Journal of proteome research

Cite This: J. Proteome Res. XXXX, XXX, XXX−XXX

Article

pubs.acs.org/jpr

A Tandem Mass Spectrometry Sequence Database Search Method for Identification of O-Fucosylated Proteins by Mass Spectrometry

Kristian E. Swearingen,*[,†] Jimmy K. Eng,[‡] David Shteynberg,[†] Vladimir Vigdorovich,[§] Timothy A. Springer,[∥] Luis Mendoza,[†] D. Noah Sather,[§] Eric W. Deutsch,[†] Stefan H. I. Kappe,[§] and Robert L. Moritz*[,†]

[†]Institute for Systems Biology, Seattle, Washington 98109, United States

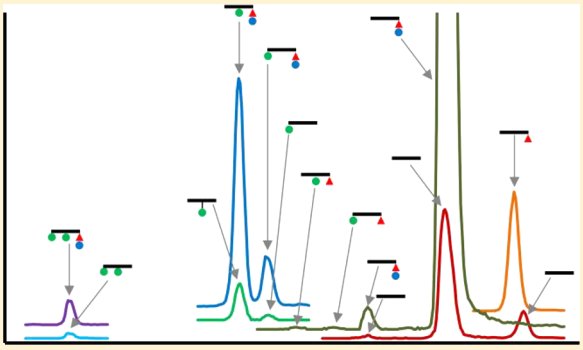
[‡]Proteomics Resource, University of Washington, Seattle, Washington 98195, United States

[§]Center for Global Infectious Disease Research, Seattle Children's Research Institute, Seattle, Washington 98101, United States

[∥]Harvard Medical School and Boston Children's Hospital, Boston, Massachusetts 02115, United States

Supporting Information

ABSTRACT: Thrombospondin type 1 repeats (TSRs), small adhesive protein domains with a wide range of functions, are usually modified with O-linked fucose, which may be extended to *O*-fucose-β1,3-glucose. Collision-induced dissociation (CID) spectra of O-fucosylated peptides cannot be sequenced by standard tandem mass spectrometry (MS/MS) sequence database search engines because O-linked glycans are highly labile in the gas phase and are effectively absent from the CID peptide fragment spectra, resulting in a large mass error. Electron transfer dissociation (ETD) preserves O-linked glycans on peptide fragments, but only a subset of tryptic peptides with low m/z can be reliably sequenced from ETD spectra compared to CID. Accordingly, studies to date that have used MS to identify O-fucosylated TSRs have required manual interpretation of CID mass spectra even when ETD was also employed. In order to facilitate high-throughput, automatic identification of O-fucosylated peptides from CID spectra, we re-engineered the MS/MS sequence database search engine Comet and the MS data analysis suite Trans-Proteomic Pipeline to enable automated sequencing of peptides exhibiting the neutral losses characteristic of labile O-linked glycans. We used our approach to reanalyze published proteomics data from *Plasmodium* parasites and identified multiple glycoforms of TSR-containing proteins.



KEYWORDS: *O-fucosylation, C-mannosylation, thrombospondin type 1 repeat, Plasmodium*

INTRODUCTION

Thrombospondin type 1 repeats (TSRs) are small (~60 amino acid residues) adhesive protein domains with a wide range of functions.[1] TSRs are often modified with two unique glycans, O-linked fucose (O-Fuc) and C-linked mannose (C-Man). The motif $CX_{2-3}(S/T)CX_2G$ is recognized by the *O*-fucosyltransferase POFUT2[2] and can be modified with O-Fuc at the serine/threonine (Ser/Thr or S/T) residue.[3] This fucose can be further extended to a fucose-β1,3-glucose disaccharide (O-Fuc-Glc) by the β1,3-glucosyltransferase B3GLCT.[4,5] TSRs also often contain WXXW and WXXC motifs that can be modified with a C-linked mannose (C-Man) at the tryptophan (Trp or W) residues.[6,7] In *C. elegans*, where the first *C*-mannosyltransferase was discovered,[8] the enzyme Dpy19 only C-mannosylates the first Trp of the WXXW motif. However, it has recently been shown that humans possess multiple homologues of Dpy19, and that Dpy19L1 preferentially modifies the first Trp of WXXW while Dpy19L3 preferentially modifies the Trp of WXXC.[9] In TSRs, all of these motifs often appear in tandem as $[WXX]_{1-3}CXX(S/T)CXXG$, and this short sequence may be modified with multiple combinations of C-Man, O-Fuc, and O-Fuc-Glc within the same organism, and even on the same protein.[3] The role of these TSR glycosylations is not fully characterized, and appears to vary by protein. In experiments where POFUT2 or Dpy19 have been deleted, certain TSR-containing proteins exhibit stability and trafficking defects, while other TSR-containing proteins show no discernible defect from loss of glycosylation.[9,10]

Given the heterogeneity of site occupancy and function observed to date for glycosylated TSRs, it is desirable to be able to detect and quantify TSR glycosylation in vivo. Mass spectrometry (MS) is typically the method of choice for high-throughput detection of protein post-translational modifications (PTMs) because a moiety covalently bound to a peptide can be directly detected as a mass shift, and the residue to which said

Special Issue: Software Tools and Resources 2019

Received: August 17, 2018
Published: December 7, 2018

Downloaded via HARVARD UNIV on December 31, 2018 at 16:23:16 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

ACS Publications © XXXX American Chemical Society

DOI: 10.1021/acs.jproteome.8b00638
J. Proteome Res. XXXX, XXX, XXX−XXX

modification is attached can usually be directly inferred from the MS^2 fragment spectra. However, detection of O-fucosylation of TSRs by MS is hindered by the fact that these O-linked glycans are highly labile in the gas phase[3] and are present at very low abundance or absent from the collision induced dissociation (CID) and higher-energy collision (HCD) MS^2 peptide fragment spectra that are typically used to sequence peptides in shotgun proteomics experiments. When the majority of fragment ions lack the glycan due to neutral loss, typical MS/MS sequence database search programs are incapable of identifying the peptide at all because of the large discrepancy between the precursor parent mass (with the glycan intact) and the masses of the deglycosylated fragments in the MS^2 spectrum. It has been demonstrated that a subset of peptides modified with O-Fuc may be sequenced by MS/MS sequence database search engines if they were fragmented with electron transfer dissociation (ETD), which leaves the O-linked glycan intact.[11,12] However, ETD tends to only produce quality fragment spectra at lower m/z, and many of the tryptic peptides that are readily sequenced from CID and HCD spectra cannot be sequenced using ETD.[13] For these reasons, the most common approach to identifying O-fucosylation of TSRs has been to manually interpret CID and HCD mass spectra of peptides bearing the predicted glycosylation motifs, elucidating the sequence of the peptide by the unmodified fragment ions, and inferring the identity of the O-linked glycan by the mass of the neutral loss.[3,14]

In order to facilitate automatic identification of O-Fuc and O-Fuc-Glc from typical CID and HCD shotgun proteomics experiments, we re-engineered several open-source software tools for proteomics data analysis in order to enable automated sequencing and annotation of peptides exhibiting the neutral losses characteristic of gas-phase-labile PTMs. To demonstrate the utility of our approach, we reanalyzed published proteomics data that had previously been manually interpreted to identify O-fucosylation of TSRs in the *Plasmodium* parasites *P. falciparum* and *P. vivax*. Our method validated these results with dozens of automatically identified and annotated corroborating spectra. We were also able to use our approach to identify the first evidence for TSR glycosylation in *P. yoelii*.

EXPERIMENTAL PROCEDURES

Sample Preparation

A segment of *P. falciparum* circumsporozoite protein (CSP; PF3D7_0304600) bearing the TSR domain was expressed in HEK293 cells and purified as described elsewhere.[15] One hundred μg of purified protein was resuspended in 100 μL of 50 mM ammonium bicarbonate (ABC) and 5 mM tris(2-carboxyethyl)phosphine (TCEP; (ThermoFisher Bond-Breaker Neutral pH solution) and incubated 15 min at 50 °C, then added to a Microcon YM-3 3 kDa molecular weight cutoff filter (Amicon Bioseparations) that had been washed with ABC. After reducing the buffer volume by centrifugation, 100 μL of 10 mM iodoacetamide (IAM) in 50 mM ABC was added to the filter and incubated 20 min at room temperature (RT) in darkness. The IAM was removed by centrifugation, the filter was washed once with 100 μL of 50 mM ABC, and 2 μg of trypsin (Promega, sequencing grade) in 50 μL of 50 mM ABC was added to the filter, mixed by vortexing 1 min at 600 rpm in a thermomixer, and incubated overnight at 37 °C. The digested peptides were collected by centrifugation, dried in vacuum concentrator, and reconstituted in 50% (v/v) methanol/0.1% (v/v) formic acid (FA) to a nominal peptide concentration of 1 pmol/μL.

Recombinant *P. falciparum* thrombospondin-related anonymous protein (TRAP; PF3D7_1335900) was expressed in HEK293 cells and purified as described elsewhere.[16] Six μg of *Pf*TRAP was digested with trypsin using an S-Trap Micro (Protifi) following the manufacturer's instruction. Briefly, to a 6 μL aliquot of a 1 μg/μL solution of purified protein in HEPES/NaCl buffer, 0.6 μL of 0.5 M TCEP was added (final concentration 45 mm) and incubated 5 min at 95 °C. The solution was cooled and 0.6 μL of 1 M IAM was added and incubated 20 min at RT in darkness. Lysis buffer (5% (w/v) sodium dodecyl sulfate (SDS) in 50 mM ABC) was added to achieve a final volume of 25 μL, followed by 2.5 μL of 12% (v/v) phosphoric acid, then 165 μL of S-Trap buffer (100 mM ABC in 90% (v/v) methanol). The colloidal protein precipitate that formed was collected on the S-Trap by centrifugation and washed three times with S-Trap buffer, after which 2 μg of trypsin in 20 μL of 50 mM ABC was placed on the trap. The trap was incubated 1 h at 47 °C and the peptides were collected by centrifugation, followed by three additional 40 μL washes of 50 mM ABC, 0.2% (v/v) FA, and 50% (v/v) acetonitrile (ACN). All elution steps were combined, dried in a vacuum concentrator, and reconstituted in 5% (v/v) ACN/1% (v/v) trifluoroacetic acid (TFA) to a nominal peptide concentration of 0.2 μg/μL.

Mass Spectrometry

Mass spectrometry (MS) was performed on an LTQ-Velos Pro Orbitrap Elite (ThermoFisher). The *Pf*CSP tryptic digest was infused directly at 500 nL/min through an empty Picofrit column (New Objective) with a 75 μm internal diameter (ID) and a 15 μm ID fritted tip. Mass spectra, including collision-induced dissociation (CID) and higher-energy dissociation (HCD) MS^n were collected in the Orbitrap with a nominal resolution of 240 000 at 400 m/z. Spectra were acquired through the direct acquisition function in the instrument tune window. Ten spectra were collected for each combination of normalized collision energy (CE) and MS^n and signal-averaged in XCalibur to produce the spectra presented here. The *Pf*TRAP tryptic digest was analyzed by nanoflow liquid chromatography (LC)−MS using an Agilent 1100 with electronically controlled split-flow coupled to the MS. Injections of nominally 0.16 or 0.2 μg of peptides were loaded onto a 150 μm ID fused silica trap column made in-house with a 10 mm bed of C18 (Dr. Maisch ReproSil Pur C18 AQ, 120 Å, 3 μm) and washed with LC load buffer (2% (v/v) ACN/0.2% (v/v) TFA) prior to separation on a column packed in-house with a 20 cm bed of the same stationary phase in a Picofrit column with a 75 μm ID and a 10 μm ID fritted tip. Mobile phase A was 0.1% (v/v) FA in H_2O and mobile phase B was 0.1% (v/v) FA in ACN. The separation gradient was as follows, all at 300 nL/min: 5% B to 35% B in 60 min, 35% B to 80% B in 10 min, 5 min at 80% B, 80% B to 5% B in 1 min, and 29 min at 5% B to re-equilibrate the column. For the collision energy (CE) scanning experiment, MS^1 scans were collected in the Orbitrap from 400−1500 m/z at a resolution of 30 000. A data-dependent acquisition (DDA) method was employed to select the eight most abundant precursors for CID with a normalized CE of 35%. CID spectra were collected in the Orbitrap at a resolution of 15 000. Monoisotopic precursor selection was enabled. Singly charged precursors and those with unassigned charge states were excluded from selection. An inclusion list was employed to preferentially select 1248.01 m/z and 1329.03 m/z at CEs of 0, 5, 10, 15, 20, 25, 30, and 35%. Dynamic exclusion was not enabled. For the experiment

DOI: 10.1021/acs.jproteome.8b00638
J. Proteome Res. XXXX, XXX, XXX−XXX

targeting the multiple glycoforms of the glycosite-containing peptide, the MS method was modified such that the MS^1 scans were collected over 1000−1500 m/z and only doubly charged precursors were selected for fragmentation. The inclusion list is provided in Table S1.

Software

The tandem mass spectrometry sequence database search engine Comet[17] was re-engineered to support the analysis of precursor mass offsets by the implementation of the "mass_offsets" parameter. Within the Comet code, this functionality is a simple extension to the precursor mass check function where each user defined mass offset value is subtracted from the experimental precursor mass such that peptides that are smaller than the experimental precursor mass by the mass offset value can still be matched to the spectrum. Multiple mass offsets can be specified within a single search. This function was first implemented in Comet version 2015.02 rev.0.

The peptide spectrum matches (PSMs) produced by Comet were analyzed using the Trans-Proteomics Pipeline (TPP) version 5.1.0 Syzygy.[18] In this version, the TPP spectrum level validation tool PeptideProphet was modified to accommodate PSMs obtained with the mass_offsets parameter enabled in Comet. Specifically, PeptideProphet has a mass difference model that accounts for mass differences between the measured mass of the precursor ion and the theoretical mass of the matched peptide. This model aims to increase the probability of PSMs that have a small mass difference between the measured and theoretical masses and decrease the probability of PSMs that have a large mass difference, with the actual parameters of the model learned from the data set using the EM algorithm. Considering the mass_offsets parameter in Comet, where the PSMs can now have a possible number of large mass offsets and still be correct, this model had to be adjusted. Given that the mass error of PSMs coming from searched peptides at a nonzero mass offset are expected to be off by the mass of the lost modification for every PSM and for every possible mass offset, the mass difference model in PeptideProphet finds the correct mass difference by considering the mass differences between the measured precursor mass and the matched peptide mass with the appropriate mass offset to correct for the neutral loss of the labile modification. After considering all possible mass offsets, the mass difference that is closest to zero is used as the corrected mass difference for consideration in the mass difference model.

Additionally, the spectrum viewing tool Lorikeet (http://uwpr.github.io/Lorikeet/), which is incorporated into the TPP, was updated to automatically annotate features specific to C-mannosylated peptides. Upon recognizing modification of Trp residues with a mass of 162.05 (hexose), the viewer now provides the option of annotating neutral loss of 120.04 (from cross-ring cleavage of C-Man) from the precursor and putatively hexosylated fragments.

Peak List Generation

The MS data and associated database and analysis files generated for this work are available at www.peptidatlas.org[19] using the identifier PASS01201. The MS data from previously reported analyses of salivary gland sporozoites[20−22] were obtained from PeptideAtlas using the identifiers PASS00095 and PASS00729 (*P. falciparum* salivary gland sporozoites), PASS00098 (*P. yoelii* salivary gland sporozoites), and PASS00976 (*P. vivax* salivary gland sporozoites). Mass spectrometer output files were converted to mzML format using msConvert version 3.0.6002 from the ProteoWizard toolkit[23] and searched with Comet version 2017.01 rev.1.[17] The precursor mass tolerance was set to ±10 ppm. For the recombinant *Pf*TRAP LC−MS data, for which the MS^2 were collected at high resolution in the Orbitrap, the fragment ions bins were set to a tolerance of 0.02 m/z, the monoisotopic mass offset was 0.0, and the theoretical fragment ions setting was set at 0 ("use flanking peaks"). For the sporozoite global proteome data, for which the MS^2 were collected at low resolution in the LTQ, the bin tolerance was set to 1.0005 m/z, the monoisotopic mass offset was set to 0.4 m/z, and the theoretical fragment ions setting was 1 ("M peak only"). Semitryptic peptides and up to 2 missed cleavages were allowed. The search parameters included a static modification of +57.021464 Da at Cys for formation of *S*-carboxamidomethyl-Cys by IAM and potential modifications of +15.994915 Da at Met and Trp for oxidation, +31.989829 Da at Trp for dioxidation, and +162.052824 Da at Trp for C-mannosylation. The mass_offset parameter was set to allow mass offsets of 0, +146.057909, and +308.1107321 Da for no neutral loss and neutral loss of deoxyhexose or hexosyldeoxyhexose, respectively. The recombinant *Pf*TRAP spectra were searched against a database comprising *P. falciparum* TRAP (PF3D7_1335900) and the common Repository of Adventitious Proteins (v.2012.01.01, The Global Proteome Machine, www.thegpm.org/cRAP, 116 entries). The sporozoite data were searched against *P. falciparum* 3D7[24] (PlasmoDB v.35, www.plasmodb.org,[25] 5548 entries), *P. yoelii yoelii* 17X[26] (PlasmoDB v.35, 6092 entries), or *P. vivax* P01[27] (PlasmoDB v.31) appended with sequence polymorphisms observed in other field isolates[22] (6670 entries), as appropriate for the sample. Each of these *Plasmodium* databases was appended with the mosquito protein database *Anopheles stephensi* Indian AsteI2.3[28] (VectorBase, www.vectorbase.org,[29] 11 789 entries), and the cRAP proteins. For each database, decoy proteins with the residues between tryptic residues randomly shuffled were generated using a tool included in the TPP and interleaved among the target entries. The MS^2 spectra were analyzed using the TPP version 5.1.0 Syzygy. PSMs were assigned scores in PeptideProphet[30] using the following options: nonparametric model (NONPARAM), use Expect value as the only contributor to the f-score (EXPECTSCORE), accurate mass binning of high-accuracy precursor mass (ACCMASS PPM), report results with minimum probability of zero (ZERO), identify decoy entries in the database and use decoys to build models (DECOY = DECOY). For the recombinant *Pf*TRAP data, only very high-quality PSMs were counted (those identified with a PeptideProphet probability of 1.0 and an Expect score less than 10^{-5}). For the reanalysis of sporozoite proteome data sets, the DECOYPROBS flag was used to assign possible nonzero probabilities to decoy entries, enabling determination of the decoy-estimated false discovery rate (FDR) among peptides identified with mass shifts. Only PSMs identified with a PeptideProphet probability corresponding to a model-estimated FDR less than 1.0% were taken for further analysis.

RESULTS AND DISCUSSION

Behavior of *O*-Fucose and *C*-Mannose in Collision-Induced Dissociation

The TSR-containing proteins circumsporozoite protein (CSP) and thrombospondin-related anonymous protein (TRAP) are major surface proteins of salivary gland sporozoites, the form of the *Plasmodium* parasite that is transmitted from the mosquito vector to the vertebrate host. Expressing CSP and TRAP in

DOI: 10.1021/acs.jproteome.8b00638
J. Proteome Res. XXXX, XXX, XXX−XXX

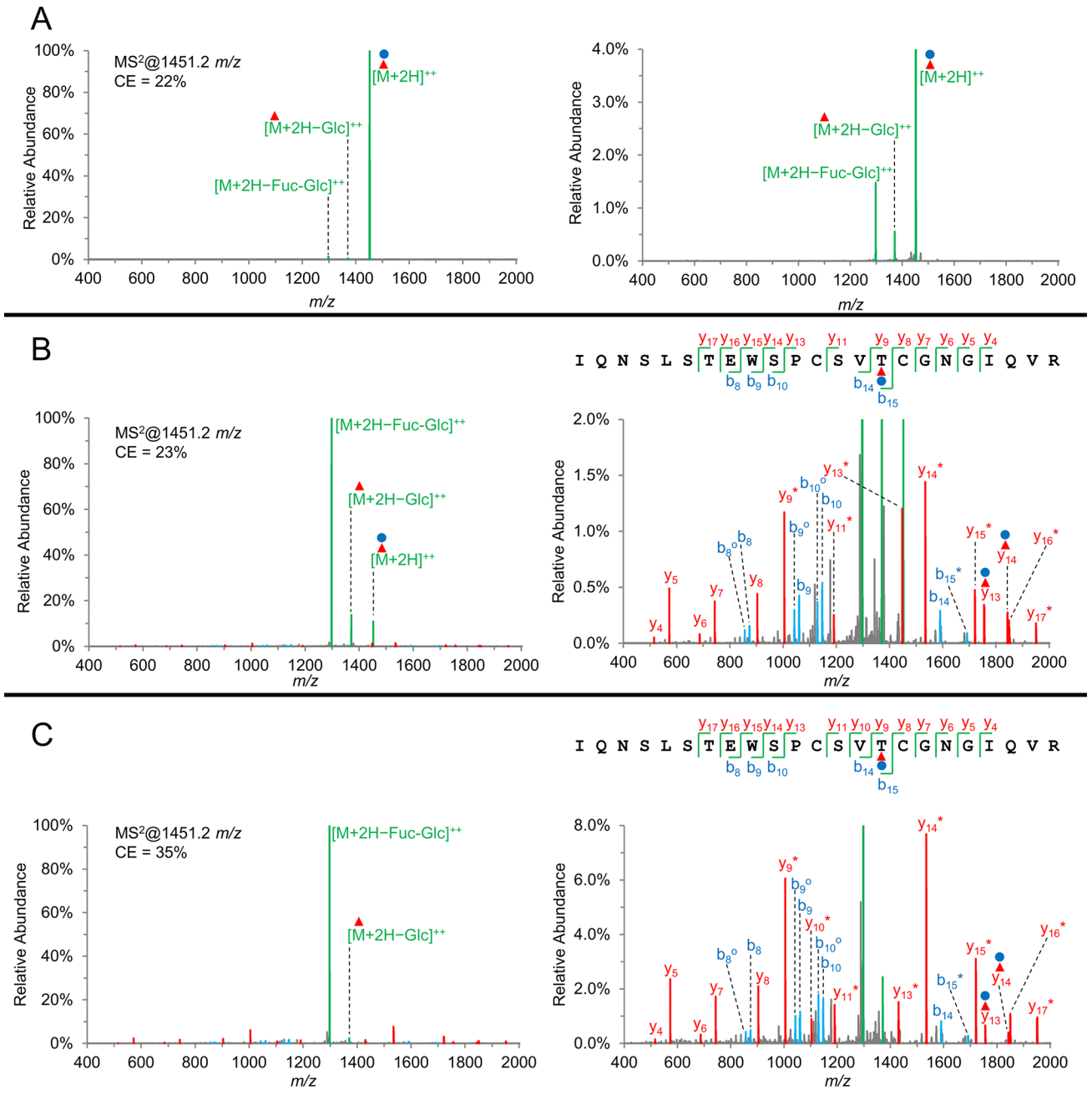


Figure 1. Collision-induced dissociation of an O-fucosylated peptide. A tryptic digest of recombinant *P. falciparum* circumsporozoite protein (CSP) was directly infused and analyzed by NSI-MS/MS. The precursor parent ion with $[M + 2H]^{2+}$ = 1451.17 *m/z*, corresponding to the mass of the potentially glycosylated peptide plus a hexose and a deoxyhexose, was fragmented by collision-induced dissociation (CID) at multiple normalized collision energies (CE). (A) At CE = 22%, the precursor ion was mostly intact, but small doubly charged peaks (detail at right) were detectable at 1370.14 *m/z* and 1297.11 *m/z*, corresponding to neutral loss of hexose (i.e., glucose) and hexose plus deoxyhexose (i.e., glucosylfucose), respectively. (B) Just a slight increase in collision energy to CE = 23% was sufficient to fragment most of the parent precursor ion. The dominant species was the 1297.11 *m/z* peak resulting from neutral loss of the O-linked disaccharide. Peptide fragment ions could be detected at very low levels (detail at right). Red triangles are O-Fuc, blue circles are Glc bound to O-Fuc by a β-1,3 linkage, neutral loss of water from fragment ions is indicated with "o", and neutral loss of O-Fuc-Glc is indicated with "*". Most fragment ions were missing the O-linked glycan. However, y_{13} and y_{14} peaks retaining the O-Fuc-Glc were detected, positively localizing the O-linked glycan to the expected Thr residue. The y_{13} and y_{14} fragments produced the highest abundance fragment ion peaks in CID spectrum of the unglycosylated peptide (Figure S1), a common feature of y-ions with N-terminal Pro residues. (C) At CE = 35% the parent precursor was completely fragmented and the relative abundance of the peptide fragment ions had increased approximately 4-fold, but the spectrum was still dominated by the intact peptide stripped of the glycan.

