dissociation (EThcD)-Enabled Intact Glycopeptide/Glycoproteome Characterization. *J Am Soc Mass Spectrom* 28, 1751-1764 (2017).
7. Caval, T., Zhu, J. & Heck, A.J.R. Simply Extending the Mass Range in Electron Transfer Higher Energy Collisional Dissociation Increases Confidence in N-Glycopeptide Identification. *Anal Chem* 91, 10401-10406 (2019).
8. Liu, M.Q. et al. pGlyco 2.0 enables precision N-glycoproteomics with comprehensive quality control and one-step mass spectrometry for intact glycopeptide identification. *Nat Commun* 8, 438 (2017).
9. Riley, N.M., Malaker, S.A. & Bertozzi, C.R. Electron-Based Dissociation Is Needed for O-Glycopeptides Derived from OpeRATOR Proteolysis. *Anal Chem* 92, 14878-14884 (2020).
10. Pap, A., Klement, E., Hunyadi-Gulyas, E., Darula, Z. & Medzihradszky, K.F. Status Report on the High-Throughput Characterization of Complex Intact O-Glycopeptide Mixtures. *J Am Soc Mass Spectrom* 29, 1210-1220 (2018).
11. Khoo, K.H. Advances toward mapping the full extent of protein site-specific O-GalNAc glycosylation that better reflects underlying glycomic complexity. *Curr Opin Struct Biol* 56, 146-154 (2019).
12. Riley, N.M., Malaker, S.A., Driessen, M.D. & Bertozzi, C.R. Optimal Dissociation Methods Differ for N- and O-Glycopeptides. *J Proteome Res* 19, 3286-3301 (2020).
13. Bern, M., Kil, Y.J. & Becker, C. Byonic: advanced peptide and protein identification software. *Curr Protoc Bioinformatics* Chapter 13, Unit13 20 (2012).
14. Strum, J.S. et al. Automated assignments of N- and O-site specific glycosylation with extensive glycan heterogeneity of glycoprotein mixtures. *Anal Chem* 85, 5666-5675 (2013).
15. Toghi Eshghi, S., Shah, P., Yang, W., Li, X. & Zhang, H. GPQuest: A Spectral Library Matching Algorithm for Site-Specific Assignment of Tandem Mass Spectra to Intact N-glycopeptides. *Anal Chem* 87, 5181-5188 (2015).
16. An, Z. et al. N-Linked Glycopeptide Identification Based on Open Mass Spectral Library Search. *Biomed Res Int* 2018, 1564136 (2018).
17. Xiao, K. & Tian, Z. GPSeeker Enables Quantitative Structural N-Glycoproteomics for Site- and

Structure-Specific Characterization of Differentially Expressed N-Glycosylation in Hepatocellular Carcinoma. *J Proteome Res* 18, 2885-2895 (2019).