HEK293 cells has been shown to modify the TSRs of these proteins with C-Man and O-Fuc-Glc.[15,31,32] We analyzed recombinant *P. falciparum* CSP and TRAP by tandem mass spectrometry (MS/MS) and nanoflow liquid chromatography (nanoLC)-MS/MS in order to characterize the behavior of O-Fuc and C-Man in CID. The recombinant *Pf*CSP was digested with trypsin and infused for direct analysis by nanospray ionization (NSI)-MS/MS. On the basis of signal intensity of the parent ions, ~92% of glycosite-bearing peptide was modified with a mass equal to hexose plus deoxyhexose ($[M + 2H]^{2+}$ =

DOI: 10.1021/acs.jproteome.8b00638
J. Proteome Res. XXXX, XXX, XXX−XXX

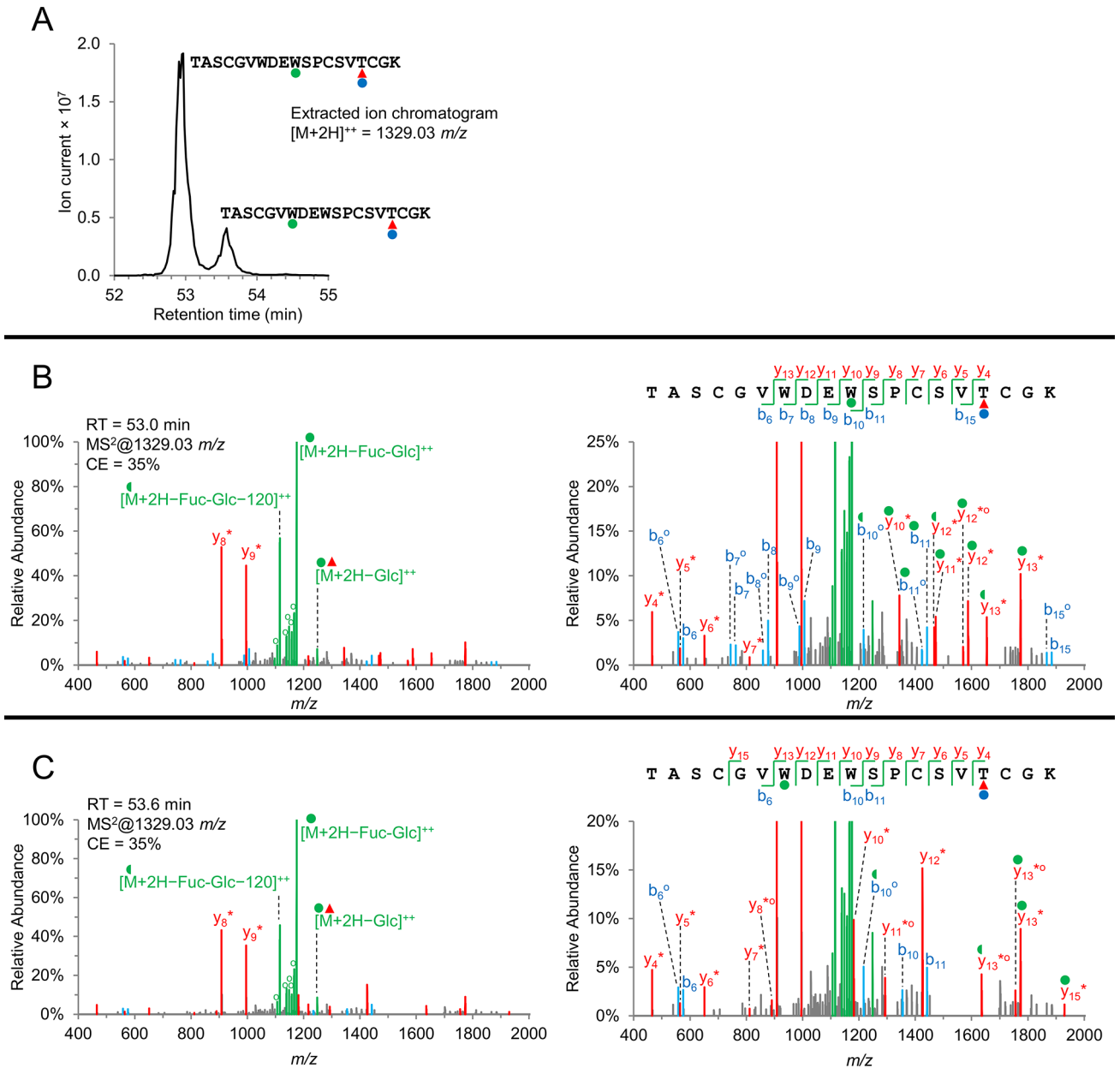


Figure 2. Collision-induced dissociation of a C-mannosylated and O-fucosylated peptide. A tryptic digest of recombinant *P. falciparum* thrombospondin-related anonymous protein (TRAP) was analyzed by LC−MS/MS. A targeted method was used that isolated the 1329.03 m/z precursor ion matching the mass of the potentially glycosylated peptide plus two hexoses and a deoxyhexose. (A) Two isobaric glycoforms of the peptide, each a positional isomer of the singly C-mannosylated peptide, exhibited distinct LC retention times. (B,C) Representative MS^2 spectra identifying the isoforms. Green circles are C-Man, green half-circles are C-Man with neutral loss of 120.04 Da due to cross-ring cleavage, red triangles are O-Fuc, blue circles are Glc bound to O-Fuc by a β-1,3 linkage. Neutral loss of water from fragment ions is indicated with "o", and neutral loss of O-Fuc-Glc is indicated with "*". At the normalized collision energy employed (CE = 35%), the parent precursor ion was completely fragmented, but the dominant peak matched that of the peptide after neutral loss of a hexose and a deoxyhexose, i.e., O-Fuc-Glc. No fragment ions retained the O-linked glycan, but the attachment site is known to be the C-terminal Thr residue in the conserved TSR motif. The remaining hexose withstood CID and could be localized to either one of the two Trp residues, consistent with C-Man. This identification and localization were further corroborated by neutral loss of 120.04 Da from cross-ring cleavage of C-Man.

1451.17 m/z), while less than 2% was modified with a deoxyhexose and ∼7% was unmodified. The 1451.17 m/z peak was isolated and the CID normalized collision energy (CE) was incrementally increased (Figure 1). At the CEs required for fragmentation of the peptide backbone, the peptide underwent neutral losses producing a dominant doubly charged ion matching the mass of the unmodified peptide ($[M + 2H]^{2+}$ = 1297.11 m/z) as well as lower-intensity doubly charged species matching the mass of the peptide modified with a single deoxyhexose (i.e., fucose, $[M + 2H]^{2+}$ = 1370.14 m/z). MS^2, MS^3 and MS^4 of these species confirmed the sequential loss of Glc and Fuc as well as the sequence of the peptide (Figure S1). Taken together, these spectra demonstrate that the peptide was modified with an O-linked hexosyldeoxyhexose, i.e., O-Fuc-Glc. Upon fragmentation at a CE typical of peptide sequencing (i.e., 35%), the precursor and fragment b- and y-ions were predominantly stripped of the labile glycan, resulting in an MS^2 spectrum effectively identifying the unmodified peptide.

DOI: 10.1021/acs.jproteome.8b00638
J. Proteome Res. XXXX, XXX, XXX−XXX

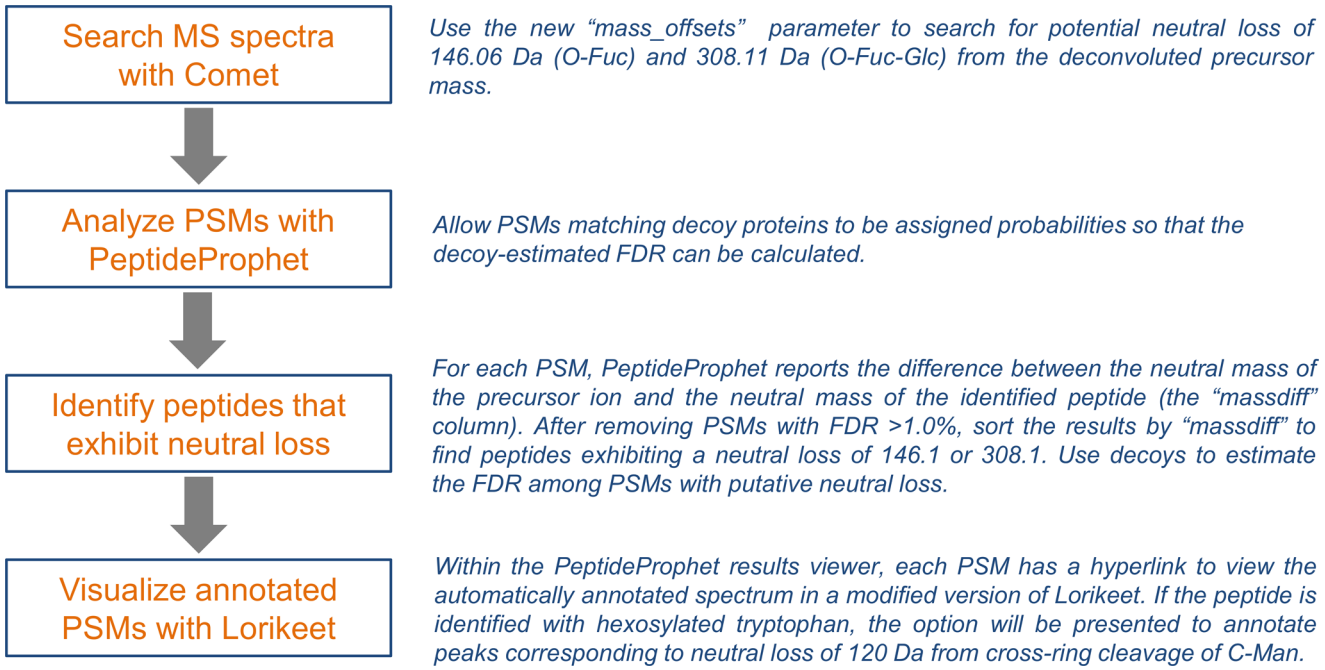


Figure 3. Summary of the data analysis workflow used to identify putatively O-fucosylated peptides via an automated sequence database search engine.

Table 1. Glycopeptides Identified from LC−MS/MS Analysis of Recombinant *Pf*TRAP[a]

modified peptide[b]	O-linked glycan[c]	$[M + 2H]^{2+}$ m/z[d]	retention time[e] (min)	proportion in sample[f] (%)	PSMs[g]
TASCGVW$_{Man}$DEW$_{Man}$SPCSVTCGK	Fuc-Glc	1410.06	50.5	1.27	9
TASCGVWDEW$_{Man}$SPCSVTCGK	Fuc-Glc	1329.03	54.7	11.3	19
TASCGVW$_{Man}$DEWSPCSVTCGK	Fuc-Glc	1329.03	55.4	2.39	11
TASCGVWDEW$_{Man}$SPCSVTCGK	Fuc	1248.01	56.1	0.09	3
TASCGVW$_{Man}$DEWSPCSVTCGK	Fuc	1248.01	57.0	0.08	1
TASCGVWDEWSPCSVTCGK	Fuc-Glc	1248.01	57.8, 59.7	78.8	120
TASCGVWDEWSPCSVTCGK	Fuc	1166.98	61.4	4.91	20
TASCGVWDEWSPCSVTCGK	None	1093.95	61.6	1.18	111

[a]A tryptic digest of recombinant *P. falciparum* TRAP expressed in mammalian cells was analyzed by LC−MS/MS, and the MS^2 spectra were searched for evidence of O-linked fucose or glucosylfucose by our automated database searching method. [b]The peptide from the TSR domain of *P. falciparum* TRAP containing potential glycosylation sites. The localization of C-Man was directly detected from the MS^2 fragment spectra. Any O-Fuc or O-Fuc-Glc is presumed to have been attached at the C-terminal Thr residue. [c]The identity of the O-linked glycan as inferred by the neutral loss of 146.06 Da (O-Fuc) or 308.11 Da (O-Fuc-Glc) from the parent mass upon MS^2 fragmentation. [d]The m/z of the doubly charged peptide parent ion. [e]The retention time of the peptide (see Figure 4). [f]The relative proportion of the peptide in the sample based on chromatographic peak height (see Figure 4). [g]The number of high-quality peptide spectrum matches (PSMs) identifying the peptide (see Table S2). Note that the total number of PSMs identifying the unglycosylated peptide (1093.95 m/z) was artificially high due to neutral loss of the O-linked glycan from the highly abundant O-fucosylated, unmannosylated species prior to MS^2 fragmentation.

However, two fragment ions retaining the glycan were detectable, enabling positive localization of the O-Fuc-Glc to the Thr residue of the conserved O-fucosylation motif.

Recombinant *Pf*TRAP was digested with trypsin and analyzed by nanoLC−MS/MS employing a data-dependent analysis (DDA) approach. An inclusion list was used to target the parent ion of the glycosite-bearing peptide modified with a mass equal to two hexoses and a deoxyhexose ($[M + 2H]^{2+}$ = 1329.03 m/z). The precursor ion was fragmented at incrementally increasing CEs as above (Figure S2). As with the recombinant *Pf*CSP, CID resulted in neutral loss of the O-linked glycan, revealing that the peptide was modified with an O-linked hexosyldeoxyhexose, i.e., O-Fuc-Glc, as well as a hexose at a single Trp residue, i.e., C-Man (Figure 2). Unlike O-linked glycans, C-Man withstands CID and was readily localized from the dominant b- and y-ions. Furthermore, C-Man can undergo cross-ring cleavage resulting in a neutral loss of 120.04 Da ($C_4H_8O_4$) and producing diagnostic precursor and fragment ion peaks.[3] Notably, two isobaric 1329.03 m/z species were present that had distinct LC retention times (Figure 2A), and had the C-Man attached at the first or the second Trp residue of the WXXW motif (Figure 2B,C).

Identification of O-Fucosylated Proteins by Automated Sequence Database Searching

Automated MS/MS sequence database search tools routinely identify protein PTMs by allowing for variable modification of peptide residues with fixed masses. However, standard automatic search approaches are unable to identify O-glycosylated peptides such as those described above because the most intense fragment ions lack the glycan and thus do not match the predicted fragmentation spectra. Nonetheless, based on the observation that the MS^2 spectra of such peptides can confidently identify the sequence of the unmodified peptide, we reasoned that these peptides could be identified by a sequence

DOI: 10.1021/acs.jproteome.8b00638
J. Proteome Res. XXXX, XXX, XXX−XXX

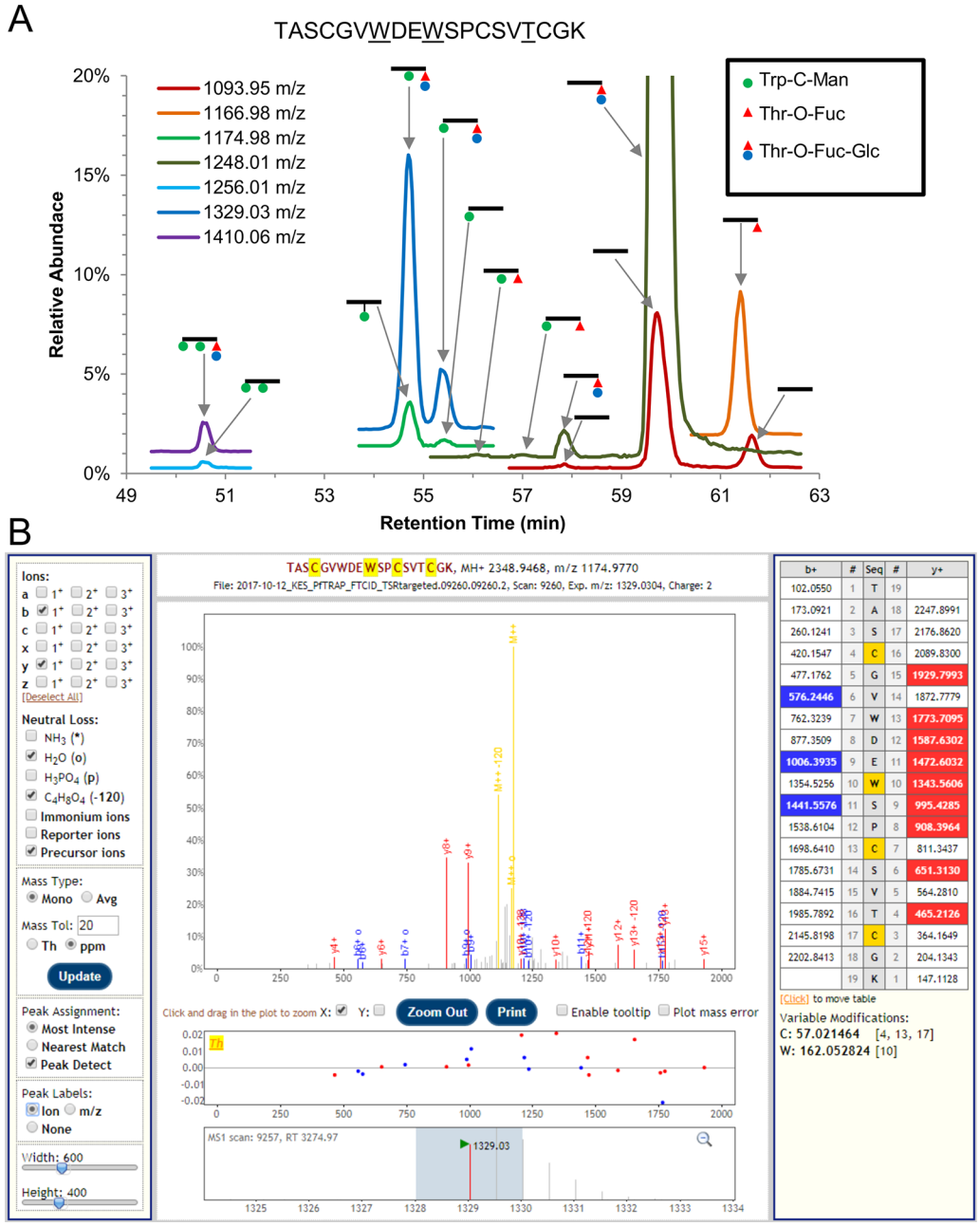


Figure 4. Automated identification of multiple glycoforms of recombinant *Pf* TRAP. A tryptic digest of recombinant *P. falciparum* TRAP was analyzed by LC−MS/MS. A targeted method was used that preferentially selected masses corresponding to the potentially glycosylated peptide modified with the various combinations of two, one, or no C-Man modifications and an O-Fuc-Glc, O-Fuc, or no O-linked glycan. The glycoforms were identified using our automated database searching approach (See summary of results in Table 1). (A) Extracted ion chromatograms of the targeted m/z values (all are $[M + 2H]^{2+}$). Traces are offset for clarity. The peptide sequence is shown with the potential C-mannosylation and O-fucosylation sites underlined. Peaks are labeled with the identified glycoform according to the inset legend. For the major species, evidence of neutral loss of the O-linked glycan due to in-source collision-induced dissociation prior to MS^2 fragmentation could be seen as lower-intensity, coeluting peaks. (B) A representative MS^2 spectrum that has been automatically annotated by a version of the spectrum visualization tool Lorikeet that was re-engineered to improve annotation of C-mannosylation and O-fucosylation and is included with the Trans-Proteomics Pipeline, the free and open-source proteomics data analysis suite used to perform the analyses described here. The difference between the observed experimental precursor mass of 1329.0304 m/z and matched peptide mass of 1174.9770 m/z indicates that the PSM was obtained by the database searching program Comet allowing neutral loss of 308.11 Da (loss of O-Fuc-Glc) in the MS^2. The dominant peak (yellow M++) at 1174.98 m/z is the peptide that has lost O-Fuc-Glc but retains C-Man. Neutral loss of 120.04 Da from cross-ring cleavage of C-Man is indicated as "−120", and neutral loss of water is indicated as "o". The fragment spectra positively localized the C-mannose to the C-terminal Trp.

database search engine if the neutral loss of the glycan could be accounted for. To this end, we modified the open-source MS/MS sequence database search tool Comet[17] by adding a new "mass_offsets" parameter that allows the user to specify neutral loss masses to be variably subtracted from precursor masses of fragment spectra. For example, in the case of the O-glucosylfucosylated *Pf* CSP peptide described above, the parent ion had a mass of 2900.32 Da ($[M + 2H]^{2+}$ = 1451.17 m/z). By

DOI: 10.1021/acs.jproteome.8b00638
J. Proteome Res. XXXX, XXX, XXX−XXX

directing Comet to variably subtract 308.11 Da (equal to neutral loss of O-Fuc-Glc) from the precursor mass, the new predicted precursor mass of the peptide would be 2592.21 Da ($[M + 2H]^{2+}$ = 1297.11 m/z), matching the mass of the peptide identified by the unmodified peptide fragment ions in the MS^2 spectrum for this species. We re-engineered the TPP to accommodate PSMs identified using the new "mass_offset" parameter, and we further re-engineered the automated spectral annotation tool Lorikeet to automatically annotate neutral loss of 120 Da from cross-ring cleavage of C-man, a PTM often found in close proximity to O-Fuc on TSRs. The data analysis workflow is summarized in Figure 3.

To demonstrate this modified search approach, the recombinant *Pf*TRAP tryptic digest was again analyzed by a DDA method, this time employing an inclusion list that targeted the glycosite-containing peptide at multiple masses corresponding to modification with the various combinations of two, one, or no C-Man and an O-Fuc-Glc, O-Fuc, or no O-linked glycan. MS^2 spectra were searched using Comet with allowed mass offsets of 0, 146.06, and 308.11 Da (no offset, loss of O-Fuc, and loss of O-Fuc-Glc, respectively), as well as variable modification of +162.05 @ Trp for C-Man. The results of the Comet search were analyzed with the TPP and peptide spectra matched with a mass offset were identified by sorting according to the "massdiff" field in the TPP's PepXML Viewer. This value was calculated for each PSM as the difference between the predicted mass of the identified peptide and the observed mass of the precursor ion (Table S2). The glycosite-containing peptide was detected with multiple combinations of glycans (Table 1). It was imperative to cross reference these results with the extracted ion chromatograms of the identified species, as loss of glycans due to in-source fragmentation gave rise to spurious identifications (Figure 4). For example, each of the two positional isoforms of the singly C-mannosylated peptide ($[M + 2H]^{2+}$ = 1174.98 m/z) coeluted with its higher-abundance, higher-mass counterpart ($[M + 2H]^{2+}$ = 1329.03 m/z), which was identified with a single C-Man as well as a mass offset of 308.11 Da (O-Fuc-Glc). This observation is consistent with the 1174.98 m/z species at that retention time arising from neutral loss of the labile O-linked glycan due to in-source CID prior to MS^2. Once properly accounted for, this in-source CID actually provided an additional line of evidence for differentiating between C-linked and O-linked glycans that are isobaric and otherwise indistinguishable by MS. Oxidation of tryptophan (likely introduced artificially during sample preparation) produced additional isobaric peptide species. For example, the peptide with a single oxidized Trp and an O-Fuc-Glc had the same mass as the peptide modified with two C-Man modifications, but these species were easily distinguished by MS^2 and retention times (Table S2, Figure S3). The one case of indistinguishable isobaric species was observed with the most abundant glycoform, i.e., the peptide modified with a glucosylfucose. This glycoform was identified from the dominant LC peak at 59.7 min, the intensity of which was more than 7-fold greater than any other species. The glycopeptide was identified from multiple high-quality spectra, as was the coeluting peptide that had lost the disaccharide due to in-source fragmentation. Interestingly, another coeluting pair of peptides with identical respective masses was identified at 57.8 min. The MS^2 spectra for these isobars were indistinguishable from their more abundant, later-eluting counterparts (Figure S4). The source of this peak splitting is unclear, as no other combination of expected glycans or oxidation states would give rise to the observed combination of precursor mass, MS^2 spectra, and coeluting peak pairs. If the dual peaks arose due to a chromatographic artifact or even conformational isomer of the same peptide, similar peak pairs would not likely be detectable for the other glycoforms identified in the sample - the minor peak intensity was only 2% of the dominant peak intensity, so additional minor peaks arising from the other, less-abundant species would likely be below the detection limit. While the major glycoforms were identified from multiple high-quality PSMs, there was one incorrect identification from a sparse, low-intensity PSM that was nonetheless assigned a high score. This PSM identified the doubly mannosylated peptide with neutral loss of O-Fuc, but manual inspection of the MS^2 spectrum revealed that it was actually an isobaric glycoform, the singly mannosylated peptide with neutral loss of O-Fuc-Glc. This correct identity could be determined from the mass of the dominant neutral loss precursor peak as well as the fact that multiple other MS^2 from the same m/z and retention time correctly identified the singly mannosylated glycoform (Figure S5). The above examples demonstrate that, as with any MS/MS sequence database search tool, the resulting list of PSMs generated using our approach should always be confirmed by manual inspection of the data, especially when considering isobaric glycoforms. However, even the many isobaric glycoforms considered above could be confidently distinguished given sufficiently high-quality spectra and chromatography.

Because our database searching strategy is amenable to standard instrumental approaches (i.e., untargeted DDA employing CID or HCD) and does not require identification of low-mass oxonium ions or alternative fragmentation techniques,[33,34] existing mass spectrometry data sets can be researched for evidence of O-fucosylation, even if the MS^2 data were collected at low resolution (e.g., using an ion trap). To demonstrate this, we reanalyzed mass spectra from previously published proteomic analyses of *P. falciparum*, *P. vivax*, and *P. yoelii* salivary gland sporozoites. Proteomic analyses of salivary gland sporozoites have recently shown that the TSR domains of CSP and TRAP are glycosylated in vivo in the human-infective *Plasmodium* species *P. falciparum*[21] and *P. vivax*.[22] Subsequent to those studies, it was demonstrated that the O-linked deoxyhexose observed in *P. falciparum* sporozoites is, in fact, O-Fuc, and the identity of the *O*-fucosyltransferase was confirmed as POFUT2.[10] It has also been confirmed that *P. falciparum* possesses a version of the Dpy19 C-mannosyltransferase capable of modifying the TSR domain of *Pf*TRAP with C-Man.[35] Because of the confounding effect of labile O-linked glycans on MS/MS sequence database searching, the evidence for in vivo O-fucosylation of *Plasmodium* parasites in the studies mentioned above was obtained by manually interpreting tandem mass spectra of peptides matching the predicted mass of glycosylated peptides. Here, we have reanalyzed those data sets with our new automated database searching approach and validated those original findings with dozens of high-quality PSMs. Furthermore, this reanalysis provides the first reported evidence of TSR glycosylation in the rodent-infective malaria parasite *P. yoelii*.

All data sets were acquired on an LTQ-Orbitrap with high resolution MS^1 and low resolution MS^2. The *P. falciparum* data, which includes a global proteome analysis[20] and a biotinylated surface-enriched sample,[21] were part of the original evidence presented for O-fucosylation of TSRs in plasmodium,[21] and the *P. vivax* data set subsequently demonstrated TSR glycosylation in that species as well.[22] In those analyses, select MS^2 spectra

DOI: 10.1021/acs.jproteome.8b00638
J. Proteome Res. XXXX, XXX, XXX−XXX

from parent ions matching predicted masses were manually annotated to provide evidence of O-fucosylation and C-mannosylation. Reanalyzing the mass spectral data with our automated database searching approach confirmed the original findings and provided extensive additional evidence for these modifications from the same data sets (Figure 5, Table 2, Table S3, Figure S6–S14). Reanalyzing the *P. yoelii* data resulted in the first reported evidence of glycosylation of TSRs in that species (Figure 4, Figure S12–14, Table S3). CSP and TRAP in *P. yoelii* were O-fucosylated at the peptides bearing the conserved glycosites in the TSR domain. Both CSP and TRAP were only detected with O-Fuc; no evidence was observed for the glycosite-bearing peptides from either protein in unmodified form or modified with a Fuc-Hex disaccharide (Figure S12, Figure S13). Both proteins were also C-mannosylated in the *P. yoelii* sample. As in *P. falciparum*, 100% of the TRAP observed in the *P. yoelii* sample was C-mannosylated, and always at the second Trp residue of the WXXW motif, in marked contrast to the recombinant *Pf* TRAP expressed in mammalian cells, which exhibited all four possible C-Man site occupancy combinations (Table 1), and *P. vivax*, in which C-Man was not observed at all on TRAP. The presence or absence and site localization of C-Man could be made unambiguously from intact C-Man on fragment ions. The majority of CSP was C-mannosylated (~96% based on LC peak height), a modification not observed on CSP in *P. falciparum* or *P. vivax*. This observation confirms that, at least in *P. yoelii*, *Plasmodium* Dpy19 is capable of mannosylating the WXXC motif even in the absence of a preceding WXXW motif.[35]

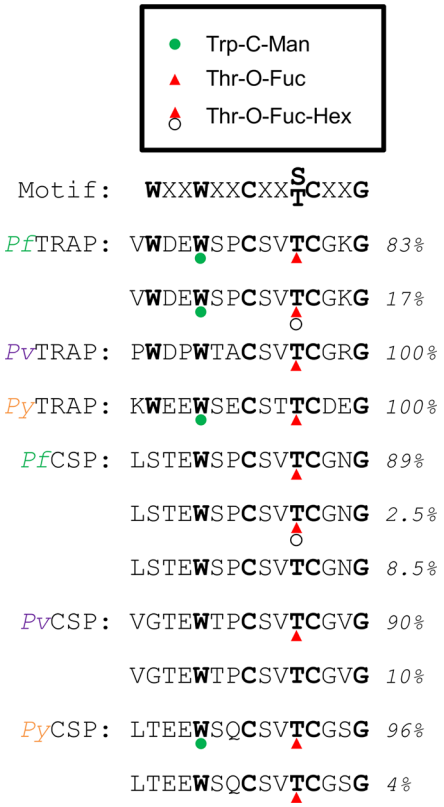


Figure 5. In vivo glycosylation of *Plasmodium* TSR domains identified from reanalysis of published data. Summarized here is the site occupancy of C-mannosylation and O-fucosylation of thrombospondin type 1 repeat (TSR) domains in circumsporozoite protein (CSP) and thrombospondin-related anonymous protein (TRAP) in *Plasmodium* salivary gland sporozoites as determined by automated database searching of previously published proteomics data. Detected glycans are indicated with triangles and circles as described in the legend. The hexose of the Fuc-Hex disaccharide is indicated as a generic hexose because neither the identity of the sugar nor its glycosyltransferase has been identified in *Plasmodium*. The approximate relative proportion of glycoforms in the analyzed samples is given as estimated by LC peak height (see Figure S6–S13). Glycosylation of CSP and TRAP has been reported in salivary gland sporozoites of *P. falciparum* (Pf)[21] and *P. vivax* (Pv).[22] Our reanalysis of published data from *P. yoelii* (Py) salivary gland sporozoites[20] has provided the first evidence of TSR glycosylation in that species.

For all of the *Plasmodium* salivary gland sporozoite data sets analyzed, the FDR calculated by PeptideProphet generally agreed with the decoy-estimated FDR; at a probability corresponding to a model-estimated FDR of 1% (which should equal a decoy-estimated FDR of approximately 0.5% after excluding the known decoys, assuming half of the incorrect answers are decoys), the decoy-estimated FDRs for the data sets ranged from 0.26% to 0.56%. Among the PSMs identified above this cutoff with a putative neutral loss of 146.06 or 308.11, none were decoy proteins. However, other real protein entries were identified that were likely false positives (Table 2, Table S3). The most common of these, appearing in four of the five data sets, were semitryptic fragments of a 39-residue human keratin peptide found in the contaminant database. The N-terminus of the fully tryptic peptide began with the sequence GSY. The mass of Gly-Ser is 144.05 Da and the mass of Gly-Ser-Tyr is 307.11. Combined with a mass error of two or one protons, respectively, arising from the incorrect isotope peak being identified as the monoisotopic peak of the precursor peptide ion, these mass errors were within the matching tolerance of the allowed neutral losses of 146.06 and 308.11 Da corresponding to O-Fuc and O-Fuc-Glc, respectively (Figure S15). Considering only fully tryptic peptides eliminated these false positives and nearly all others (Table 2), and further considering only the highest quality PSM (those with PeptideProphet probabilities above 0.99 and Expect scores below 10^{-5}) eliminated the remaining spurious fully tryptic peptides. The above observations underscore the importance of manual validation of results obtained from this database searching approach. In the case of the samples analyzed here, the vast majority of PSMs identified with putative neutral loss of an O-linked glycan were peptides containing the expected glycosylation sites, and manual inspection of the MS^2 spectra confirmed the quality of the matches.

Taken together, these data demonstrate a marked variety in glycan site occupancy on conserved TSR domains within and across *Plasmodium* species, inviting the question of what roles these glycans play in TSR proteins and whether these roles are protein- and species-specific. Importantly, these data also begin to suggest a recognition motif for the *Plasmodium* C-mannosyltransferase. Further work will be required to confirm this motif and whether the recognition sequence is conserved across *Plasmodium* species. Also of interest is the variable detection of an additional hexose (presumably glucose) modifying O-Fuc. In the data sets analyzed here, the large majority of CSP and TRAP in salivary gland sporozoites of all species was O-fucosylated, but the disaccharide was only seen in *P. falciparum* and only on a small percentage of the CSP and TRAP present (Figure 5). Further work will be required to determine whether the failure to detect the disaccharide in the *P. vivax* and *P. yoelii* data sets is due to instrumental detection limit or the true absence of the glycan. In other organisms this

DOI: 10.1021/acs.jproteome.8b00638
J. Proteome Res. XXXX, XXX, XXX−XXX

Table 2. Evidence for in Vivo Glycosylation of CSP and TRAP[a]

sample[b]	total PSMs[c]	CSP PSMs[d]	TRAP PSMs[e]	decoy PSMs[f]	nondecoy false positives[g]	fully tryptic nondecoy false positives[h]
P. vivax VK210[22]	352	259	83	0	10	1
P. vivax VK247[22]	250	171	33	0	46	0
P. falciparum global proteome[20]	57	54	0	0	3	1
P. falciparum surface-enriched[21]	12	0	4	0	8	1
P. yoelii[20]	26	18	7	0	1	0

[a]Previously published proteomics data from analyses of *Plasmodium* salivary gland sporozoites were re-analyzed using our automated database searching approach for identification of neutral loss of O-linked fucose or glucosylfucose. Peptide spectrum matches (PSMs) in this table were identified with evidence for neutral loss of −146.06 (equal to deoxyhexose, i.e., fucose) or −308.11 (equal to hexosyldeoxyhexose, i.e., glucosylfucose). [b]Sources of data are cited. VK210 and VK247 refer to different *P. vivax* haplotypes. No evidence for O-fucosylation of TRAP was observed in the *P. falciparum* global proteome data set because the precursor ion, though highly abundant, was never selected for fragmentation because the instrument software was unable to assign a charge state to the isotope envelope, and the ion was accordingly excluded from monoisotopic precursor selection. A gel fraction containing TRAP from a surface-enriched *P. falciparum* sample was analyzed for spectral evidence of O-fucosylation of TRAP. [c]Total number of peptide spectrum matches (PSMs) identified with a neutral loss putatively corresponding to an O-linked glycan. [d]Total number of PSMs identifying the glycosite-containing peptide of CSP with evidence for O-fucosylation. [e]Total number of PSMs identifying the glycosite-containing peptide of TRAP with evidence for O-fucosylation. [f]The number of PSMs identified from decoy proteins. [g]The number PSMs identifying likely incorrect nondecoy proteins. [h]The number PSMs identifying fully tryptic peptides mapping to likely incorrect nondecoy proteins.

glycan extension is performed by β-1,3-glucosyltransferase (B3GLCT). Although there is a *P. falciparum* gene with some sequence homology to B3GLCT[20] (parasite-infected erythrocyte surface protein (PIESP1); PF3D7_0310400), this function has not been verified, and the biological importance of this modification for the parasite is unknown. Finally, it is important to note that the glycosite occupancy observed for *Plasmodium* proteins in vivo differed considerably from that observed for the recombinant proteins expressed in mammalian cells (compare Table 1 with Figure 5). The TSR domains of both *Pf* CSP[36] and *Pf* TRAP[37] are part of subunit malaria vaccine candidates.[38] Since glycosylation effectively changes antibody recognition epitopes,[39] it is important to be able to match the glycosylation patterns of the recombinant protein subunit to those observed in vivo.[15,31,32,40] The methods we have described here enable simple analysis of TSR glycosylation using standard proteomics techniques and open-source analysis software.

In conclusion, the open-source MS/MS sequence database search tool Comet was re-engineered to enable automatic detection of gas-phase labile O-linked mono- and disaccharides. The analysis tools of the TPP were re-engineered to accommodate the new information obtained from this search strategy, enabling assignment of probabilities and FDRs to identified glycosylated peptides as well as visualization of the MS^2 spectra, including neutral losses from cross-ring cleavage of C-Man. We used this search strategy to reanalyze published proteomics data sets, confirming manually annotated spectra that had identified O-fucosylated peptides in *P. falciparum* and *P. vivax*, as well as identifying previously unreported glycosylation in *P. yoelii*. While the workflow described here was developed in order to identify a very specific type of glycosylation (i.e., O-fucosylation of TSRs), the new "mass_offsets" parameter in Comet can be set by the user to search for any predicted neutral loss. Combined with the support of the TPP analysis suite, this workflow can in theory be used to detect peptides exhibiting neutral loss of any moiety that behaves in a similar fashion to O-Fuc, e.g., O-GlcNAc. These tools will enable researchers to identify these and other gas-phase labile PTMs using standard proteomics techniques, as well as to reanalyze existing data sets for evidence of glycosylation that was not previously searched for.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.8b00638.

Figures S1−S15 (PDF)
Tables S1−S3 (XLSX)

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: kristian.swearingen@systemsbiology.org.
*E-mail: robert.moritz@systemsbiology.org.

ORCID

Kristian E. Swearingen: 0000-0002-6756-4471
Jimmy K. Eng: 0000-0001-6352-6737
Timothy A. Springer: 0000-0001-6627-2904
Luis Mendoza: 0000-0003-0128-8643
Eric W. Deutsch: 0000-0001-8732-0928
Robert L. Moritz: 0000-0002-3216-9447

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Research reported in this publication was supported by the National Institutes of Health National Institute of Allergy and Infectious Disease under award number K25AI119229 (KES), by the University of Washington's Proteomics Resource under award number UWPR95794 (JKE), by the National Institutes of Health National Cancer Institute under award number R01CA31798 (TAS), by the National Institutes of Health National Institute of General Medical Sciences under award numbers R24GM127667 (EWD) and R01GM087221 (RLM), by the National Institutes of Health National Institute of Biomedical Imaging and Bioengineering under award number U54EB020406 (EWD), and by the National Science Foundation under award number 0923536 (RLM, KES). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, University of Washington's Proteomics Resource, or the National Science Foundation. The funders had no role in

DOI: 10.1021/acs.jproteome.8b00638
J. Proteome Res. XXXX, XXX, XXX−XXX

study design, data collection and analysis, decision to publish, or preparation of the manuscript.

■ ABBREVIATIONS

TSR, thrombospondin type 1 repeat; POFUT2, protein *O*-fucosyltransferase 2; B3GLCT, β-1,3-glucosyltransferase.

■ REFERENCES

(1) Tucker, R. P. The thrombospondin type 1 repeat superfamily. *Int. J. Biochem. Cell Biol.* **2004**, *36* (6), 969−74.

(2) Luo, Y.; Koles, K.; Vorndam, W.; Haltiwanger, R. S.; Panin, V. M. Protein O-fucosyltransferase 2 adds O-fucose to thrombospondin type 1 repeats. *J. Biol. Chem.* **2006**, *281* (14), 9393−9.

(3) Hofsteenge, J.; Huwiler, K. G.; Macek, B.; Hess, D.; Lawler, J.; Mosher, D. F.; Peter-Katalinic, J. C-mannosylation and O-fucosylation of the thrombospondin type 1 module. *J. Biol. Chem.* **2001**, *276* (9), 6485−98.

(4) Sato, T.; Sato, M.; Kiyohara, K.; Sogabe, M.; Shikanai, T.; Kikuchi, N.; Togayachi, A.; Ishida, H.; Ito, H.; Kameyama, A.; Gotoh, M.; Narimatsu, H. Molecular cloning and characterization of a novel human beta1,3-glucosyltransferase, which is localized at the endoplasmic reticulum and glucosylates O-linked fucosylglycan on thrombospondin type 1 repeat domain. *Glycobiology* **2006**, *16* (12), 1194−206.

(5) Kozma, K.; Keusch, J. J.; Hegemann, B.; Luther, K. B.; Klein, D.; Hess, D.; Haltiwanger, R. S.; Hofsteenge, J. Identification and characterization of abeta1,3-glucosyltransferase that synthesizes the Glc-beta1,3-Fuc disaccharide on thrombospondin type 1 repeats. *J. Biol. Chem.* **2006**, *281* (48), 36742−51.

(6) Hofsteenge, J.; Muller, D. R.; de Beer, T.; Loffler, A.; Richter, W. J.; Vliegenthart, J. F. New type of linkage between a carbohydrate and a protein: C-glycosylation of a specific tryptophan residue in human RNase Us. *Biochemistry* **1994**, *33* (46), 13524−30.

(7) Julenius, K. NetCGlyc 1.0: prediction of mammalian C-mannosylation sites. *Glycobiology* **2007**, *17* (8), 868−76.

(8) Buettner, F. F.; Ashikov, A.; Tiemann, B.; Lehle, L.; Bakker, H. C. elegans DPY-19 is a C-mannosyltransferase glycosylating thrombospondin repeats. *Mol. Cell* **2013**, *50* (2), 295−302.

(9) Shcherbakova, A.; Tiemann, B.; Buettner, F. F.; Bakker, H. Distinct C-mannosylation of netrin receptor thrombospondin type 1 repeats by mammalian DPY19L1 and DPY19L3. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114* (10), 2574−2579.

(10) Lopaticki, S.; Yang, A. S. P.; John, A.; Scott, N. E.; Lingford, J. P.; O'Neill, M. T.; Erickson, S. M.; McKenzie, N. C.; Jennison, C.; Whitehead, L. W.; Douglas, D. N.; Kneteman, N. M.; Goddard-Borger, E. D.; Boddey, J. A. Protein O-fucosylation in Plasmodium falciparum ensures efficient infection of mosquito and vertebrate hosts. *Nat. Commun.* **2017**, *8* (1), 561.

(11) Bandini, G.; Haserick, J. R.; Motari, E.; Ouologuem, D. T.; Lourido, S.; Roos, D. S.; Costello, C. E.; Robbins, P. W.; Samuelson, J. O-fucosylated glycoproteins form assemblies in close proximity to the nuclear pore complexes of Toxoplasma gondii. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113* (41), 11567−11572.

(12) Sorvillo, N.; Kaijen, P. H.; Matsumoto, M.; Fujimura, Y.; van der Zwaan, C.; Verbij, F. C.; Pos, W.; Fijnheer, R.; Voorberg, J.; Meijer, A. B. Identification of N-linked glycosylation and putative O-fucosylation, C-mannosylation sites in plasma derived ADAMTS13. *J. Thromb. Haemostasis* **2014**, *12* (5), 670−9.

(13) Swaney, D. L.; McAlister, G. C.; Coon, J. J. Decision tree-driven tandem mass spectrometry for shotgun proteomics. *Nat. Methods* **2008**, *5* (11), 959−64.

(14) Kakuda, S.; Haltiwanger, R. S. Analyzing the posttranslational modification status of Notch using mass spectrometry. *Methods Mol. Biol.* **2014**, *1187*, 209−21.

(15) Doud, M. B.; Koksal, A. C.; Mi, L. Z.; Song, G.; Lu, C.; Springer, T. A. Unexpected fold in the circumsporozoite protein target of malaria vaccines. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (20), 7817−22.

(16) Carbonetti, S.; Oliver, B. G.; Vigdorovich, V.; Dambrauskas, N.; Sack, B.; Bergl, E.; Kappe, S. H. I.; Sather, D. N. A method for the isolation and characterization of functional murine monoclonal antibodies by single B cell cloning. *J. Immunol. Methods* **2017**, *448*, 66−73.

(17) Eng, J. K.; Jahan, T. A.; Hoopmann, M. R. Comet: an open-source MS/MS sequence database search tool. *Proteomics* **2013**, *13* (1), 22−4.

(18) Deutsch, E. W.; Mendoza, L.; Shteynberg, D.; Slagel, J.; Sun, Z.; Moritz, R. L. Trans-Proteomic Pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. *Proteomics: Clin. Appl.* **2015**, *9* (7−8), 745−54.

(19) Kusebauch, U.; Deutsch, E. W.; Campbell, D. S.; Sun, Z.; Farrah, T.; Moritz, R. L. Using PeptideAtlas, SRMAtlas, and PASSEL: Comprehensive Resources for Discovery and Targeted Proteomics. *Curr. Protoc Bioinformatics* **2014**, *46*, 13.25.1−28.

(20) Lindner, S. E.; Swearingen, K. E.; Harupa, A.; Vaughan, A. M.; Sinnis, P.; Moritz, R. L.; Kappe, S. H. Total and putative surface proteomics of malaria parasite salivary gland sporozoites. *Mol. Cell. Proteomics* **2013**, *12* (5), 1127−43.

(21) Swearingen, K. E.; Lindner, S. E.; Shi, L.; Shears, M. J.; Harupa, A.; Hopp, C. S.; Vaughan, A. M.; Springer, T. A.; Moritz, R. L.; Kappe, S. H.; Sinnis, P. Interrogating the Plasmodium Sporozoite Surface: Identification of Surface-Exposed Proteins and Demonstration of Glycosylation on CSP and TRAP by Mass Spectrometry-Based Proteomics. *PLoS Pathog.* **2016**, *12* (4), No. e1005606.

(22) Swearingen, K. E.; Lindner, S. E.; Flannery, E. L.; Vaughan, A. M.; Morrison, R. D.; Patrapuvich, R.; Koepfli, C.; Muller, I.; Jex, A.; Moritz, R. L.; Kappe, S. H. I.; Sattabongkot, J.; Mikolajczak, S. A. Proteogenomic analysis of the total and surface-exposed proteomes of Plasmodium vivax salivary gland sporozoites. *PLoS Neglected Trop. Dis.* **2017**, *11* (7), No. e0005791.

(23) Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **2008**, *24* (21), 2534−6.

(24) Gardner, M. J.; Hall, N.; Fung, E.; White, O.; Berriman, M.; Hyman, R. W.; Carlton, J. M.; Pain, A.; Nelson, K. E.; Bowman, S.; Paulsen, I. T.; James, K.; Eisen, J. A.; Rutherford, K.; Salzberg, S. L.; Craig, A.; Kyes, S.; Chan, M. S.; Nene, V.; Shallom, S. J.; Suh, B.; Peterson, J.; Angiuoli, S.; Pertea, M.; Allen, J.; Selengut, J.; Haft, D.; Mather, M. W.; Vaidya, A. B.; Martin, D. M.; Fairlamb, A. H.; Fraunholz, M. J.; Roos, D. S.; Ralph, S. A.; McFadden, G. I.; Cummings, L. M.; Subramanian, G. M.; Mungall, C.; Venter, J. C.; Carucci, D. J.; Hoffman, S. L.; Newbold, C.; Davis, R. W.; Fraser, C. M.; Barrell, B. Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* **2002**, *419* (6906), 498−511.

(25) Aurrecoechea, C.; Brestelli, J.; Brunk, B. P.; Dommer, J.; Fischer, S.; Gajria, B.; Gao, X.; Gingle, A.; Grant, G.; Harb, O. S.; Heiges, M.; Innamorato, F.; Iodice, J.; Kissinger, J. C.; Kraemer, E.; Li, W.; Miller, J. A.; Nayak, V.; Pennington, C.; Pinney, D. F.; Roos, D. S.; Ross, C.; Stoeckert, C. J., Jr.; Treatman, C.; Wang, H. PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids Res.* **2009**, *37*, D539−43.

(26) Otto, T. D.; Bohme, U.; Jackson, A. P.; Hunt, M.; Franke-Fayard, B.; Hoeijmakers, W. A.; Religa, A. A.; Robertson, L.; Sanders, M.; Ogun, S. A.; Cunningham, D.; Erhart, A.; Billker, O.; Khan, S. M.; Stunnenberg, H. G.; Langhorne, J.; Holder, A. A.; Waters, A. P.; Newbold, C. I.; Pain, A.; Berriman, M.; Janse, C. J. A comprehensive evaluation of rodent malaria parasite genomes and gene expression. *BMC Biol.* **2014**, *12*, 86.

(27) Auburn, S.; Bohme, U.; Steinbiss, S.; Trimarsanto, H.; Hostetler, J.; Sanders, M.; Gao, Q.; Nosten, F.; Newbold, C. I.; Berriman, M.; Price, R. N.; Otto, T. D. A new Plasmodium vivax reference sequence with improved assembly of the subtelomeres reveals an abundance of pir genes. *Wellcome Open Res.* **2016**, *1*, 4.

(28) Jiang, X.; Peery, A.; Hall, A. B.; Sharma, A.; Chen, X. G.; Waterhouse, R. M.; Komissarov, A.; Riehle, M. M.; Shouche, Y.; Sharakhova, M. V.; Lawson, D.; Pakpour, N.; Arensburger, P.; Davidson, V. L.; Eiglmeier, K.; Emrich, S.; George, P.; Kennedy, R. C.; Mane, S. P.; Maslen, G.; Oringanje, C.; Qi, Y.; Settlage, R.; Tojo, M.; Tubio, J. M.; Unger, M. F.; Wang, B.; Vernick, K. D.; Ribeiro, J. M.;

DOI: 10.1021/acs.jproteome.8b00638
J. Proteome Res. XXXX, XXX, XXX−XXX

James, A. A.; Michel, K.; Riehle, M. A.; Luckhart, S.; Sharakhov, I. V.; Tu, Z. Genome analysis of a major urban malaria vector mosquito, Anopheles stephensi. *Genome Biol.* **2014**, *15* (9), 459.
(29) Giraldo-Calderon, G. I.; Emrich, S. J.; MacCallum, R. M.; Maslen, G.; Dialynas, E.; Topalis, P.; Ho, N.; Gesing, S.; VectorBase, C.; Madey, G.; Collins, F. H.; Lawson, D. VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. *Nucleic Acids Res.* **2015**, *43*, D707−13.
(30) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **2002**, *74* (20), 5383−92.
(31) Song, G.; Koksal, A. C.; Lu, C.; Springer, T. A. Shape change in the receptor for gliding motility in Plasmodium sporozoites. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (52), 21420−5.
(32) Song, G.; Springer, T. A. Structures of the Toxoplasma gliding motility adhesin. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111* (13), 4862−7.
(33) Hahne, H.; Kuster, B. A novel two-stage tandem mass spectrometry approach and scoring scheme for the identification of O-GlcNAc modified peptides. *J. Am. Soc. Mass Spectrom.* **2011**, *22* (5), 931−42.
(34) Hahne, H.; Gholami, A. M.; Kuster, B. Discovery of O-GlcNAc-modified proteins in published large-scale proteome data. *Mol. Cell. Proteomics* **2012**, *11* (10), 843−50.
(35) Hoppe, C. M.; Albuquerque-Wendt, A.; Bandini, G.; Leon, D. R.; Shcherbakova, A.; Buettner, F. F. R.; Izquierdo, L.; Costello, C. E.; Bakker, H.; Routier, F. H. Apicomplexan C-Mannosyltransferases Modify Thrombospondin Type I-containing Adhesins of the TRAP Family. *Glycobiology* **2018**, *28* (5), 333−343.
(36) Olotu, A.; Fegan, G.; Wambua, J.; Nyangweso, G.; Leach, A.; Lievens, M.; Kaslow, D. C.; Njuguna, P.; Marsh, K.; Bejon, P. Seven-Year Efficacy of RTS,S/AS01 Malaria Vaccine among Young African Children. *N. Engl. J. Med.* **2016**, *374* (26), 2519−29.
(37) Moorthy, V. S.; Imoukhuede, E. B.; Milligan, P.; Bojang, K.; Keating, S.; Kaye, P.; Pinder, M.; Gilbert, S. C.; Walraven, G.; Greenwood, B. M.; Hill, A. S. A randomised, double-blind, controlled vaccine efficacy trial of DNA/MVA ME-TRAP against malaria infection in Gambian adults. *PLoS Med.* **2004**, *1* (2), No. e33.
(38) Kester, K. E.; Gray Heppner, D., Jr.; Moris, P.; Ofori-Anyinam, O.; Krzych, U.; Tornieporth, N.; McKinney, D.; Delchambre, M.; Ockenhouse, C. F.; Voss, G.; Holland, C.; Beckey, J. P.; Ballou, W. R.; Cohen, J.; Rts, S. T. G. Sequential Phase 1 and Phase 2 randomized, controlled trials of the safety, immunogenicity and efficacy of combined pre-erythrocytic vaccine antigens RTS,S and TRAP formulated with AS02 Adjuvant System in healthy, malaria naive adults. *Vaccine* **2014**, *32* (49), 6683−91.
(39) Luca, V. C.; Jude, K. M.; Pierce, N. W.; Nachury, M. V.; Fischer, S.; Garcia, K. C. Structural biology. Structural basis for Notch1 engagement of Delta-like 4. *Science* **2015**, *347* (6224), 847−53.
(40) Goddard-Borger, E. D.; Boddey, J. A. Implications of Plasmodium glycosylation on vaccine efficacy and design. *Future Microbiol.* **2018**, *13*, 609−612.

DOI: 10.1021/acs.jproteome.8b00638
J. Proteome Res. XXXX, XXX, XXX−XXX

SUPPORTING INFORMATION

A tandem mass spectrometry sequence database search method for identification of O-fucosylated proteins by mass spectrometry.