18. Polasky, D.A., Yu, F., Teo, G.C. & Nesvizhskii, A.I. Fast and comprehensive N- and O-glycoproteomics analysis with MSFragger-Glyco. *Nat Methods* 17, 1125-1132 (2020).
19. Lu, L., Riley, N.M., Shortreed, M.R., Bertozzi, C.R. & Smith, L.M. O-Pair Search with MetaMorpheus for O-glycopeptide characterization. *Nat Methods* 17, 1133-1138 (2020).
20. Chi, H. et al. pFind-Alioth: A novel unrestricted database search algorithm to improve the interpretation of high-resolution MS/MS data. *J Proteomics* 125, 89-97 (2015).
21. Wu, S.W., Pu, T.H., Viner, R. & Khoo, K.H. Novel LC-MS(2) product dependent parallel data acquisition function and data analysis workflow for sequencing and identification of intact glycopeptides. *Anal Chem* 86, 5478-5486 (2014).
22. Stadlmann, J. et al. Comparative glycoproteomics of stem cells identifies new players in ricin toxicity. *Nature* 549, 538-542 (2017).
23. Lynn, K.S. et al. MAGIC: an automated N-linked glycoprotein identification tool using a Y1-ion pattern matching algorithm and in silico MS(2) approach. *Anal Chem* 87, 2466-2473 (2015).
24. Mao, J. et al. A New Searching Strategy for the Identification of O-Linked Glycopeptides. *Anal Chem* 91, 3852-3859 (2019).
25. Praissman, J.L. & Wells, L. Getting more for less: new software solutions for glycoproteomics. *Nat Methods* 17, 1081-1082 (2020).
26. Zeng, W.F. et al. pGlyco: a pipeline for the identification of intact N-glycopeptides by using HCD- and CID-MS/MS and MS3. *Sci Rep* 6, 25102 (2016).
27. Fang, P. et al. A streamlined pipeline for multiplexed quantitative site-specific N-glycoproteomics. *Nat Commun* 11, 5268 (2020).
28. Ranzinger, R., Herget, S., von der Lieth, C.W. & Frank, M. GlycomeDB--a unified database for carbohydrate structures. *Nucleic Acids Res* 39, D373-376 (2011).
29. Zhang, Y. et al. Identification of Glycopeptides with Multiple Hydroxylysine O-Glycosylation Sites by Tandem Mass Spectrometry. *J Proteome Res* 14, 5099-5108 (2015).
30. Ceroni, A. et al. GlycoWorkbench: a tool for the computer-assisted annotation of mass spectra of glycans. *J Proteome Res* 7, 1650-1659 (2008).
31. Yang, W., Ao, M., Hu, Y., Li, Q.K. & Zhang, H. Mapping the O-glycoproteome using site-specific extraction of O-linked glycopeptides (EXoO). *Mol Syst Biol* 14, e8486 (2018).
32. Huang, J. et al. OGP: A Repository of Experimentally Characterized O-Glycoproteins to Facilitate Studies on O-Glycosylation. *Biorxiv preprint* (2020).
33. Wulff-Fuentes, E. et al. The human O-GlcNAcome database and meta-analysis. *Sci Data* 8, 25 (2021).
34. Chi, H. et al. Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine. *Nat Biotechnol* (2018).
35. Chen, Z.L. et al. A high-speed search engine pLink 2 with systematic evaluation for proteome-scale identification of cross-linked peptides. *Nat Commun* 10, 3404 (2019).
36. Zielinska, D.F., Gnad, F., Wisniewski, J.R. & Mann, M. Precision mapping of an in vivo N-glycoproteome reveals rigid topological and sequence constraints. *Cell* 141, 897-907 (2010).
37. Malaker, S.A. et al. The mucin-selective protease StcE enables molecular and functional analysis of human cancer-associated mucins. *Proc Natl Acad Sci U S A* 116, 7278-7287 (2019).


a
User-Customized Glycan Structures
Glycan canonicalization
start end
N
N
H
H
H
N
F
GlycoWorkbench
(N(N(H(H(N))(H(F)))))
pGlyco-Built-in N/O-Glycan Databases (Human, Mouse, Plant)
Online Modified/Labelled Glycan Generation
Canonicalization-based Glycans and Modified Glycans
MS2 of sceHCD or sceHCD+ETD/EThcD
Spectrum Preprocessing
Glycan Search and Filtration
Glycan Fragment Ion Indexing
Ion-Indexing
Protein Database
Digestion and Modified Peptide Generation
Peptide Search and Glycopeptide Fine-Scoring
Post-processing (Removing Unreliable Mixed GPSMs, e.g. NeuAc+1=2×Fuc)
Glycan, Peptide, and Glycopeptide FDR Analysis
pGlycoSite
O-Glycosylated Site Localization using ETD/EThcD Data
SEAFASATSSK
×3
SEAFASATSSK
b
Ion-Indexing-based Fast Glycan-First Search:
MS/MS Spectrum
204 m/z?
Yes
No
Search Matched Y Ions and Core Y Ions with Glycan Ion-Indexing
Not Glycopeptide Spectrum
Check Core Ions
Check Diagnostic Ions
Keep Top-K Scored Glycans
Peptide Search and Further Processing ...
Ignore N-Glycans with <2 Core Y-Ion Matched (<1 for O-Glycans)
Ignore X-Glycans without X-Diagnostic Ions (e.g., 274, 292 for X=NeuAc; 290, 308 for X=NeuGc; ...)

Ion	Masses	Glycan IDs
Fuc(1)	146	4,8,18,...
Hex(1)	162	2,5,15,...
...	...	...