Kristian E. Swearingen[1], Jimmy K. Eng[2], David Shteynberg[1], Vladimir Vigdorovich[3], Timothy A. Springer[4], Luis Mendoza[1], Noah D. Sather[3], Eric W. Deutsch[1], Stefan H. I. Kappe[3,] , Robert L. Moritz[1*]*

[1] Institute for Systems Biology, Seattle, WA, USA

[2] Proteomics Resource, University of Washington, Seattle, WA, USA

[3] Center for Global Infectious Disease Research, Seattle Children's Research Institute, Seattle, WA, USA

[4]Harvard Medical School and Children's Hospital, Boston, MA, USA

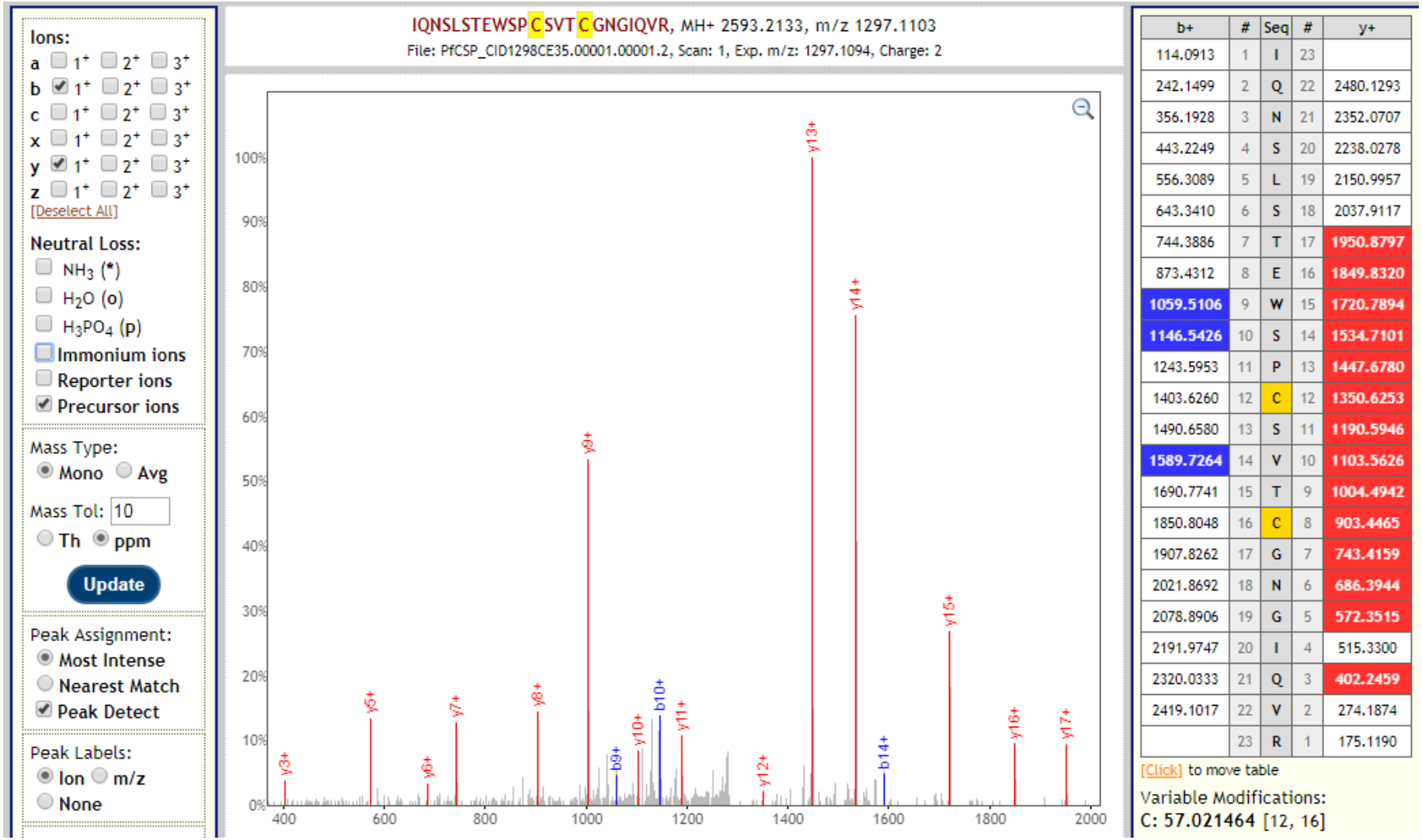
*To whom correspondence should be addressed.

kristian.swearingen@systemsbiology.org

robert.moritz@systemsbiology.org

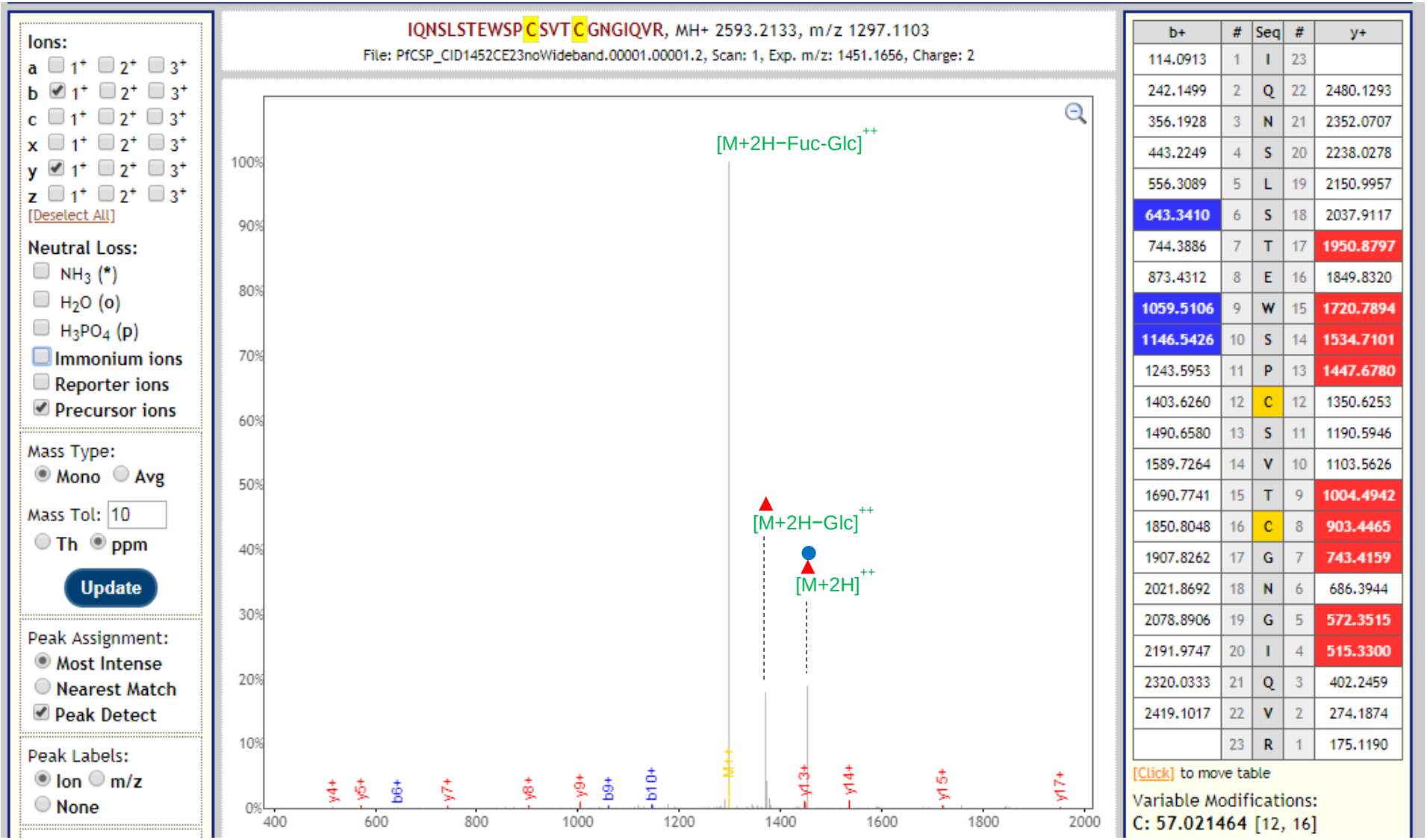
TABLE OF CONTENTS

Figure S1. MS^n confirming the identity of fragment spectra arising from neutral loss of O-linked glycans. A tryptic digest of recombinant *P. falciparum* circumsporozoite protein (CSP) TSR domain was infused into an LTQ-Orbitrap Elite by nanospray. Precursor ions for the unmodified or O-glucosylfucosylated peptide were isolated and fragmented. All spectra are collision-induced dissociation (CID) collected at high resolution in the Orbitrap. CE = normalized collision energy. Annotated fragment spectra were visualized using a development version of Lorikeet as implemented in the Trans-Proteomic Pipeline.



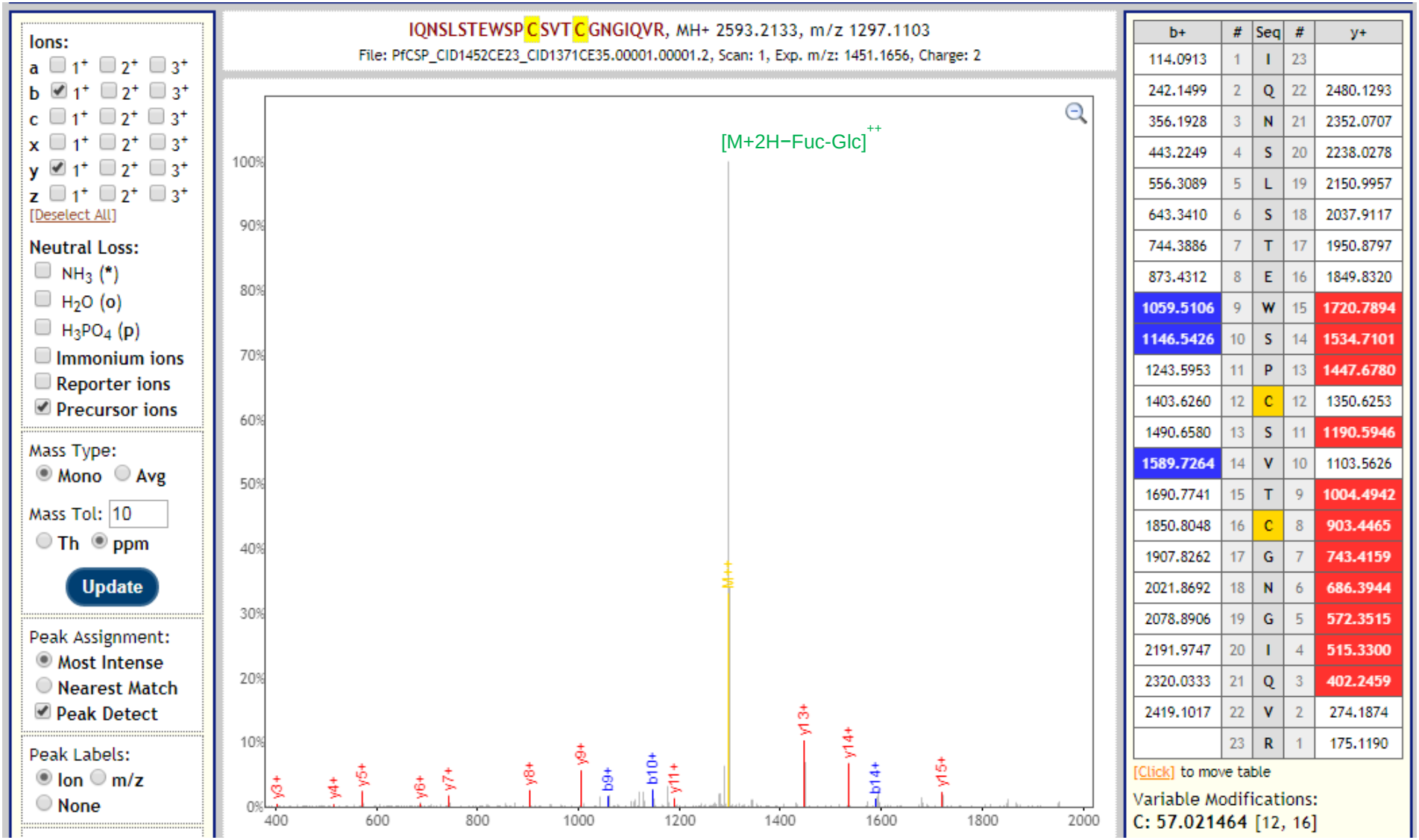
b+	#	Seq	#	y+
114.0913	1	I	23	
242.1499	2	Q	22	2480.1293
356.1928	3	N	21	2352.0707
443.2249	4	S	20	2238.0278
556.3089	5	L	19	2150.9957
643.3410	6	S	18	2037.9117
744.3886	7	T	17	1950.8797
873.4312	8	E	16	1849.8320
1059.5106	9	W	15	1720.7894
1146.5426	10	S	14	1534.7101
1243.5953	11	P	13	1447.6780
1403.6260	12	C	12	1350.6253
1490.6580	13	S	11	1190.5946
1589.7264	14	V	10	1103.5626
1690.7741	15	T	9	1004.4942
1850.8048	16	C	8	903.4465
1907.8262	17	G	7	743.4159
2021.8692	18	N	6	686.3944
2078.8906	19	G	5	572.3515
2191.9747	20	I	4	515.3300
2320.0333	21	Q	3	402.2459
2419.1017	22	V	2	274.1874
	23	R	1	175.1190

S1a. CID of the unmodified peptide, $[M+2H]^{++}$ = 1297 *m/z*, CE = 35 %. Note that the dominant peaks are y_{13} and y_{14}, the y-ion peaks containing the Pro residue. The y_{13} and y_{14} peaks are the only species visible in the spectra in Figure 1 still bearing the O-linked disaccharide.



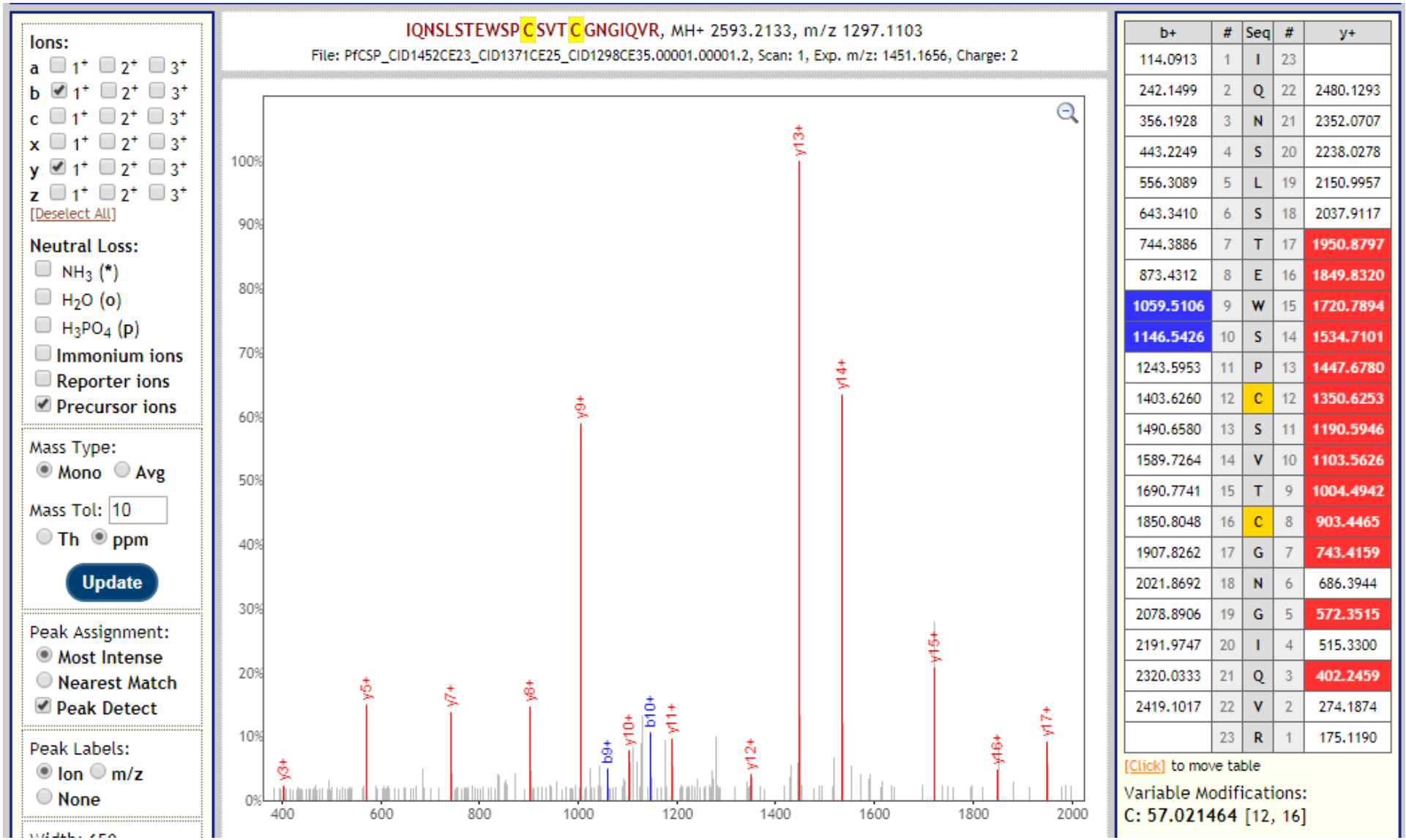
b+	#	Seq	#	y+
114.0913	1	I	23	
242.1499	2	Q	22	2480.1293
356.1928	3	N	21	2352.0707
443.2249	4	S	20	2238.0278
556.3089	5	L	19	2150.9957
643.3410	6	S	18	2037.9117
744.3886	7	T	17	1950.8797
873.4312	8	E	16	1849.8320
1059.5106	9	W	15	1720.7894
1146.5426	10	S	14	1534.7101
1243.5953	11	P	13	1447.6780
1403.6260	12	C	12	1350.6253
1490.6580	13	S	11	1190.5946
1589.7264	14	V	10	1103.5626
1690.7741	15	T	9	1004.4942
1850.8048	16	C	8	903.4465
1907.8262	17	G	7	743.4159
2021.8692	18	N	6	686.3944
2078.8906	19	G	5	572.3515
2191.9747	20	I	4	515.3300
2320.0333	21	Q	3	402.2459
2419.1017	22	V	2	274.1874
	23	R	1	175.1190

Figure S1b. CID of the peptide modified with O-Fuc-Glc, $[M+2H]^{++}$ = 1451.17 *m/z*, CE = 23 %. At this CE, some of the precursor is still intact ($[M+2H]^{++}$, red triangle = O-Fuc, blue circle = β1,3-Glc). Neutral loss of Glc can be seen as a peak at 1370.14 *m/z* ($[M+2H-Glc]^{++}$, red triangle = O-Fuc). The dominant peak at 1297.11 *m/z* is the precursor with neutral loss of the O-Fuc-Glc disaccharide ($[M+2H-Fuc-Glc]^{++}$). Peptide fragments are present at low abundance, identifying the sequence of the unmodified peptide.



b+	#	Seq	#	y+
114.0913	1	I	23	
242.1499	2	Q	22	2480.1293
356.1928	3	N	21	2352.0707
443.2249	4	S	20	2238.0278
556.3089	5	L	19	2150.9957
643.3410	6	S	18	2037.9117
744.3886	7	T	17	1950.8797
873.4312	8	E	16	1849.8320
1059.5106	9	W	15	**1720.7894**
1146.5426	10	S	14	**1534.7101**
1243.5953	11	P	13	**1447.6780**
1403.6260	12	C	12	1350.6253
1490.6580	13	S	11	**1190.5946**
1589.7264	14	V	10	1103.5626
1690.7741	15	T	9	**1004.4942**
1850.8048	16	C	8	**903.4465**
1907.8262	17	G	7	**743.4159**
2021.8692	18	N	6	**686.3944**
2078.8906	19	G	5	**572.3515**
2191.9747	20	I	4	**515.3300**
2320.0333	21	Q	3	**402.2459**
2419.1017	22	V	2	274.1874
	23	R	1	175.1190

Figure S1c. MS^3 (CE = 35 %) of the 1370.14 *m/z* fragment seen in Figure S1b produces a dominant peak at 1297.11 *m/z* matching the mass of the unmodified peptide. Many unmodified peptide fragment peaks confirm the sequence of the peptide. This spectrum confirms that the 1370.14 *m/z* peak seen in Figure S1b arises from neutral loss of Glc from the O-Fuc-Glc dissacharide.



b+	#	Seq	#	y+
114.0913	1	I	23	
242.1499	2	Q	22	2480.1293
356.1928	3	N	21	2352.0707
443.2249	4	S	20	2238.0278
556.3089	5	L	19	2150.9957
643.3410	6	S	18	2037.9117
744.3886	7	T	17	1950.8797
873.4312	8	E	16	1849.8320
1059.5106	9	W	15	1720.7894
1146.5426	10	S	14	1534.7101
1243.5953	11	P	13	1447.6780
1403.6260	12	C	12	1350.6253
1490.6580	13	S	11	1190.5946
1589.7264	14	V	10	1103.5626
1690.7741	15	T	9	1004.4942
1850.8048	16	C	8	903.4465
1907.8262	17	G	7	743.4159
2021.8692	18	N	6	686.3944
2078.8906	19	G	5	572.3515
2191.9747	20	I	4	515.3300
2320.0333	21	Q	3	402.2459
2419.1017	22	V	2	274.1874
	23	R	1	175.1190

Figure S1d. MS^4 of the 1298.11 *m/z* peak seen in Figure S1c (MS^2 CE = 23 %, MS^3 CE = 25 %, MS^4 CE = 35 %). This peak arises from neutral loss of O-Fuc after neutral loss of Glc from the O-Fuc-Glc disaccharide. Fragment ions confirm the sequence of peptide. This spectrum further confirms that the 1370.14 *m/z* peak seen in Figure S1b arises from neutral loss of Glc from the O-Fuc-Glc dissacharide.

Figure S2. Neutral loss of O-linked and C-linked glycans at increasing collision energies.
A tryptic digest of recombinant *P. falciparum* thrombospondin related anonymous protein (TRAP) was analyzed by LC-MS/MS with an LTQ-Orbitrap Elite. Precursor ions for the TSR peptide modified with C-Man and O-Fuc-Glc were isolated and fragmented at sequentially increasing normalized collision energies (CE). All spectra are collision-induced dissociation (CID) collected at high resolution in the Orbitrap. Annotated fragment spectra were visualized using a development version of Lorikeet as implemented in the Trans-Proteomic Pipeline.

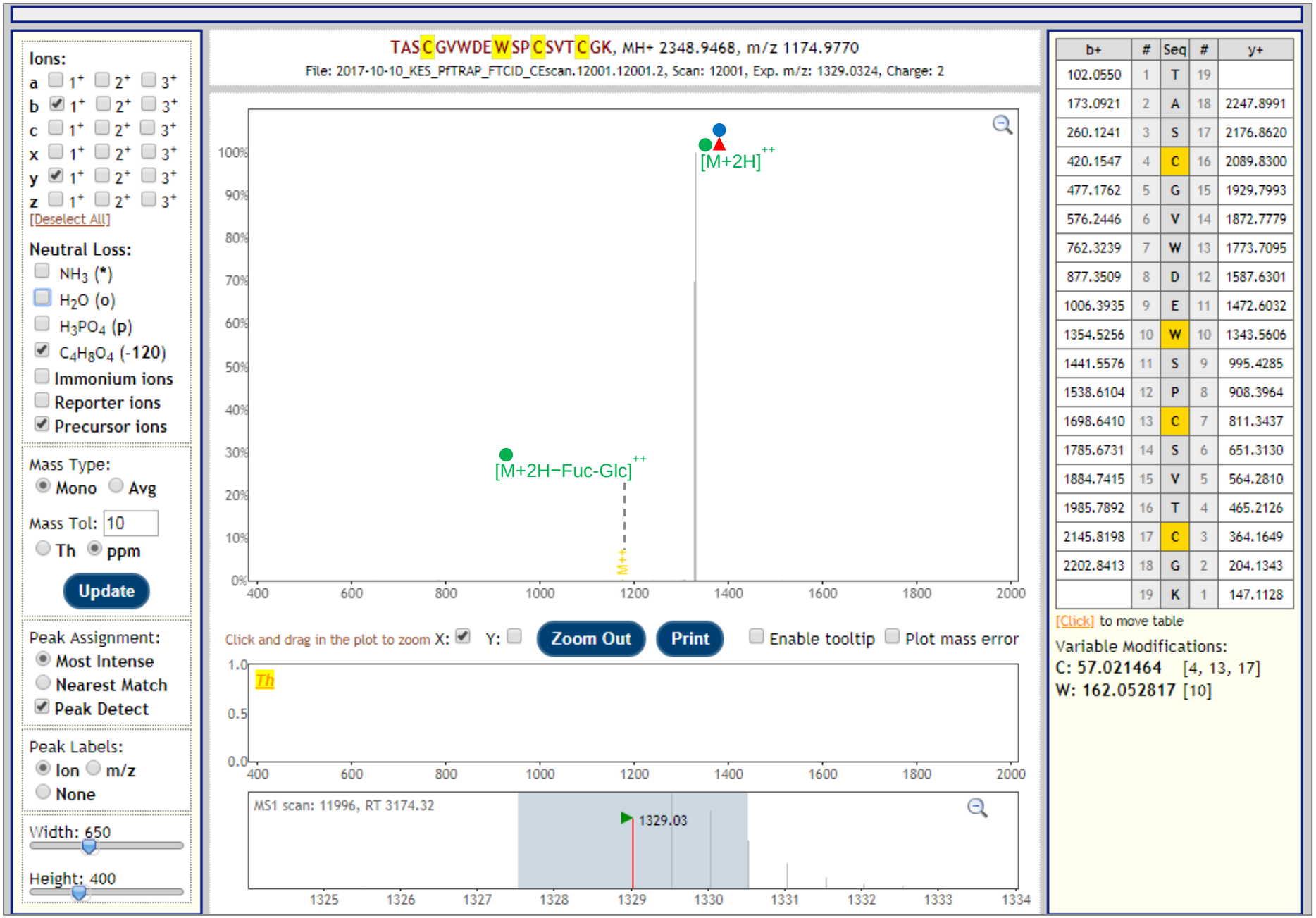


Figure S2a. CID at 1329.03 *m/z*, CE = 20 %. The precursor peak is largely unreacted ([M+2H]$^{++}$, green circle = C-Man, red triangle = O-Fuc, blue circle = β1,3-Glc), though a small peak (~0.3% relative abundance) matches the mass of the peptide after neutral loss of the O-linked disaccharide (1174.98 *m/z,* yellow, automatically annotated by Lorikeet as the precursor peak M++).

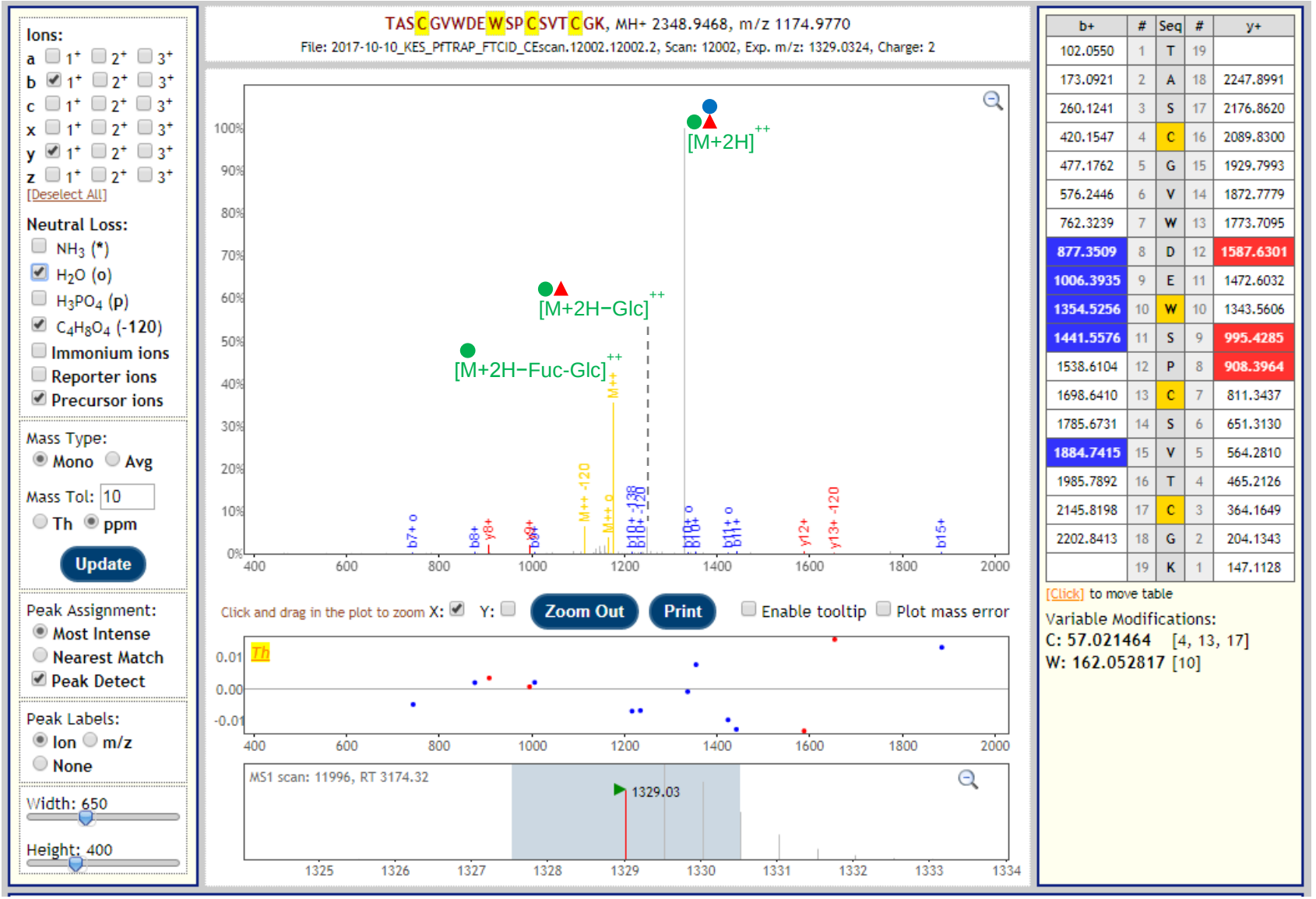


Figure S2b. CID at 1329.03 *m/z*, CE = 25 %. The dominant peak is still unreacted precursor ($[M+2H]^{++}$, green circle = C-Man, red triangle = O-Fuc, blue circle = β1,3-Glc). A smaller peak at 1248.01 *m/z* matches the mass of the precursor after neutral loss of Glc from the O-Fuc-Glc disaccharide ($[M+2H-Glc]^{++}$). A third peak at 1174.98 *m/z* (yellow, automatically annotated by Lorikeet as the precursor peak M++) matches the mass of the peptide precursor ion after neutral loss of the O-Fuc-Glc disaccharide ($[M+2H-Fuc-Glc]^{++}$). Note that peptide fragment ions appearing at this CE identify the sequence of the peptide with C-Man intact, positively localizing the glycan to the C-terminal Trp residue. Neutral loss of 120.04 Da from cross-ring cleavage is seen on the precursor and fragment peaks.

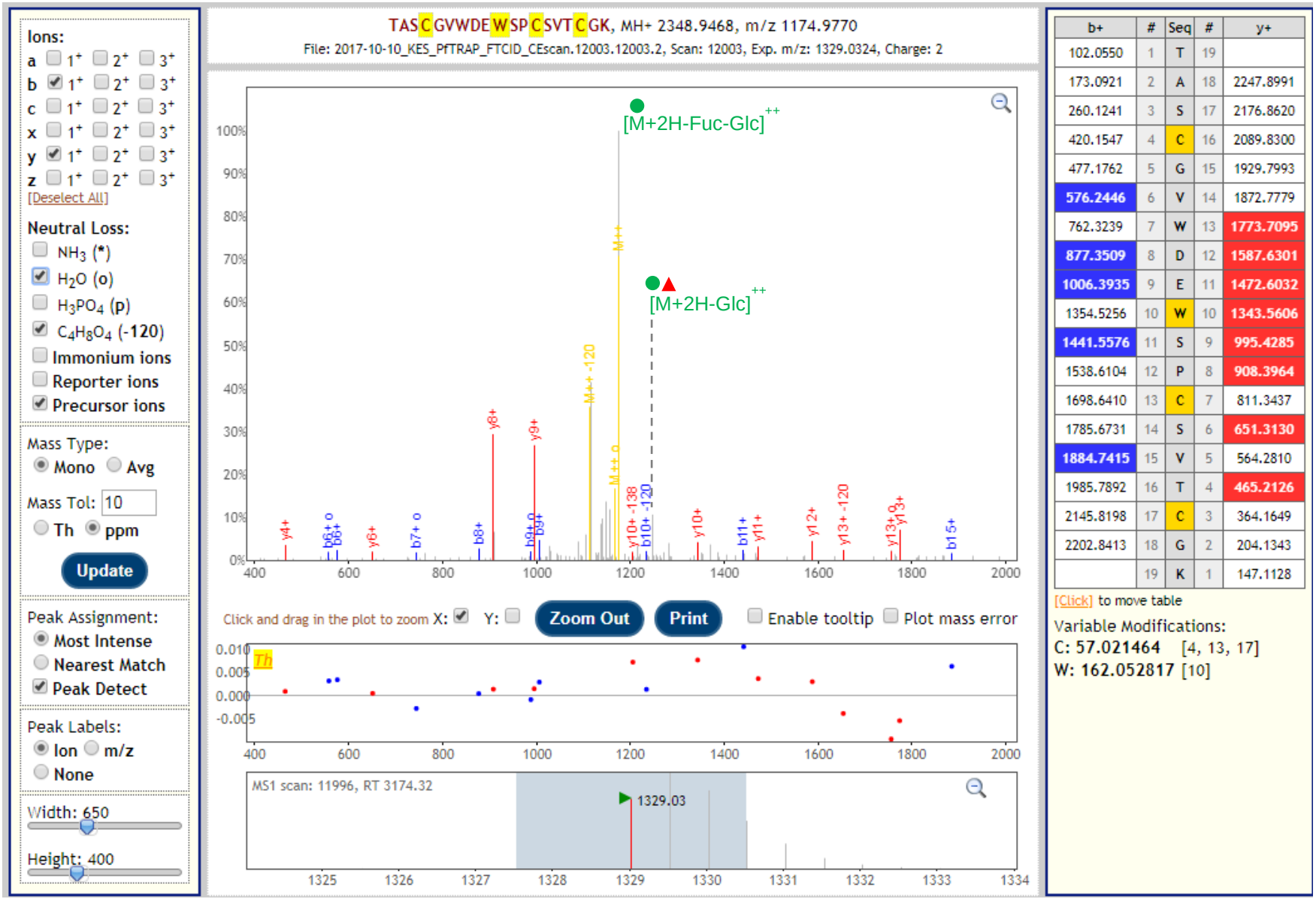
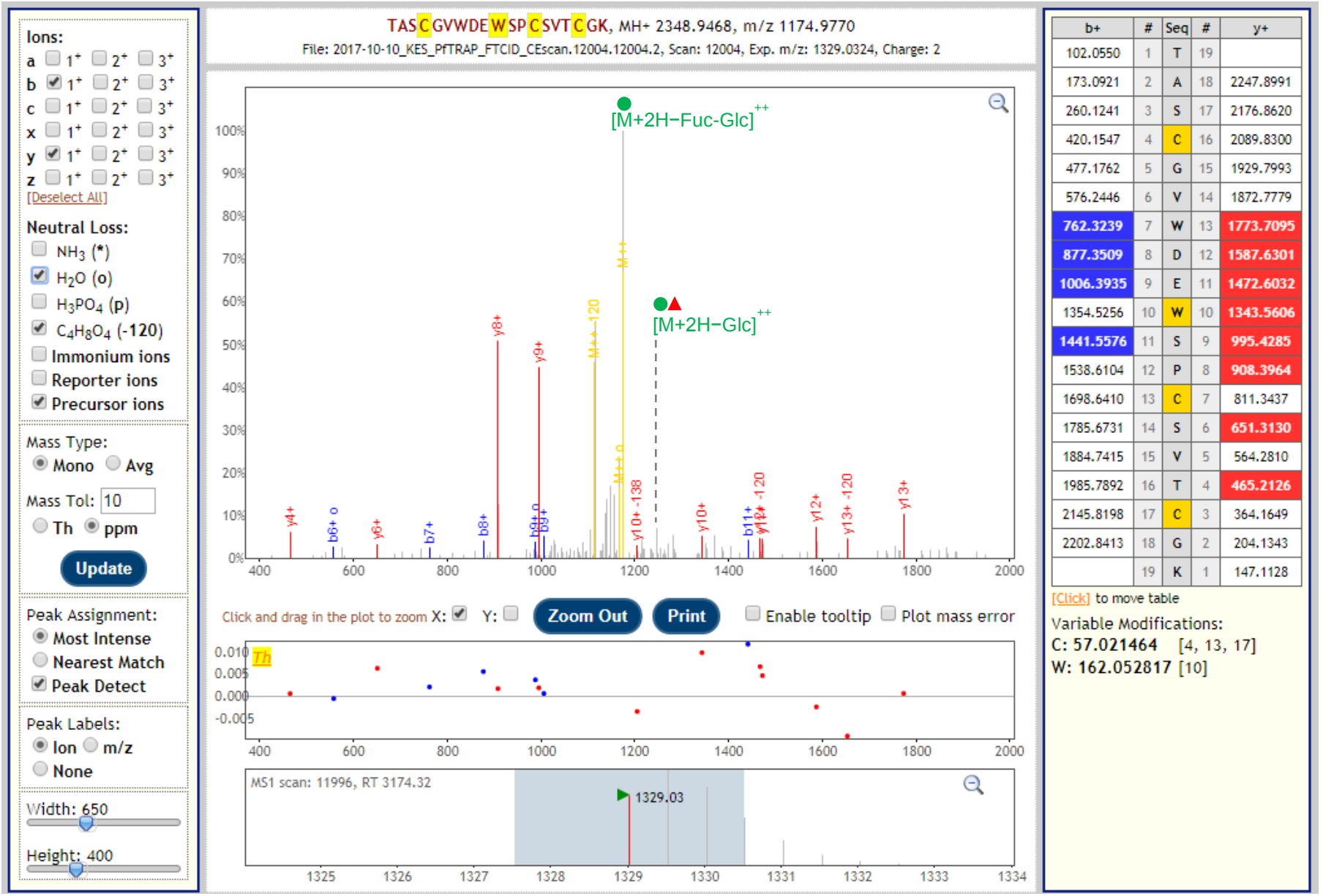


Figure S2c. CID at 1329.03 *m/z*, CE = 30 %. At this CE all parent precursor ion has been fragmented. The small peak at 1248.01 *m/z* remains, the mass of the precursor after neutral loss of Glc from the O-Fuc-Glc disaccharide ([M+2H−Glc]$^{++}$). The dominant peak is now 1174.98 *m/z* (yellow M++), the peptide precursor ion after neutral loss of the O-Fuc-Glc disaccharide ([M+2H−Fuc-Glc]$^{++}$). Further peptide fragment ions confidently identify the sequence of the peptide with C-Man intact.



b+	#	Seq	#	y+
102.0550	1	T	19	
173.0921	2	A	18	2247.8991
260.1241	3	S	17	2176.8620
420.1547	4	C	16	2089.8300
477.1762	5	G	15	1929.7993
576.2446	6	V	14	1872.7779
762.3239	7	W	13	1773.7095
877.3509	8	D	12	1587.6301
1006.3935	9	E	11	1472.6032
1354.5256	10	W	10	1343.5606
1441.5576	11	S	9	995.4285
1538.6104	12	P	8	908.3964
1698.6410	13	C	7	811.3437
1785.6731	14	S	6	651.3130
1884.7415	15	V	5	564.2810
1985.7892	16	T	4	465.2126
2145.8198	17	C	3	364.1649
2202.8413	18	G	2	204.1343
	19	K	1	147.1128

[Click] to move table

Variable Modifications:
C: 57.021464 [4, 13, 17]
W: 162.052817 [10]

Figure S2d. CID at 1329.03 *m/z*, CE = 35 %. This is the CE typically used in shotgun proteomics methods. The C-mannosylated peptide precursor lacking the O-linked glycan remains the dominant peak, but peptide fragment ions are abundant enough to confidently sequence the peptide. Unlike the recombinant *Pf*CSP spectra in Figure 1 and supplemental Figure S1b, no peptide fragment ions with the O-linked glycan intact are visible at any CE. This is likely a signal-to-noise issue; the spectra in Figure 1 and Figure S1 were acquired by direct infusion of the peptide, allowing signal and ion fill time to be maximized, whereas these spectra were acquired on-the-fly by LC-MS/MS.

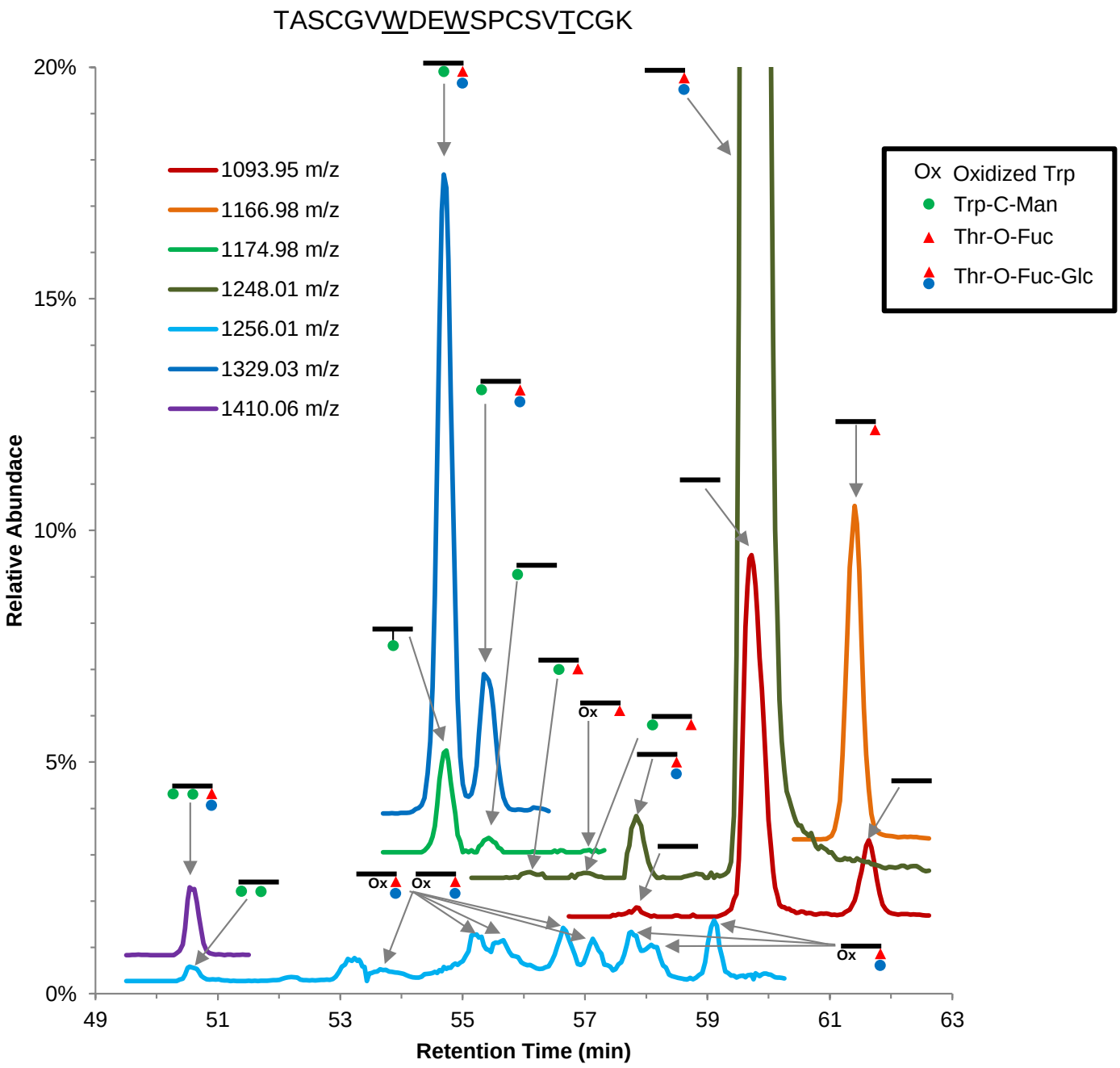
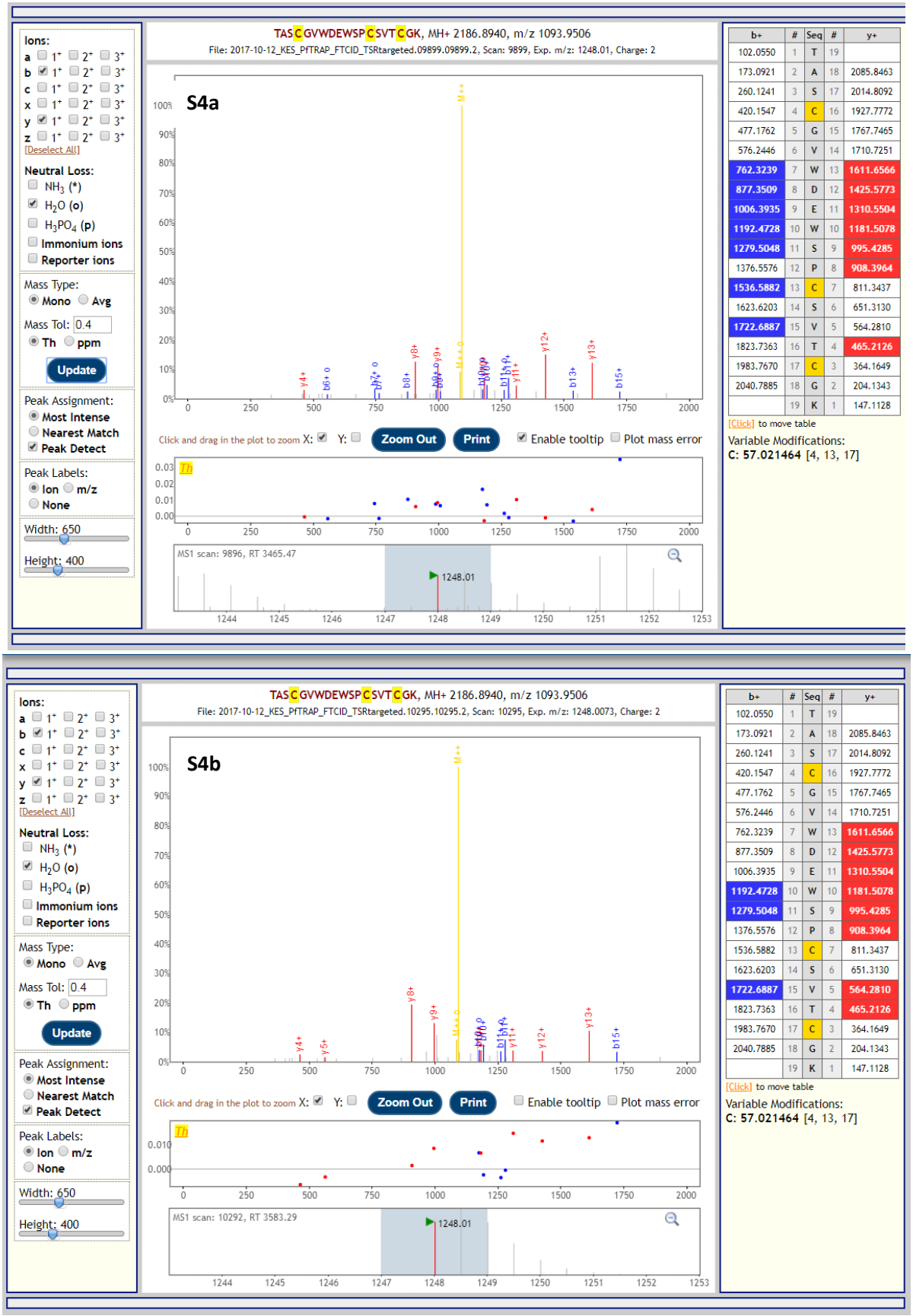


Figure S3. Glycoforms identified by automated sequence database search of data from LC-MS/MS analysis of recombinant *P. falciparum* TRAP. The variably C-mannosylated or oxidized Trp residues and the variably O-fucosylated Thr are underlined. Traces are offset for clarity. Peaks indicated with arrows were positively identified by PSMs. Both positional isoforms of the oxidized peptide were identified at several co-eluting retention times. See table S2.

Figure S4. Indistinguishable isobaric glycopeptides from recombinant *P. falciparum* TRAP.



b+	#	Seq	#	y+
102.0550	1	T	19	
173.0921	2	A	18	2085.8463
260.1241	3	S	17	2014.8092
420.1547	4	C	16	1927.7772
477.1762	5	G	15	1767.7465
576.2446	6	V	14	1710.7251
762.3239	7	W	13	1611.6566
877.3509	8	D	12	1425.5773
1006.3935	9	E	11	1310.5504
1192.4728	10	W	10	1181.5078
1279.5048	11	S	9	995.4285
1376.5576	12	P	8	908.3964
1536.5882	13	C	7	811.3437
1623.6203	14	S	6	651.3130
1722.6887	15	V	5	564.2810
1823.7363	16	T	4	465.2126
1983.7670	17	C	3	364.1649
2040.7885	18	G	2	204.1343
	19	K	1	147.1128

[Click] to move table

b+	#	Seq	#	y+
102.0550	1	T	19	
173.0921	2	A	18	2085.8463
260.1241	3	S	17	2014.8092
420.1547	4	C	16	1927.7772
477.1762	5	G	15	1767.7465
576.2446	6	V	14	1710.7251
762.3239	7	W	13	1611.6566
877.3509	8	D	12	1425.5773
1006.3935	9	E	11	1310.5504
1192.4728	10	W	10	1181.5078
1279.5048	11	S	9	995.4285
1376.5576	12	P	8	908.3964
1536.5882	13	C	7	811.3437
1623.6203	14	S	6	651.3130
1722.6887	15	V	5	564.2810
1823.7363	16	T	4	465.2126
1983.7670	17	C	3	364.1649
2040.7885	18	G	2	204.1343
	19	K	1	147.1128

[Click] to move table

Figure S4a,b. Representative MS^2 of the two 1248.01 *m/z* peaks eluting at **a)** 57.8 min and **b)** 58.9 min. The MS^2 from both species appear to identify the same species, i.e. the peptide modified with a gas-phase-labile moiety having a mass matching that of O-Fuc-Glc. The dominant species of the MS^2 spectra, 1093.95 *m/z*, matches the mass of the unmodified peptide. Additional PSMs are listed in Table S2.

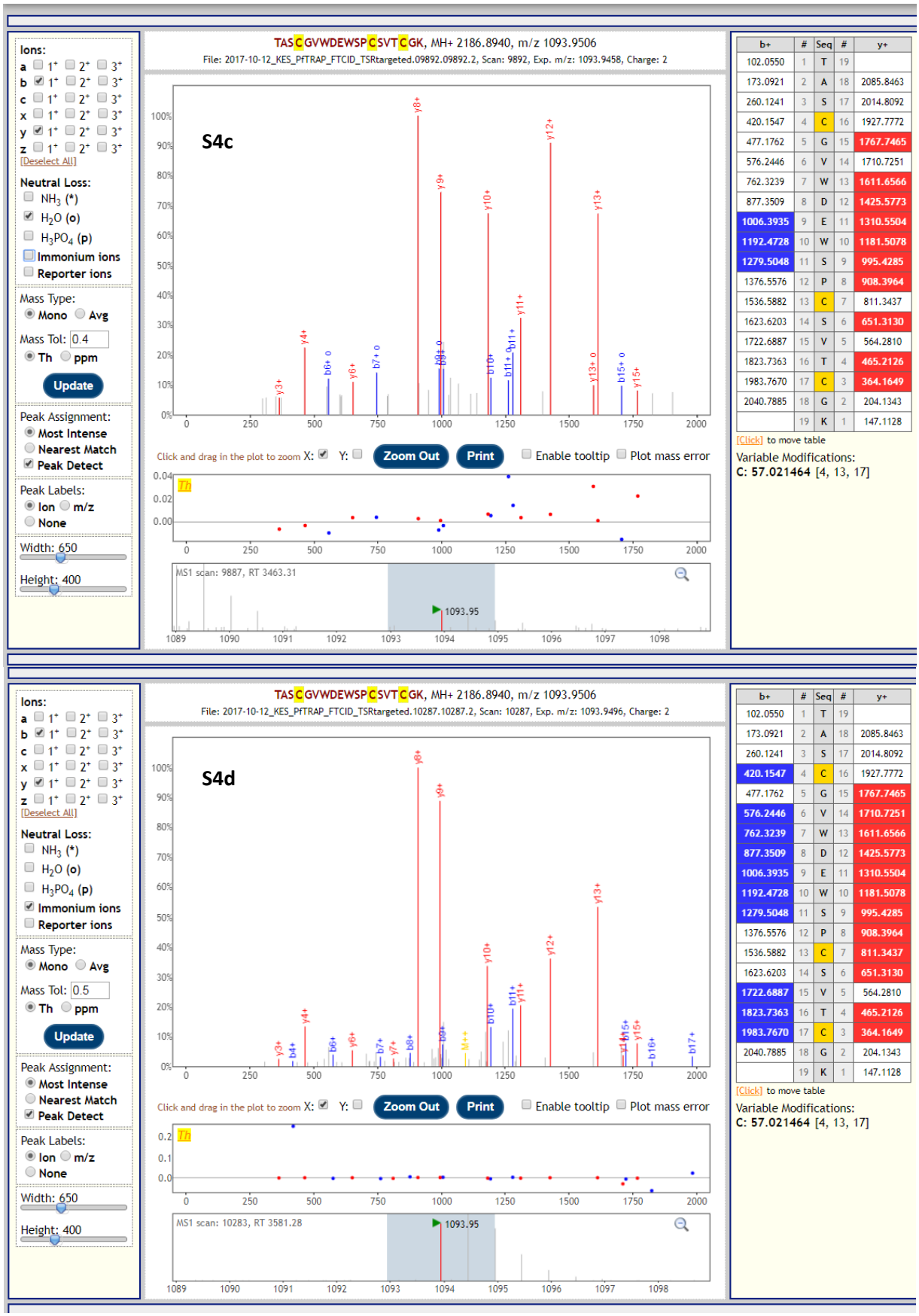


Figure S4c,d. Representative MS^2 of the two 1093.5 *m*/z peaks eluting at **c)** 57.8 min and **d)** 58.9 min. The MS^2 from both species appear to identify the unmodified peptide. Both species co-elute with the 1248.01 m/z species identified in Fig. S4a and b, consistent with loss of the gas-phase-labile O-Fuc-Glc from in-source fragmentation. Additional PSMs are listed in Table S2.