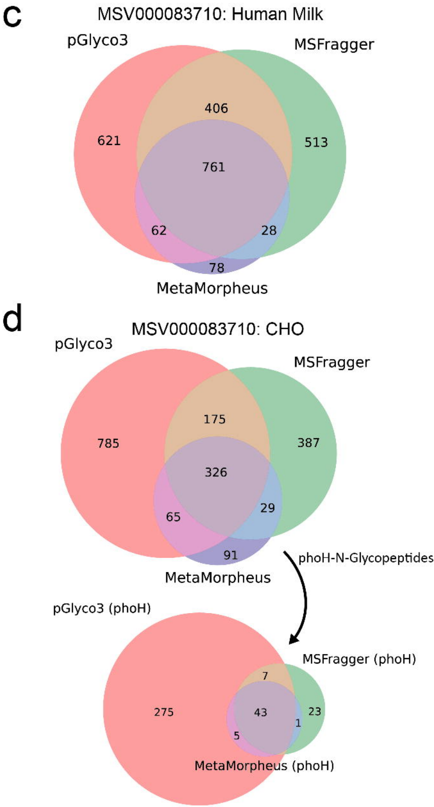


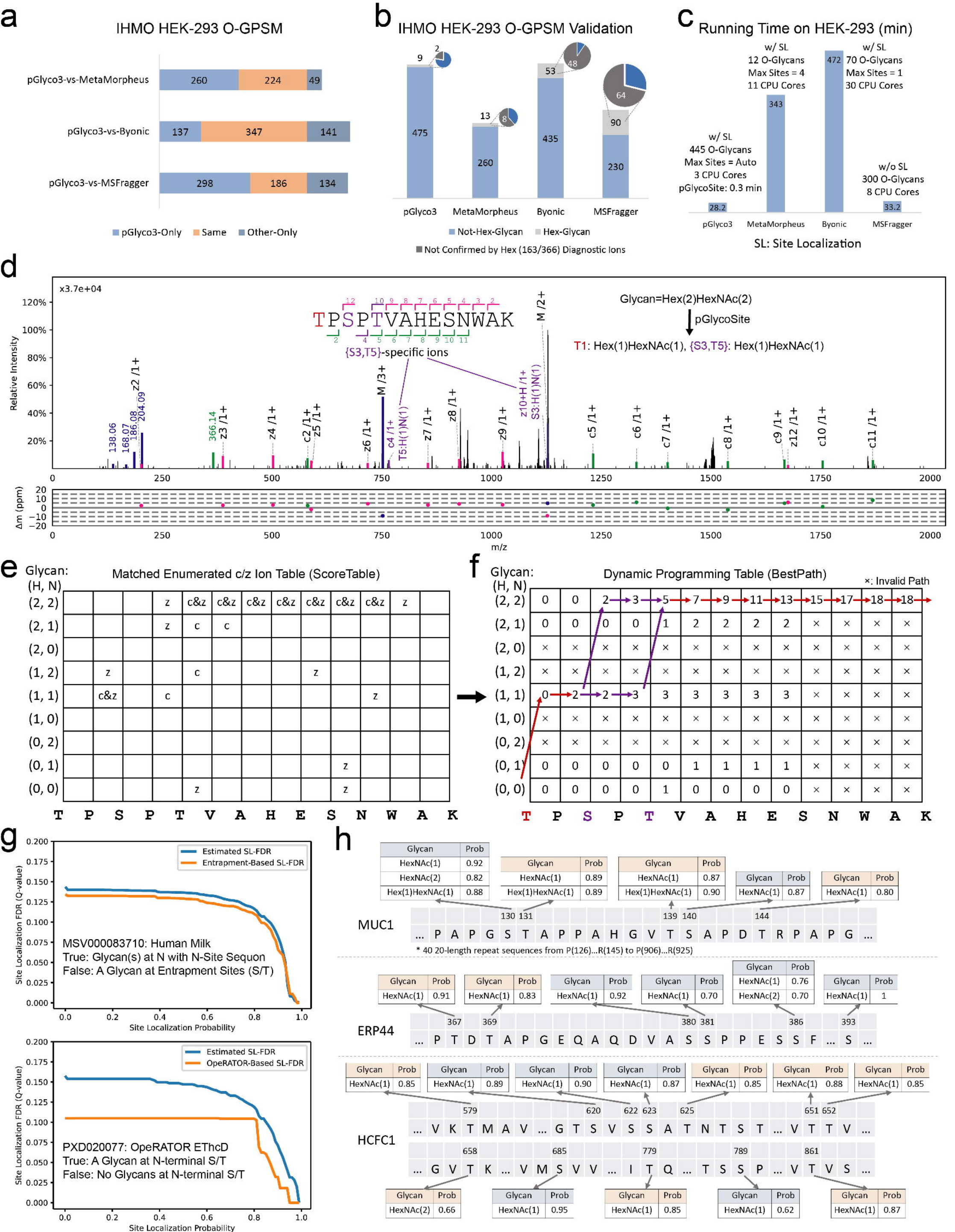
	pGlyco3	Byonic	MetaMorpheus	MSFragger
🟦 All GPSM	3405	3406	2877	3945
🟧 No-15N-Evidence	14	87	48	135
⬜ No-13C-Evidence	30	120	46	256
🟨 No-15N&13C-Evidence	39	145	76	329