S5. Misidentification of a glycoform in LC-MS/MS of recombinant *P. falciparum* TRAP

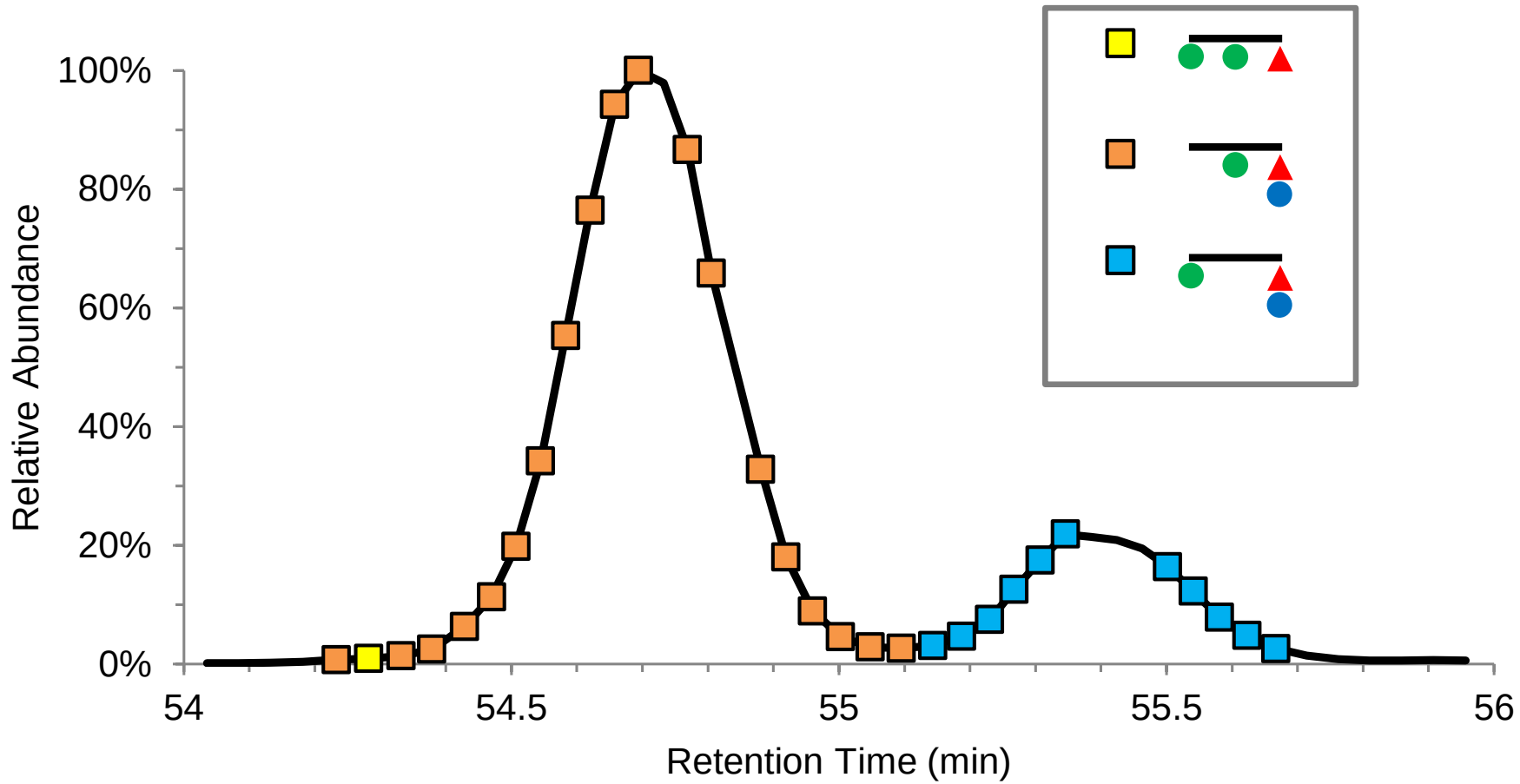
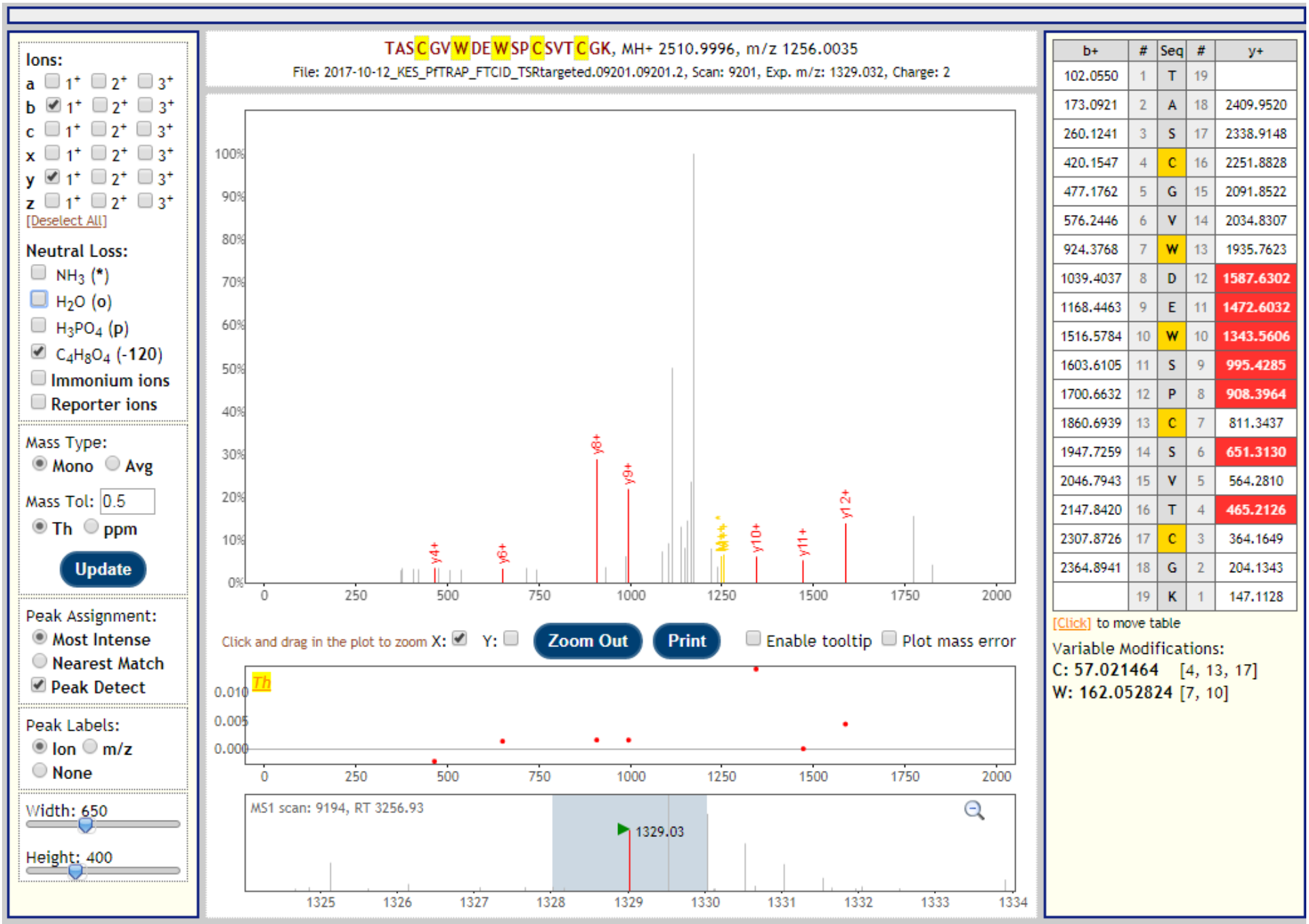


Figure S5a. Extracted Ion Chromatogram (XIC) of the precursor ion matching multiple isobaric glycoforms. The XIC of 1329.03 *m/z* (black line) is overlaid with squares showing points where MS^2 resulted in high-quality peptide spectrum matches (PSMs) identifying glycoforms of the recombinant *Pf*TRAP TSR peptide TASCGVWDEWSPCSVTCGK modified with combinations of C-linked mannose at tryptophan (green circle), O-linked fucose at the C-terminal threonine (red triangle) and β1,3-linked glucose (blue circle) added to the fucose. See Table S2.



b+	#	Seq	#	y+
102.0550	1	T	19	
173.0921	2	A	18	2409.9520
260.1241	3	S	17	2338.9148
420.1547	4	C	16	2251.8828
477.1762	5	G	15	2091.8522
576.2446	6	V	14	2034.8307
924.3768	7	W	13	1935.7623
1039.4037	8	D	12	**1587.6302**
1168.4463	9	E	11	**1472.6032**
1516.5784	10	W	10	**1343.5606**
1603.6105	11	S	9	**995.4285**
1700.6632	12	P	8	**908.3964**
1860.6939	13	C	7	811.3437
1947.7259	14	S	6	**651.3130**
2046.7943	15	V	5	564.2810
2147.8420	16	T	4	**465.2126**
2307.8726	17	C	3	364.1649
2364.8941	18	G	2	204.1343
	19	K	1	147.1128

[Click] to move table

Variable Modifications:
C: 57.021464 [4, 13, 17]
W: 162.052824 [7, 10]

Figure S5b A single PSM (the yellow square in Figure S5a) was incorrectly identified as the glycoform of the peptide featuring double mannosylation and an O-Fuc, with the MS^2 annotated as shown here.

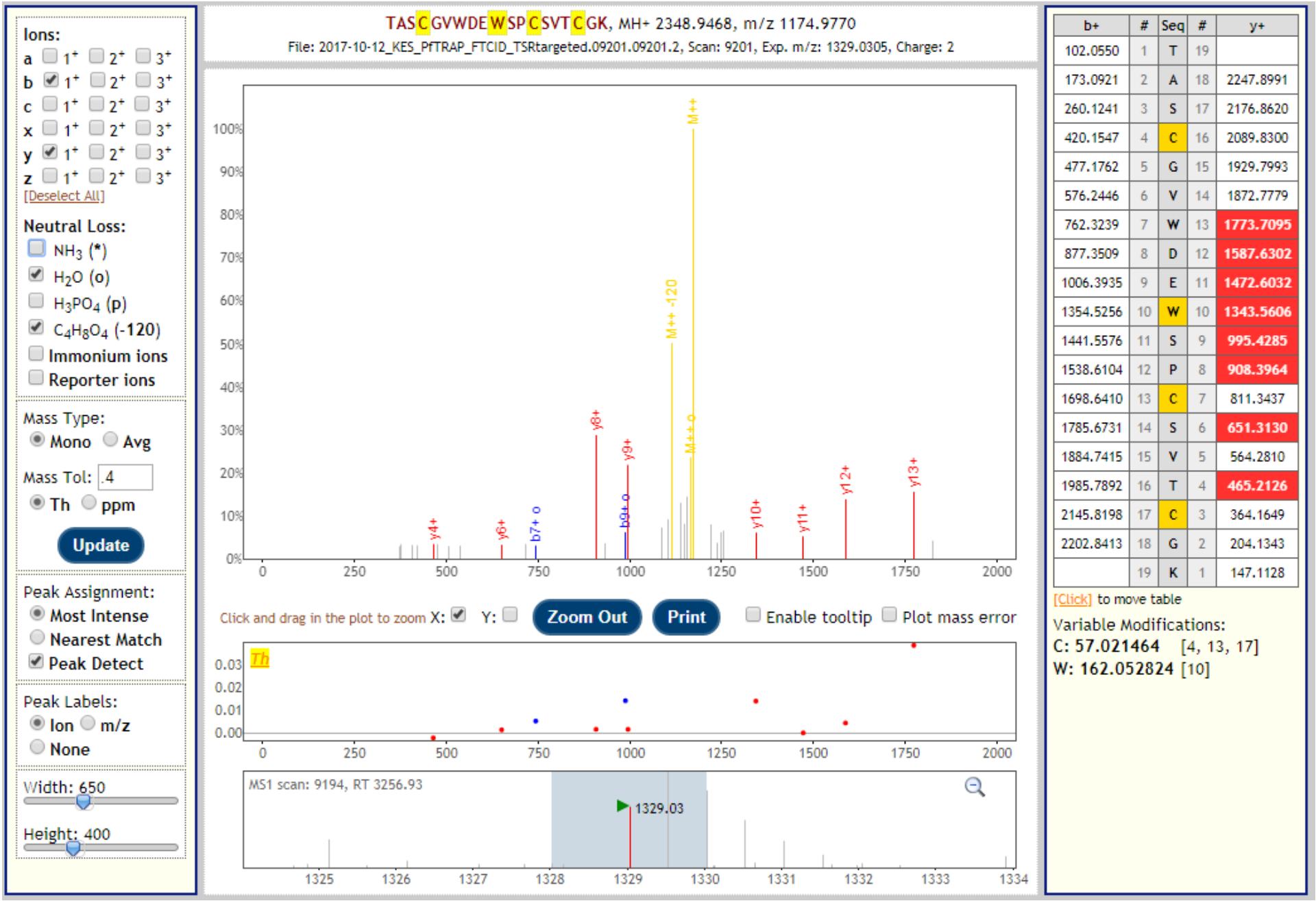


Figure S5c The same MS^2 from Figure S5b, re-annotated with the correct glycoform. While most of the fragment ions are the same as for the incorrect annotation (except y_{13}), the precursor fragment exhibiting neutral loss of O-Fuc-Glc (the yellow M++ peak) as well as cross-ring cleavage of C-man (yellow M++-120 peak) confirm that this is the correct annotation.

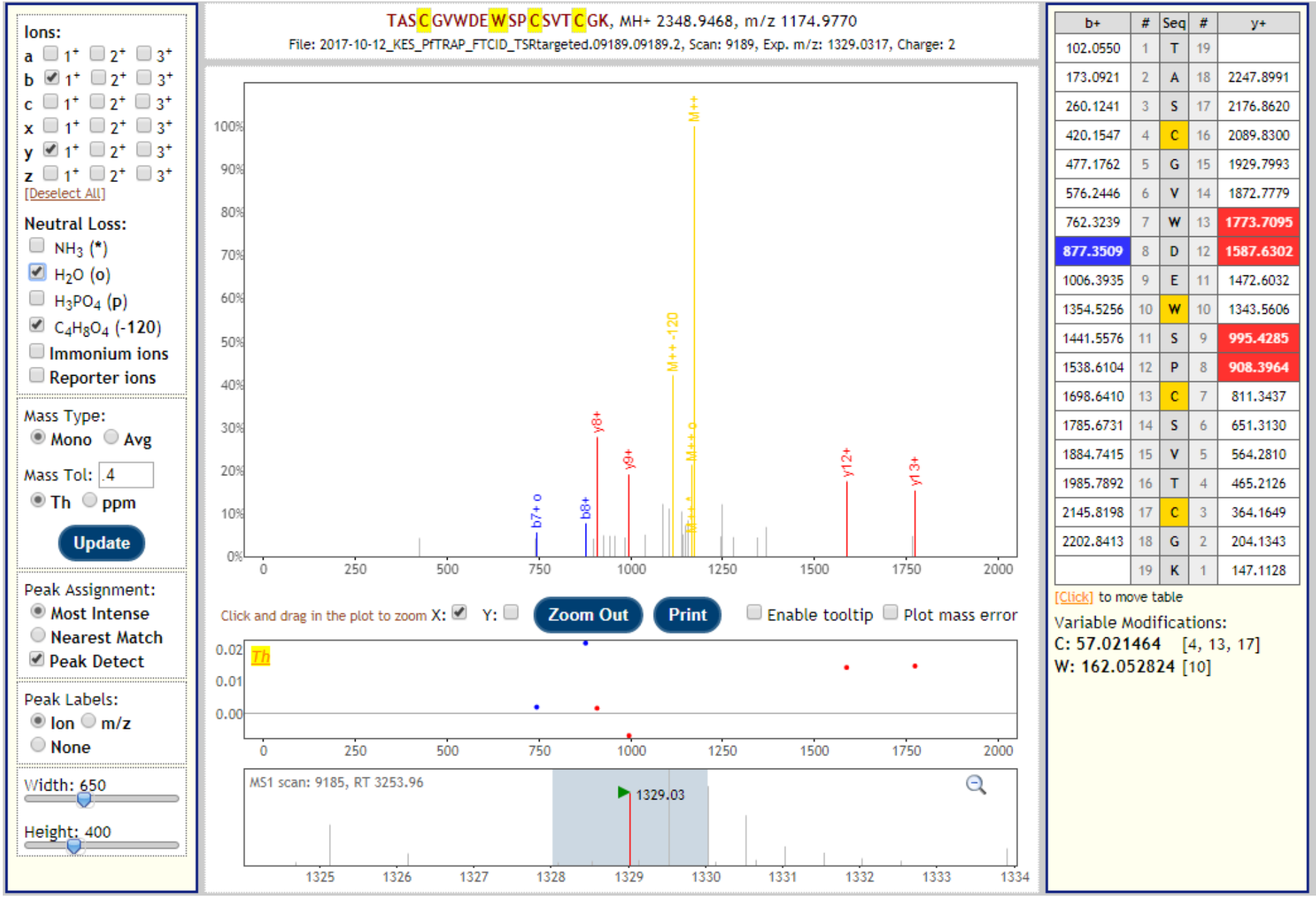


Figure S5d. The MS^2 for the same parent ion that immediately precedes the incorrect PSM, though sparse, identifies the correct glycoform.

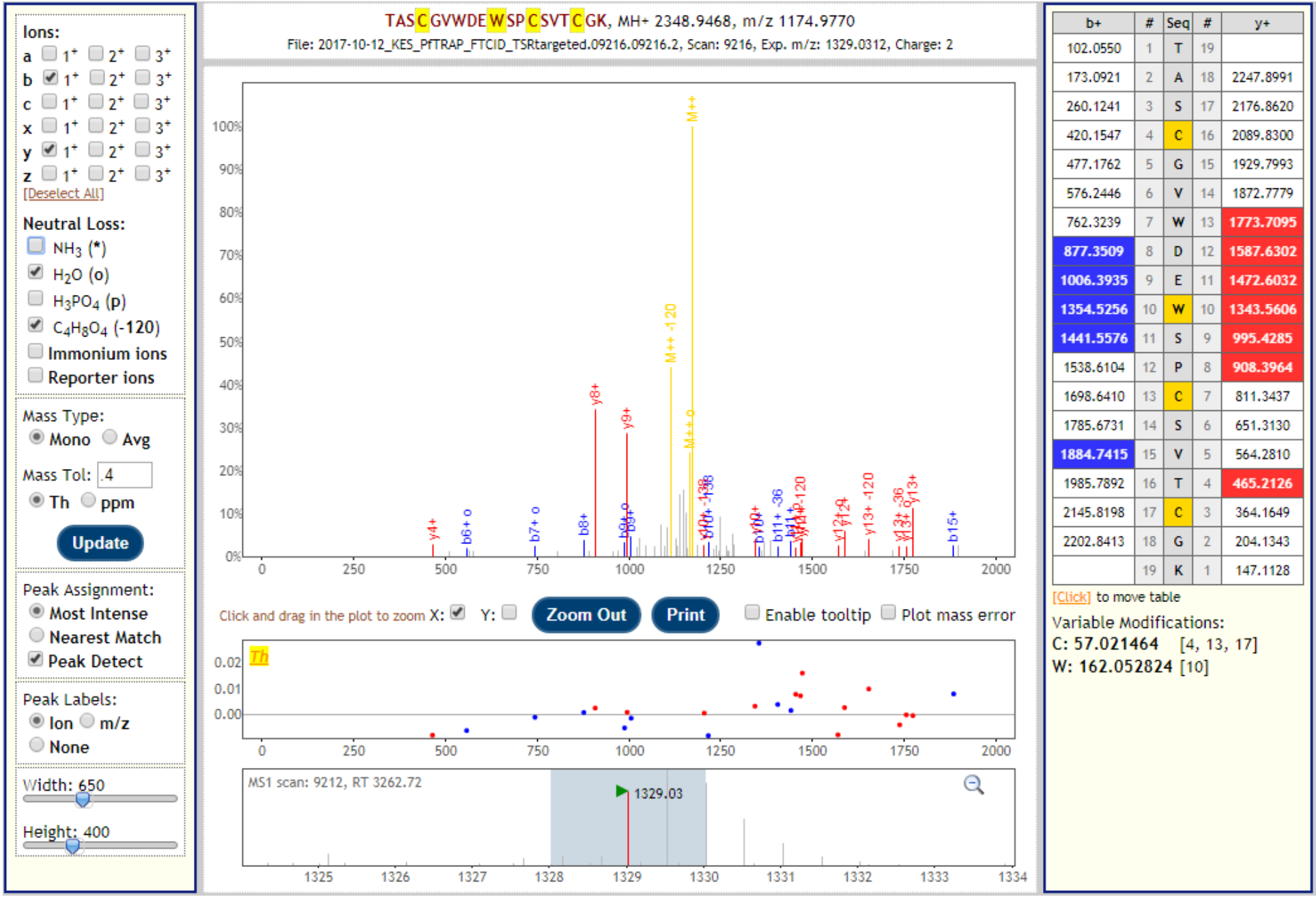


Figure S5e. The MS^2 for the same parent ion immediately following the incorrect PSM identifies the correct glycoform, as do the remaining PSMs for the parent ion in the same chromatographic peak.

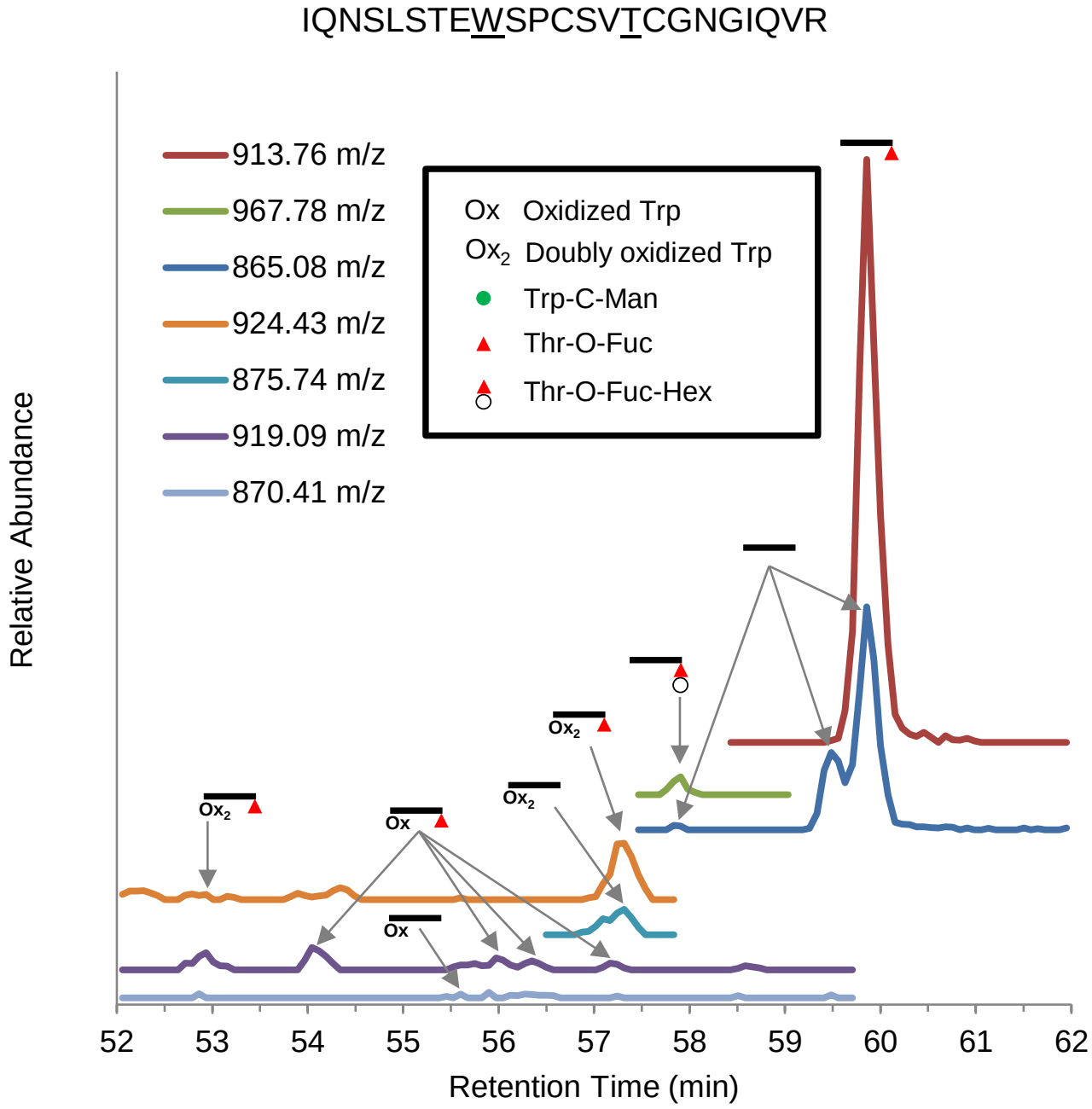


Figure S6. Representative XIC of glycosylated CSP from *P. falciparum* salivary gland sporozoites. The variably oxidized Trp and the variably O-fucosylated Thr are underlined. Traces are offset for clarity. Peaks indicated with arrows were positively identified by PSMs. The Hex of the O-Fuc-Hex disaccharide is indicated as a generic hexose because the identity of the glycan has not been confirmed.

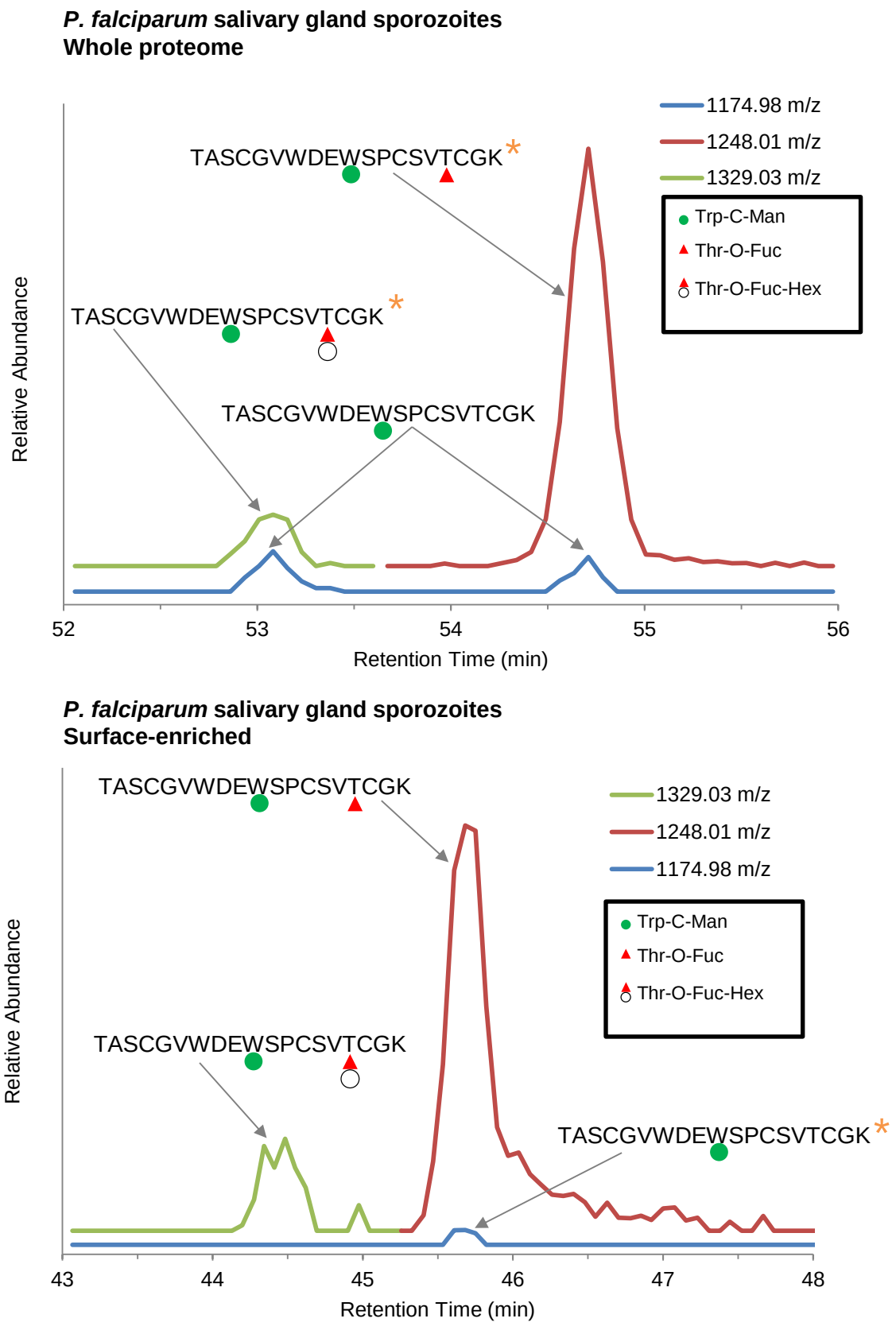


Figure S7. Representative XIC of glycosylated TRAP from *P. falciparum* salivary gland sporozoites. Traces are offset for clarity. Peaks marked with an orange asterisk (*) were not identified by PSMs because the precursor was not selected for MS^2, but their retention time and mass correspond with the indicted glycoform. The Hex of the O-Fuc-Hex disaccharide is indicated as a generic hexose because the identity of the glycan has not been confirmed.

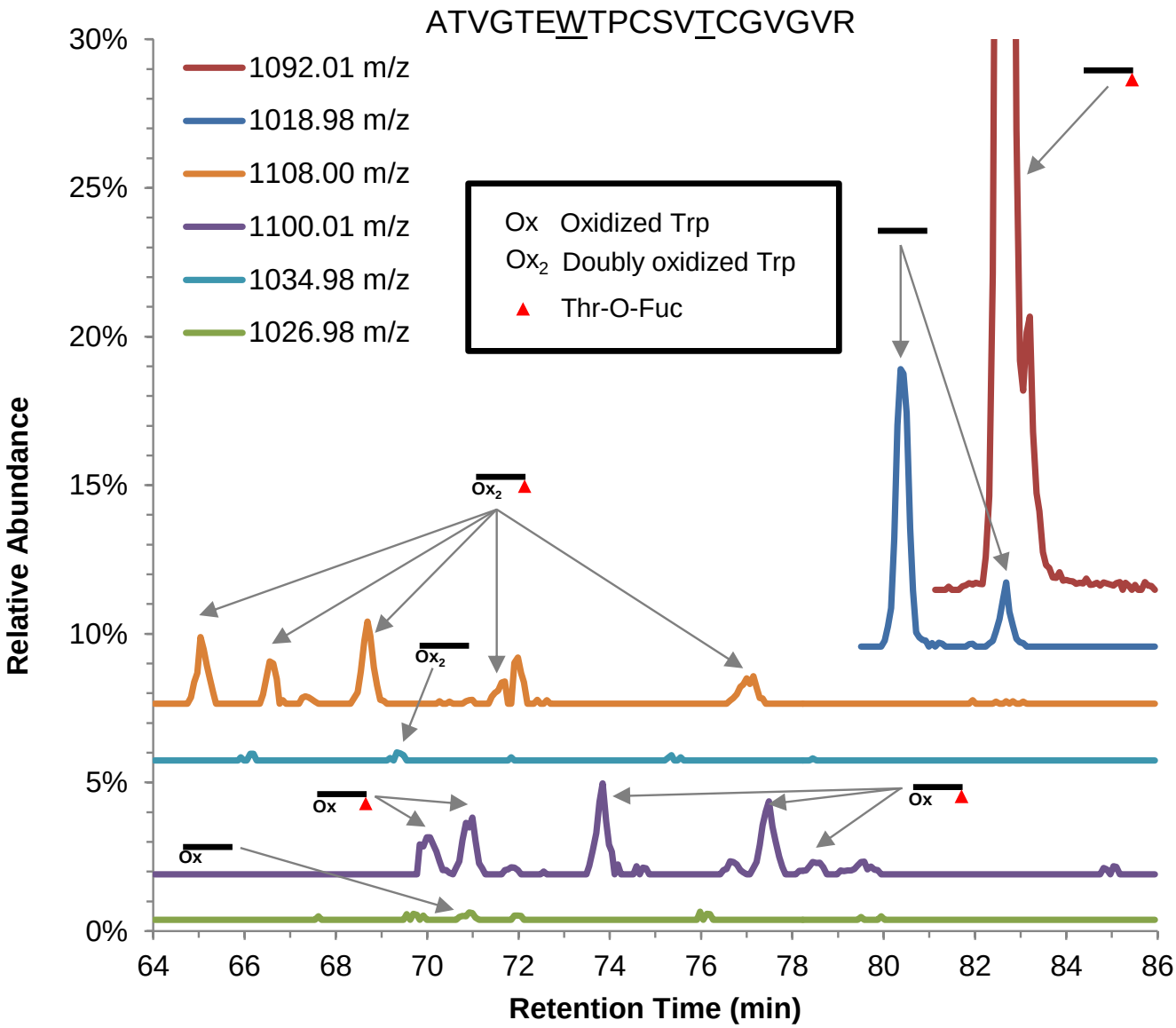


Figure S8. Representative XIC of glycosylated CSP from *P. vivax* VK210 salivary gland sporozoites. The variably oxidized Trp and the variably O-fucosylated Thr are underlined. Traces are offset for clarity. Peaks indicated with arrows were positively identified by PSMs.

VANCGPWDPWTACSVTCGR

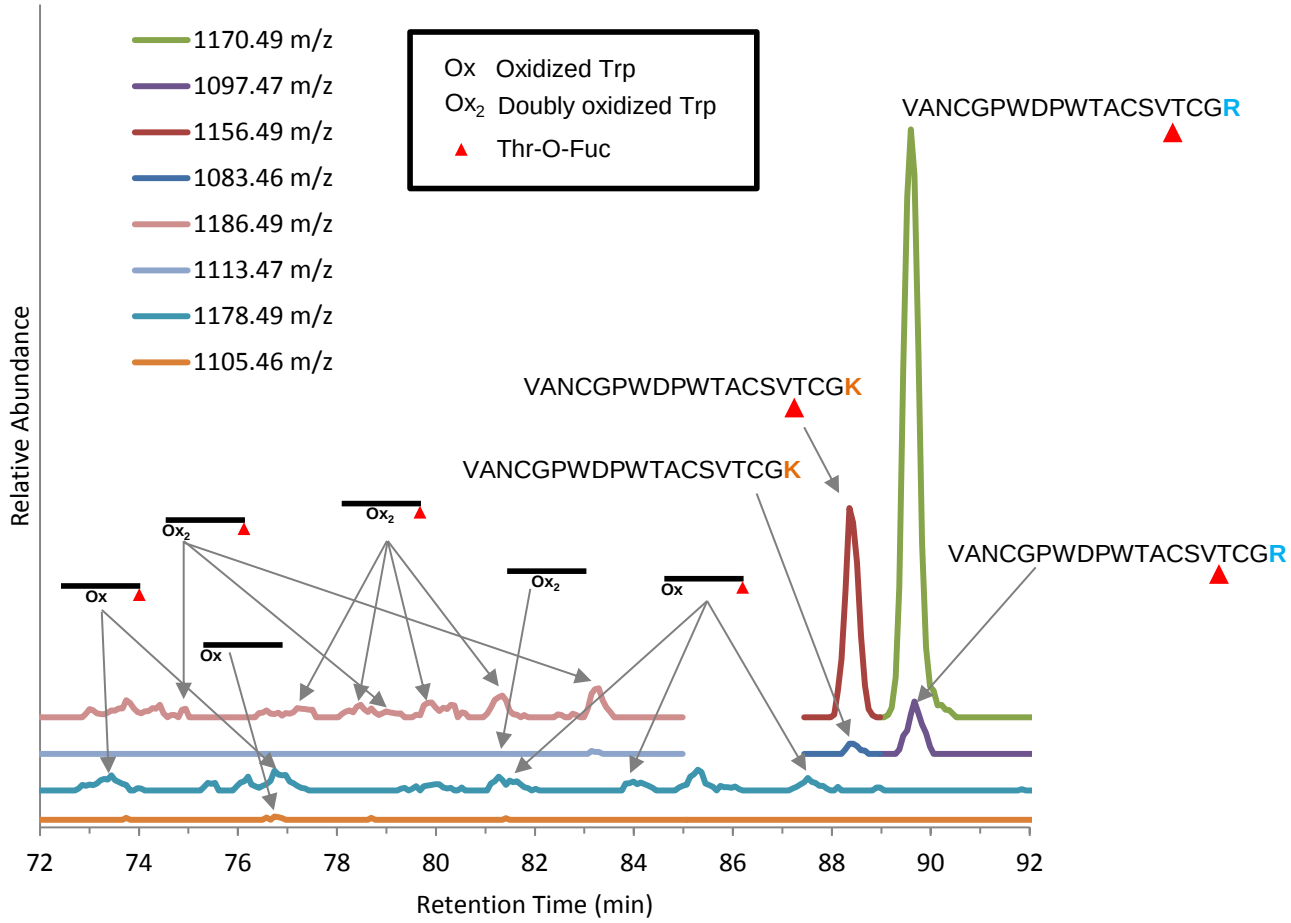


Figure S9. Representative XIC of glycosylated TRAP from *P. vivax* VK210 salivary gland sporozoites. The variably oxidized Trp residues and the variably O-fucosylated Thr are underlined. Traces are offset for clarity. Peaks indicated with arrows were positively identified by PSMs. Note that this sample contained a mixture of two field isolates, each of which carried a different version of the TRAP gene with either a Lys or Arg at the C-terminus of the glycopeptide.

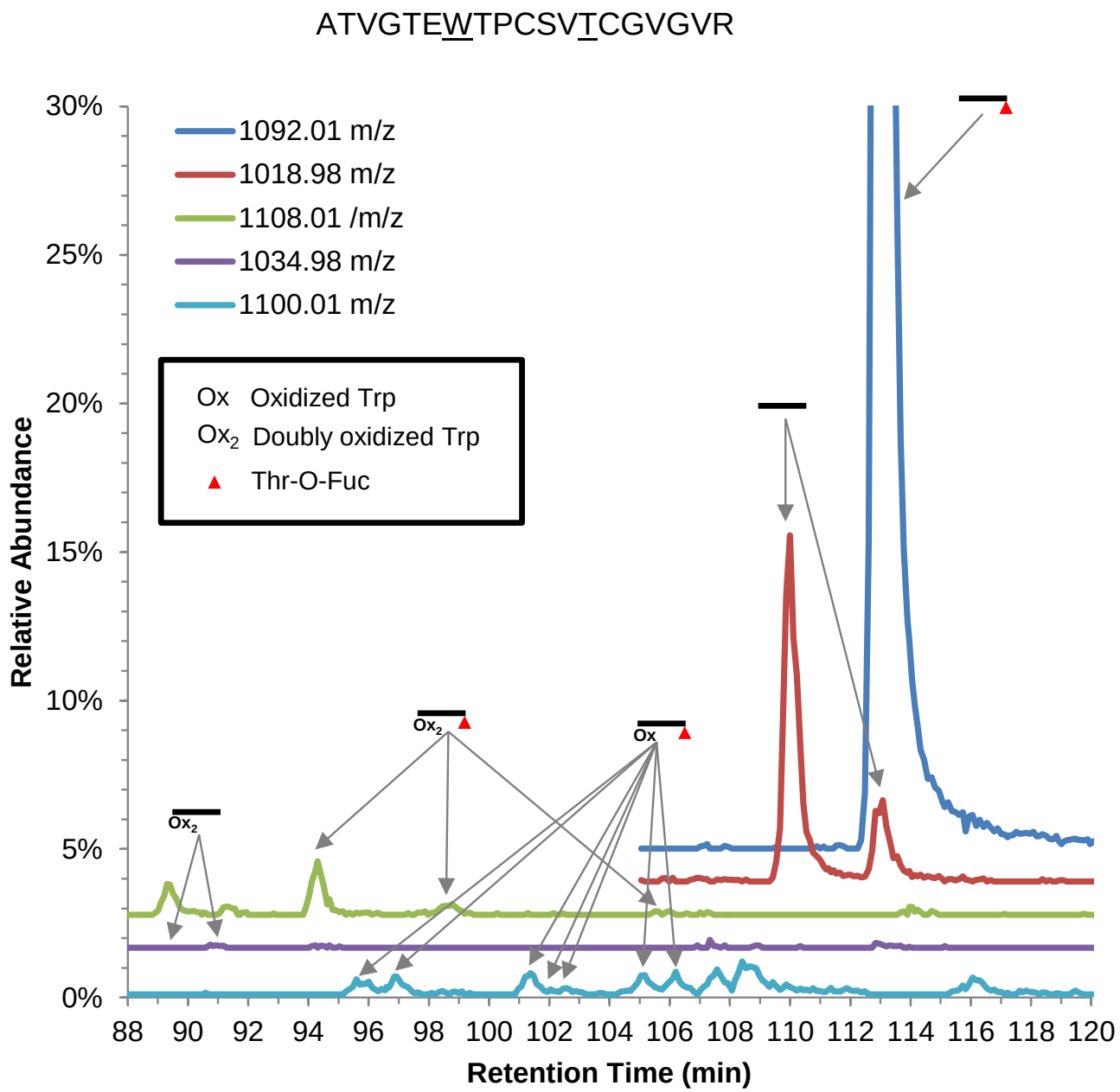


Figure S10. Representative XIC of glycosylated CSP from *P. vivax* VK247 salivary gland sporozoites. The variably oxidized Trp and the variably O-fucosylated Thr are underlined. Traces are offset for clarity. Peaks indicated with arrows were positively identified by PSMs.

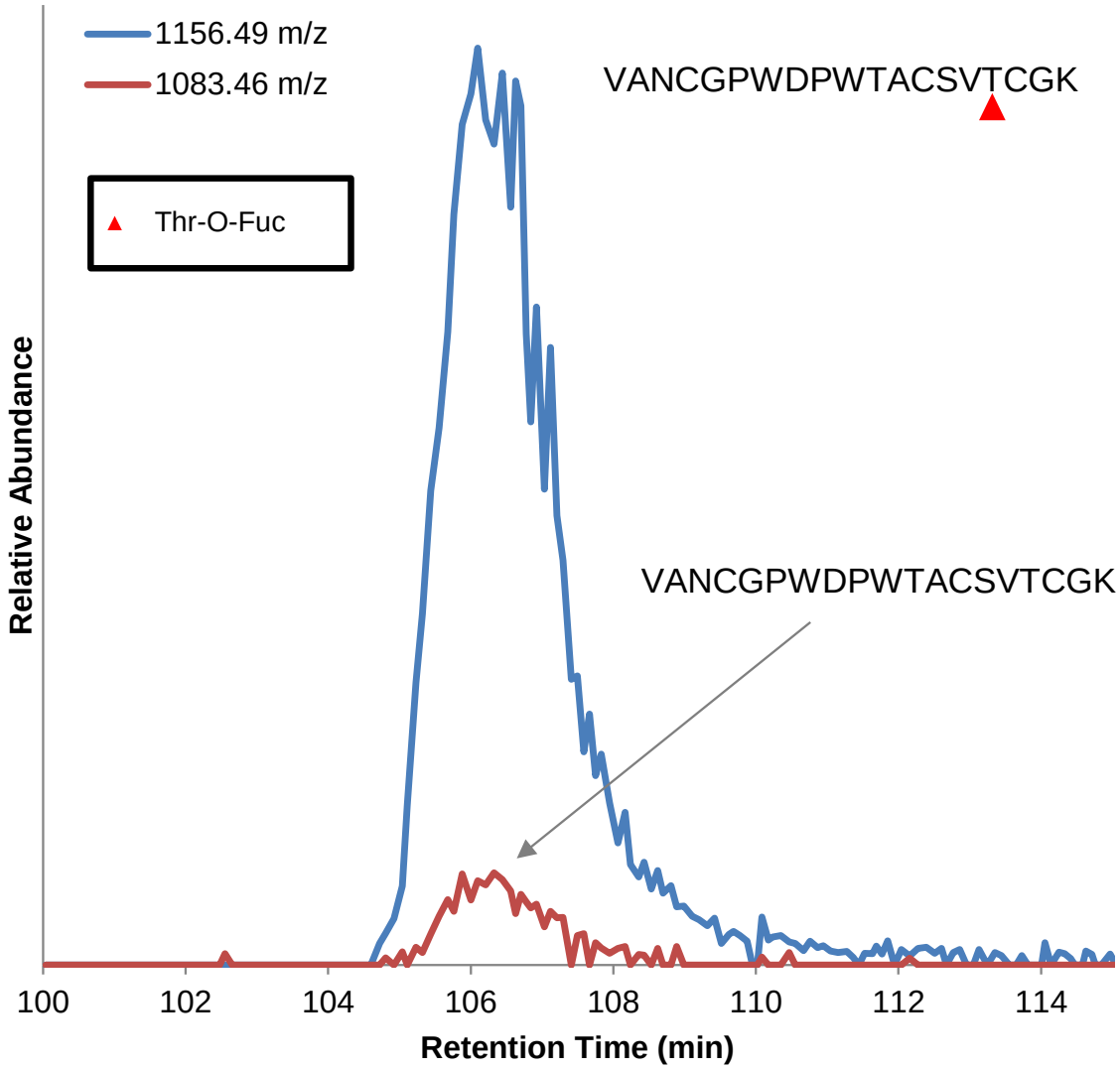


Figure S11. Representative XIC of glycosylated TRAP from *P. vivax* VK247 salivary gland sporozoites.

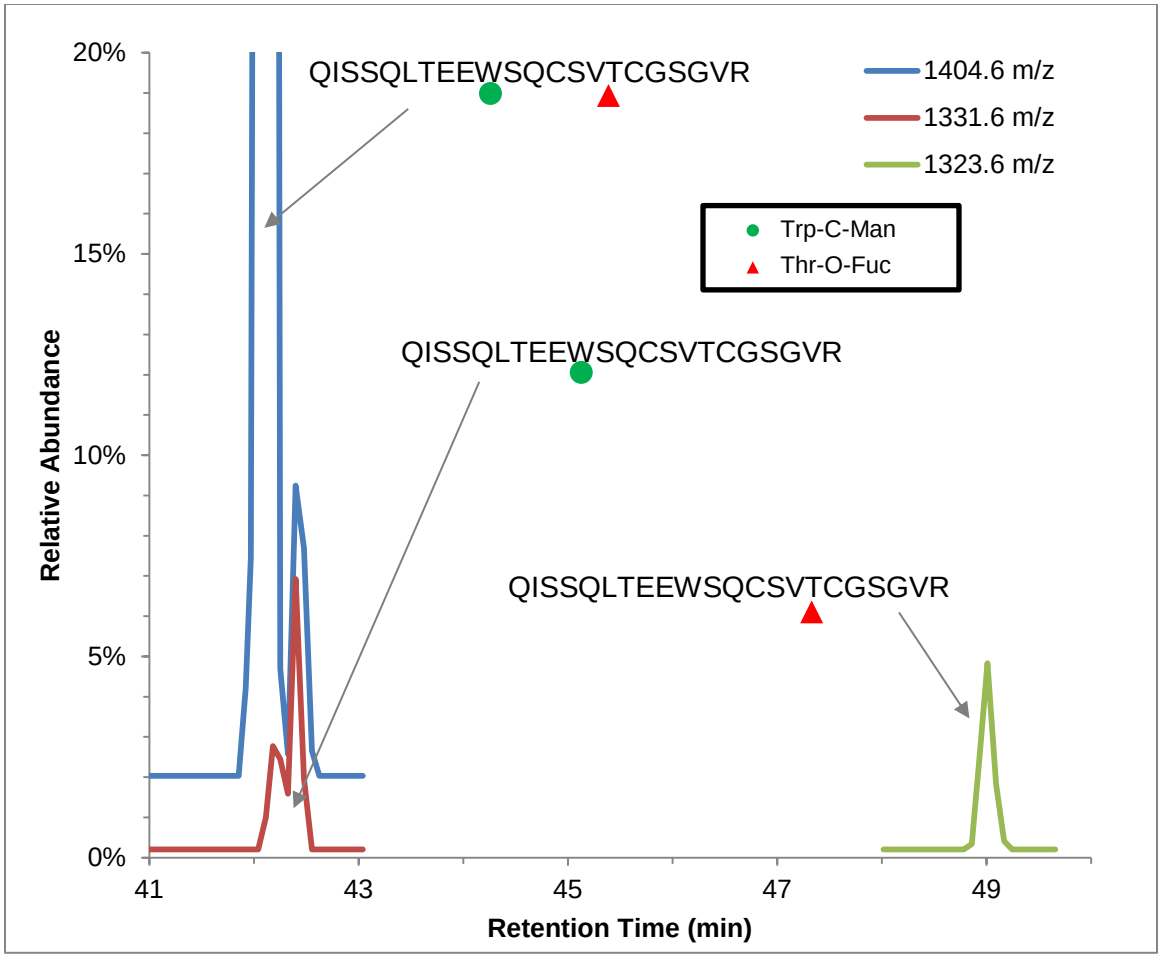


Figure S12. Representative XIC of glycosylated CSP from *P. yoelii* salivary gland sporozoites. Traces are offset for clarity. The reason for the doublet shape of the 1404.6 *m/z* and co-eluting 1331.6 *m/z* peaks is unknown, and may reflect spray instability or closely eluting isobaric species. Only the indicated glycopeptides were identified from MS^2 of these masses collected during the elution window of the peaks.