b

Software	Mode	MS File	**Total Time (min)**
pGlyco3	Mouse N-Glycan (Large), 1622 Glycan Compositions	MGF	47
		RAW	116
Byonic	N-Glycan, 309 Glycan Compositions (Biggest DB in Byonic)	RAW	390
	N-Glycan, Glycan Compositions Same as pGlyco3	RAW	1,500
MSFragger	N-HCD, 182 Glycan Compositions	mzML	240
		RAW	490
	N-HCD-open	mzML	370
		RAW	650
MetaMorpheus	N-Glycan, 182 Glycan Compositions	RAW	1,900

Server: Dell PowerEdge R840 with Intel Xeon 6252 x 8, 512GB RAM
Configuration: 30 CPU cores (processors) for all software tools
MS data: 30 x 6h mouse raw files from Liu et al. 2017





Matched Enumerated c/z Ion Table (ScoreTable)

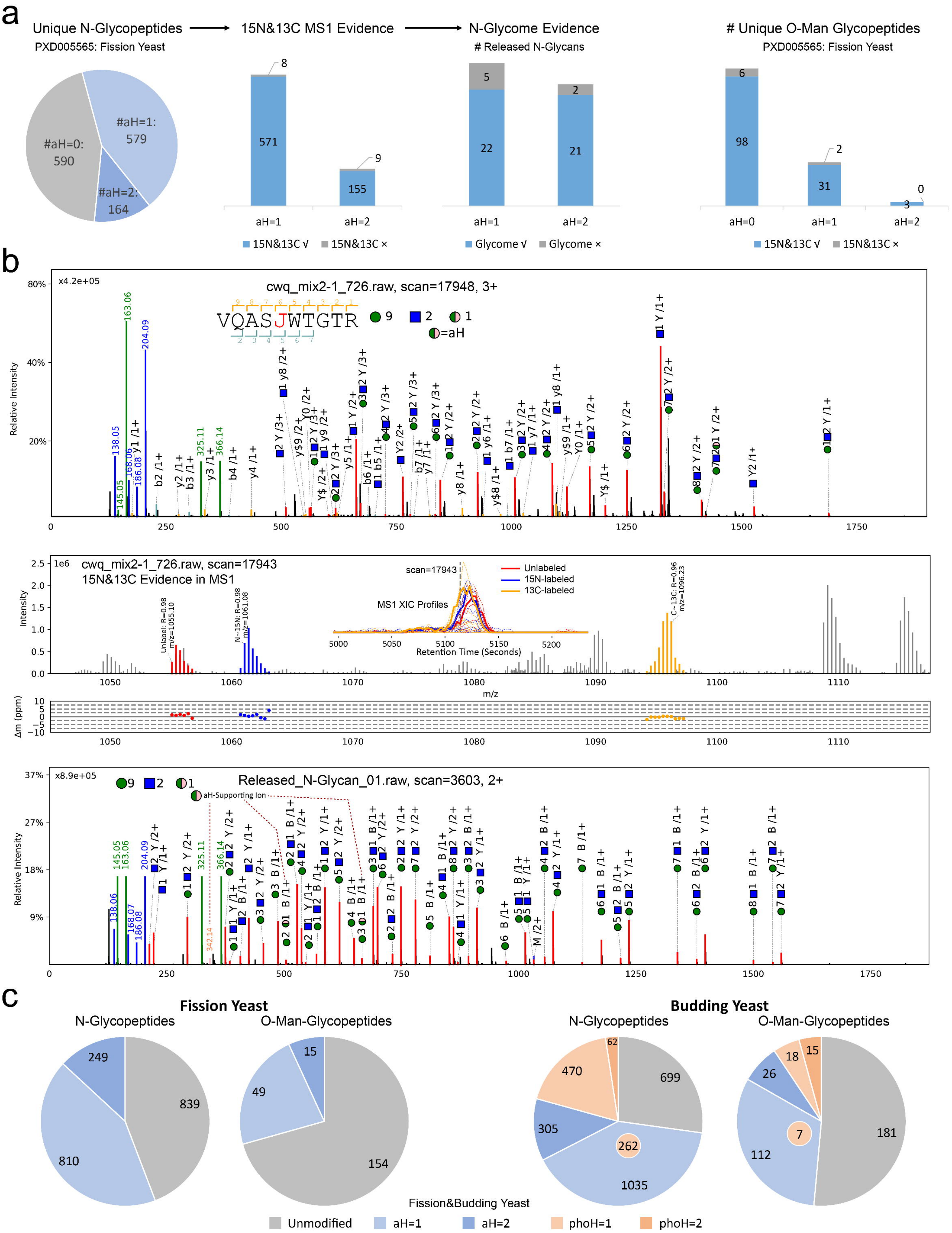
Glycan: (H, N)													
(2, 2)				z	c&z	c&z	c&z	c&z	c&z	c&z	c&z	z	
(2, 1)				z	c	c							
(2, 0)													
(1, 2)		z			c				z				
(1, 1)		c&z		c							z		
(1, 0)													
(0, 2)													
(0, 1)										z			
(0, 0)					z					z			

T P S P T V A H E S N W A K

Dynamic Programming Table (BestPath)

Glycan: (H, N)													
(2, 2)	0	0	2	3	5	7	9	11	13	15	17	18	18
(2, 1)	0	0	0	0	1	2	2	2	2	×	×	×	×
(2, 0)	×	×	×	×	×	×	×	×	×	×	×	×	×
(1, 2)	×	×	×	×	×	×	×	×	×	×	×	×	×
(1, 1)	0	2	2	3	3	3	3	3	3	×	×	×	×
(1, 0)	×	×	×	×	×	×	×	×	×	×	×	×	×
(0, 2)	×	×	×	×	×	×	×	×	×	×	×	×	×
(0, 1)	0	0	0	0	0	1	1	1	1	×	×	×	×
(0, 0)	0	0	0	0	1	0	0	0	0	×	×	×	×

T P S P T V A H E S N W A K


a
Unique N-Glycopeptides
PXD005565: Fission Yeast
#aH=1: 579
#aH=0: 590
#aH=2: 164
15N&13C MS1 Evidence
571
8
155
9
aH=1
aH=2
15N&13C √
15N&13C ×
N-Glycome Evidence
# Released N-Glycans
22
5
21
2
aH=1
aH=2
Glycome √
Glycome ×
# Unique O-Man Glycopeptides
PXD005565: Fission Yeast
98
6
31
2
3
0
aH=0
aH=1
aH=2
15N&13C √
15N&13C ×
b
x4.2e+05
cwq_mix2-1_726.raw, scan=17948, 3+
VQASJWTGTR
9
2
1
=aH
Relative Intensity
80%
40%
20%
0
250
500
750
1000
1250
1500
1750
163.06
204.09
138.05
145.05
168.06
186.08
325.11
366.14
y1 /1+
b2 /1+
y2 /1+
b3 /1+
y3 /1+
b4 /1+
y4 /1+
1 y8 /2+
1 Y /1+
2 Y /2+
Y0 /2+
Y2 /1+
Y$ /1+
1e6
cwq_mix2-1_726.raw, scan=17943
15N&13C Evidence in MS1
scan=17943
Unlabeled
15N-labeled
13C-labeled
MS1 XIC Profiles
Retention Time (Seconds)
5000
5050
5100
5150
5200
Unlabel: R=0.98 m/z=1055.10
N→15N: R=0.98 m/z=1061.08
C→13C: R=0.96 m/z=1096.23
Intensity
2.5
2.0
1.5
1.0
0.5
0.0
1050
1060
1070
1080
1090
1100
1110
m/z
Δm (ppm)
10
5
0
−5
−10
x8.9e+05
9
2
1
Released_N-Glycan_01.raw, scan=3603, 2+
aH-Supporting Ion
342.14
37%
27%
18%
9%
145.05
163.06
204.09
138.06
168.07
186.08
325.11
366.14
M /2+
0
250
500
750
1000
1250
1500
1750
c
Fission Yeast
N-Glycopeptides
249
839
810
O-Man-Glycopeptides
15
49
154
Budding Yeast
N-Glycopeptides
62
470
699
305
262
1035
O-Man-Glycopeptides
15
18
26
7
181
112
Fission&Budding Yeast
Unmodified
aH=1
aH=2
phoH=1
phoH=2