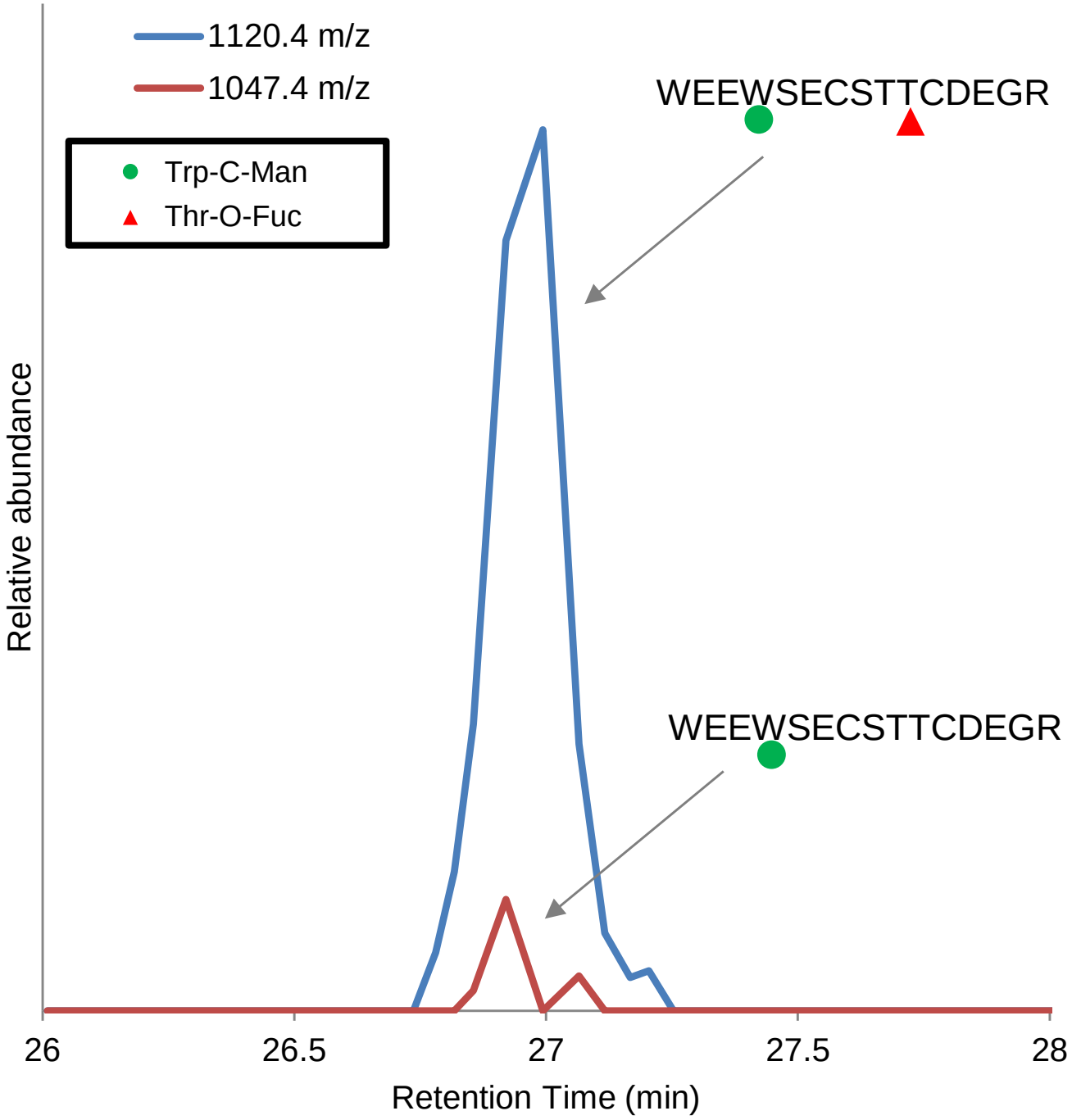


Figure S13. Representative XIC of glycosylated TRAP from *P. yoelii* salivary gland sporozoites.

Ions:
a ☐ 1+ ☐ 2+ ☐ 3+
b ☑ 1+ ☐ 2+ ☐ 3+
c ☐ 1+ ☐ 2+ ☐ 3+
x ☐ 1+ ☐ 2+ ☐ 3+
y ☑ 1+ ☐ 2+ ☐ 3+
z ☐ 1+ ☐ 2+ ☐ 3+
[Deselect All]

Neutral Loss:
☐ NH_3 (*)
☑ H_2O (o)
☐ H_3PO_4 (p)
☑ $C_4H_8O_4$ (-120)
☐ Immonium ions
☐ Reporter ions
☑ Precursor ions

Mass Type:
◉ Mono ○ Avg

Mass Tol: 0.5
◉ Th ○ ppm

Update

Peak Assignment:
◉ Most Intense
○ Nearest Match
☑ Peak Detect

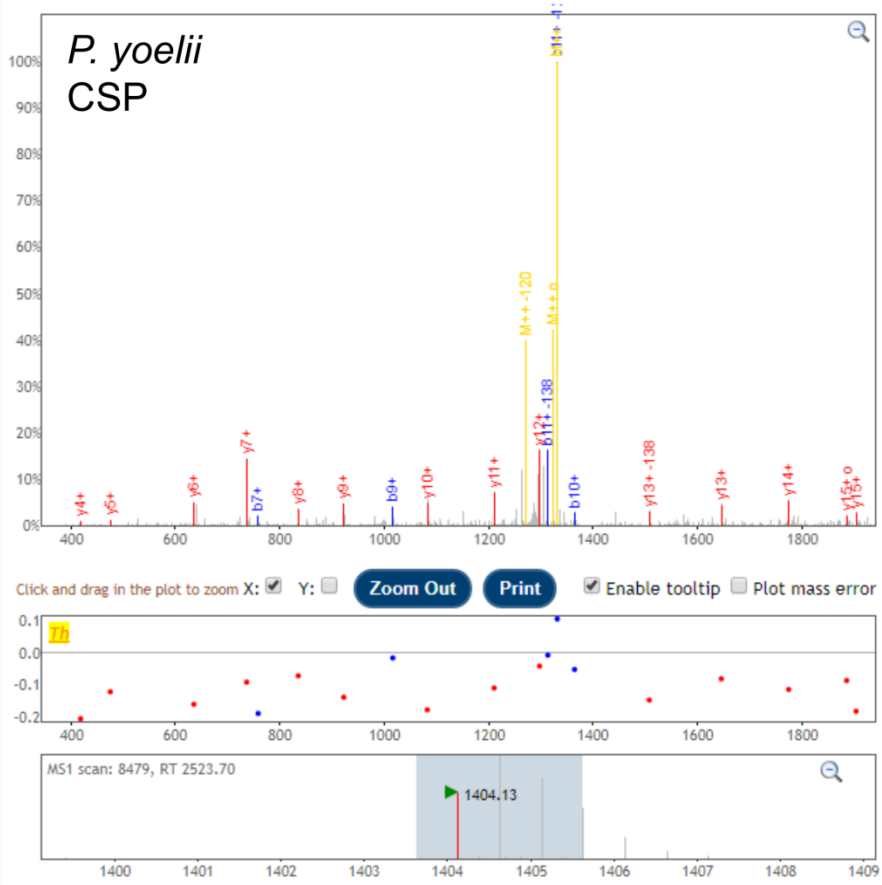
Peak Labels:
◉ Ion ○ m/z
○ None

Width: 650

Height: 400

QISSQLTEEWSQCSVTCGSGVR, MH+ 2661.1767, m/z 1331.0920

File: V2_2011_1111_KES_Pyoelii_11_2.08482.08482.2, Scan: 8482, Exp. m/z: 1404.1322, Charge: 2



Click and drag in the plot to zoom X: ☑ Y: ☐ Zoom Out Print ☑ Enable tooltip ☐ Plot mass error

b+	#	Seq	#	y+
129.0659	1	Q	22	
242.1499	2	I	21	2533.1181
329.1819	3	S	20	2420.0340
416.2140	4	S	19	2333.0020
544.2726	5	Q	18	2245.9700
657.3566	6	L	17	2117.9114
758.4043	7	T	16	2004.8273
887.4469	8	E	15	1903.7797
1016.4895	9	E	14	1774.7371
1364.6216	10	W	13	1645.6945
1451.6536	11	S	12	1297.5623
1579.7122	12	Q	11	1210.5303
1739.7429	13	C	10	1082.4717
1826.7749	14	S	9	922.4411
1925.8433	15	V	8	835.4091
2026.8910	16	T	7	736.3406
2186.9216	17	C	6	635.2930
2243.9431	18	G	5	475.2623
2330.9751	19	S	4	418.2409
2387.9966	20	G	3	331.2088
2487.0650	21	V	2	274.1874
	22	R	1	175.1190

[Click] to move table

Variable Modifications:
C: 57.021464 [13, 17]
W: 162.052824 [10]

Ions:
a ☐ 1+ ☐ 2+ ☐ 3+
b ☑ 1+ ☐ 2+ ☐ 3+
c ☐ 1+ ☐ 2+ ☐ 3+
x ☐ 1+ ☐ 2+ ☐ 3+
y ☑ 1+ ☐ 2+ ☐ 3+
z ☐ 1+ ☐ 2+ ☐ 3+
[Deselect All]

Neutral Loss:
☐ NH_3 (*)
☑ H_2O (o)
☐ H_3PO_4 (p)
☑ $C_4H_8O_4$ (-120)
☐ Immonium ions
☐ Reporter ions
☑ Precursor ions

Mass Type:
◉ Mono ○ Avg

Mass Tol: .4
◉ Th ○ ppm

Update

Peak Assignment:
◉ Most Intense
○ Nearest Match
☑ Peak Detect

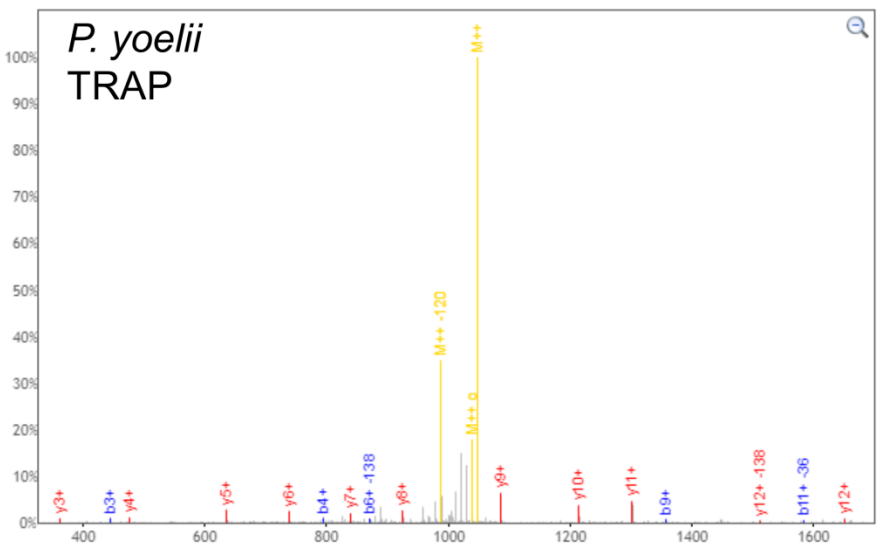
Peak Labels:
◉ Ion ○ m/z
○ None

Width: 650

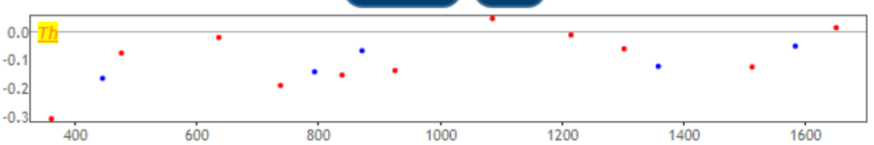
Height: 400

WEEWSECSTTCDEGR, MH+ 2093.7699, m/z 1047.3886

File: V2_2011_1111_KES_Pyoelii_07_2.05953.05953.2, Scan: 5953, Exp. m/z: 1120.4241, Charge: 2



Click and drag in the plot to zoom X: ☑ Y: ☐ Zoom Out Print ☐ Enable tooltip ☐ Plot mass error



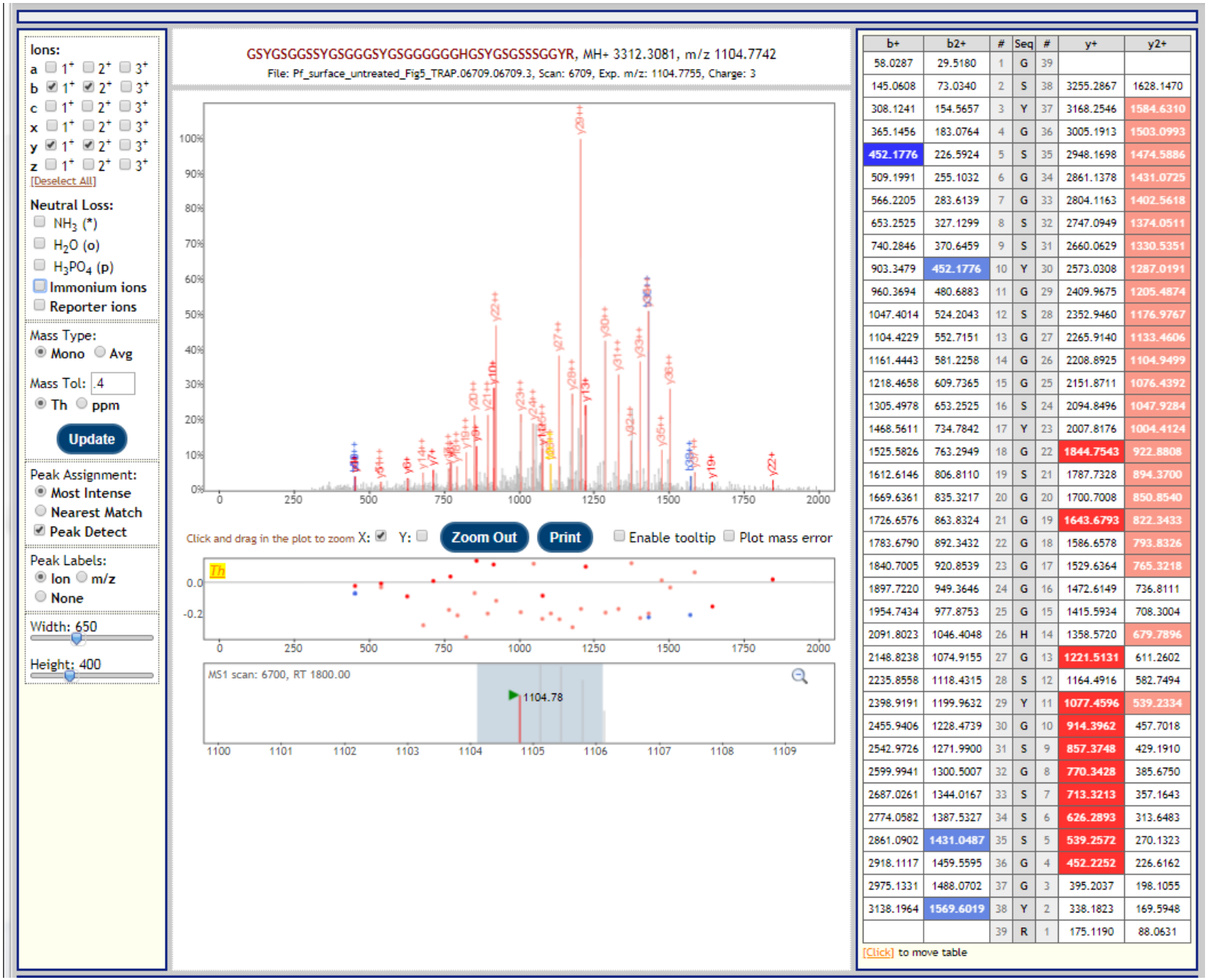
b+	#	Seq	#	y+
187.0866	1	W	15	
316.1292	2	E	14	1907.6906
445.1718	3	E	13	1778.6480
793.3039	4	W	12	1649.6054
880.3359	5	S	11	1301.4733
1009.3785	6	E	10	1214.4412
1169.4092	7	C	9	1085.3986
1256.4412	8	S	8	925.3680
1357.4889	9	T	7	838.3360
1458.5366	10	T	6	737.2883
1618.5672	11	C	5	636.2406
1733.5942	12	D	4	476.2100
1862.6368	13	E	3	361.1830
1919.6582	14	G	2	232.1404
	15	R	1	175.1190

[Click] to move table

Variable Modifications:
C: 57.021464 [7, 11]
W: 162.052824 [4]

Figure S14. Automated identification of TSR glycosylation in *P. yoelii* salivary gland sporozoites. A previously published proteomic analysis of salivary gland sporozoites of the rodent-infective *Plasmodium* species *P. yoelii* (Lindner and Swearingen *et al*. 2013 PMID:23325771) was re-analyzed with the mass offset approach described here, providing the first reported evidence of O-fucosylation and C-mannosylation of TSR domains in that species. Sample MS^2 spectra for the CSP peptide (top) and the TRAP peptide (bottom) are shown as automatically annotated by the Lorikeet spectrum viewer incorporated into the TPP. Both examples show a mass difference of -146.06 Da between the observed precursor *m/z* and that of the matched peptide, corresponding to neutral loss of O-Fuc. The C-mannosylated peptide precursor after neutral loss of O-Fuc is the dominant peak in both spectra (yellow peak annotated M++. In the CSP spectrum, this peak is isobaric with the b11-120 Da peak). Neutral loss of 120.04 Da from the M++ peak due to cross-ring cleavage of C-Man is also indicated.

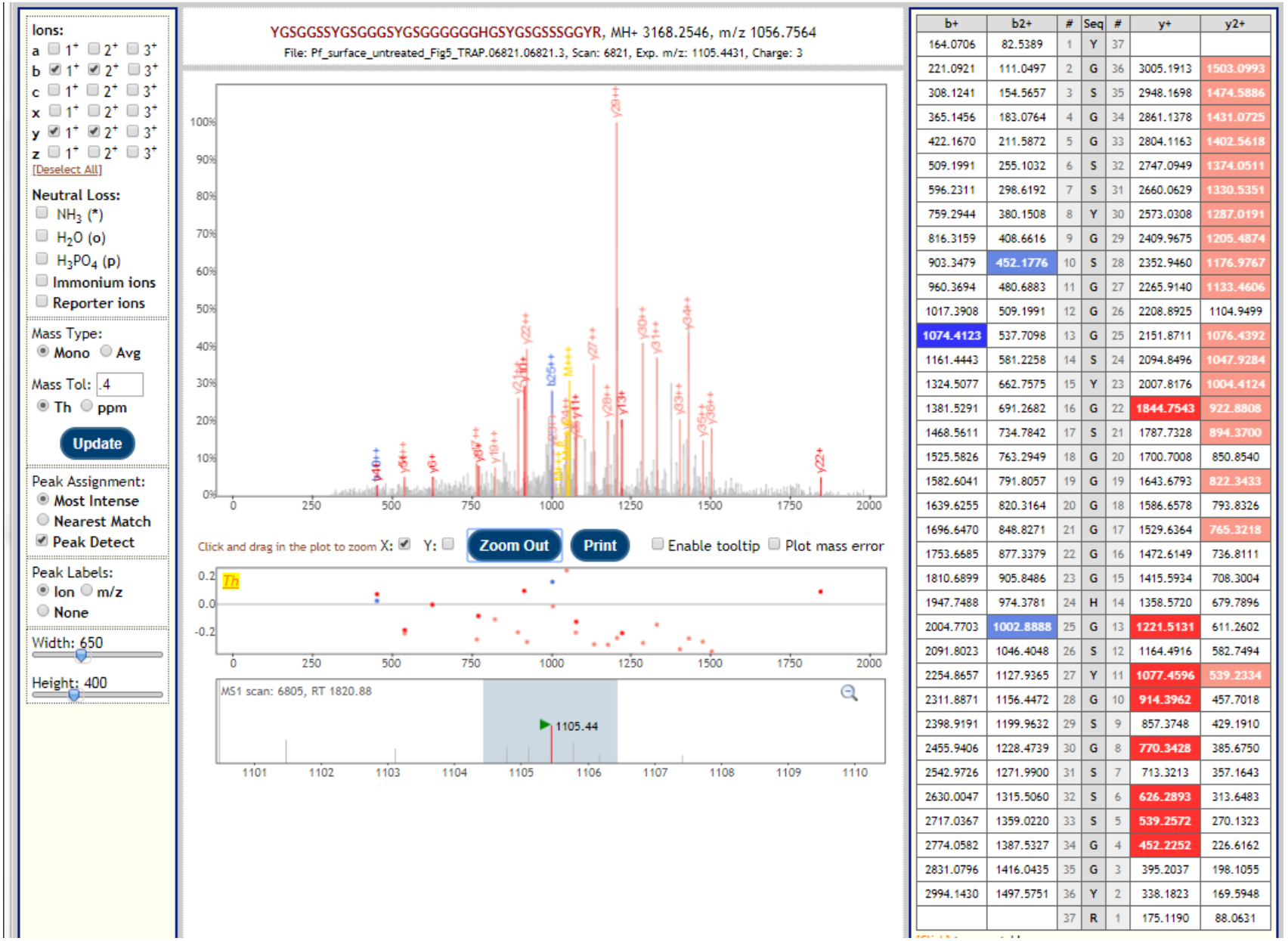
Figure S15. Example false positive peptides showing neutral loss of O-linked glycans.



b+	b2+	#	Seq	#	y+	y2+
58.0287	29.5180	1	G	39		
145.0608	73.0340	2	S	38	3255.2867	1628.1470
308.1241	154.5657	3	Y	37	3168.2546	1584.6310
365.1456	183.0764	4	G	36	3005.1913	1503.0993
452.1776	226.5924	5	S	35	2948.1698	1474.5886
509.1991	255.1032	6	G	34	2861.1378	1431.0725
566.2205	283.6139	7	G	33	2804.1163	1402.5618
653.2525	327.1299	8	S	32	2747.0949	1374.0511
740.2846	370.6459	9	S	31	2660.0629	1330.5351
903.3479	452.1776	10	Y	30	2573.0308	1287.0191
960.3694	480.6883	11	G	29	2409.9675	1205.4874
1047.4014	524.2043	12	S	28	2352.9460	1176.9767
1104.4229	552.7151	13	G	27	2265.9140	1133.4606
1161.4443	581.2258	14	G	26	2208.8925	1104.9499
1218.4658	609.7365	15	G	25	2151.8711	1076.4392
1305.4978	653.2525	16	S	24	2094.8496	1047.9284
1468.5611	734.7842	17	Y	23	2007.8176	1004.4124
1525.5826	763.2949	18	G	22	1844.7543	922.8808
1612.6146	806.8110	19	S	21	1787.7328	894.3700
1669.6361	835.3217	20	G	20	1700.7008	850.8540
1726.6576	863.8324	21	G	19	1643.6793	822.3433
1783.6790	892.3432	22	G	18	1586.6578	793.8326
1840.7005	920.8539	23	G	17	1529.6364	765.3218
1897.7220	949.3646	24	G	16	1472.6149	736.8111
1954.7434	977.8753	25	G	15	1415.5934	708.3004
2091.8023	1046.4048	26	H	14	1358.5720	679.7896
2148.8238	1074.9155	27	G	13	1221.5131	611.2602
2235.8558	1118.4315	28	S	12	1164.4916	582.7494
2398.9191	1199.9632	29	Y	11	1077.4596	539.2334
2455.9406	1228.4739	30	G	10	914.3962	457.7018
2542.9726	1271.9900	31	S	9	857.3748	429.1910
2599.9941	1300.5007	32	G	8	770.3428	385.6750
2687.0261	1344.0167	33	S	7	713.3213	357.1643
2774.0582	1387.5327	34	S	6	626.2893	313.6483
2861.0902	1431.0487	35	S	5	539.2572	270.1323
2918.1117	1459.5595	36	G	4	452.2252	226.6162
2975.1331	1488.0702	37	G	3	395.2037	198.1055
3138.1964	1569.6019	38	Y	2	338.1823	169.5948
		39	R	1	175.1190	88.0631

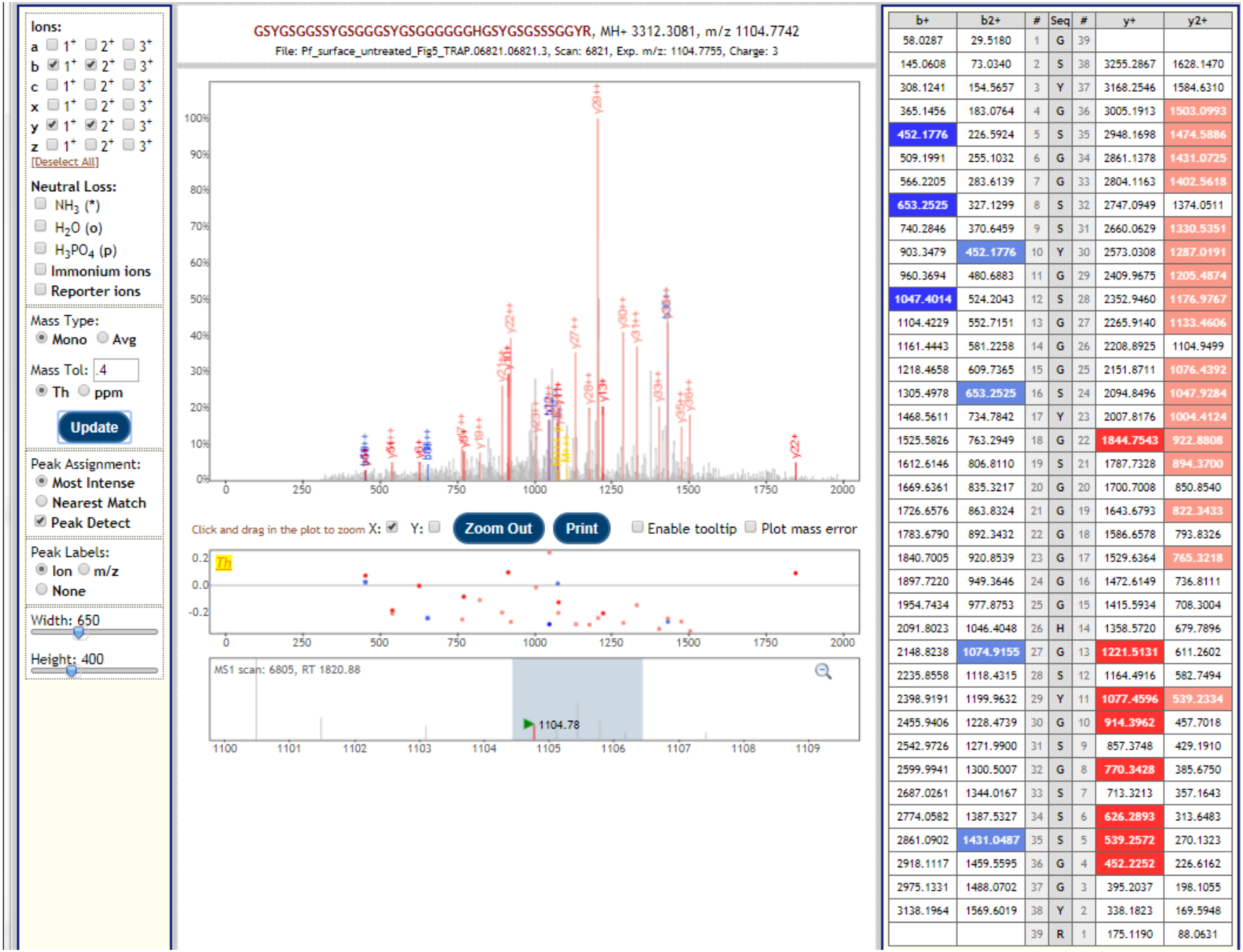
[Click] to move table

Figure S15a. A sample high-quality PSM identified the fully tryptic human keratin peptide GSYGSGGSSYGSGGGSYGSGGGGGGHGSYGSGSSSGGYR.



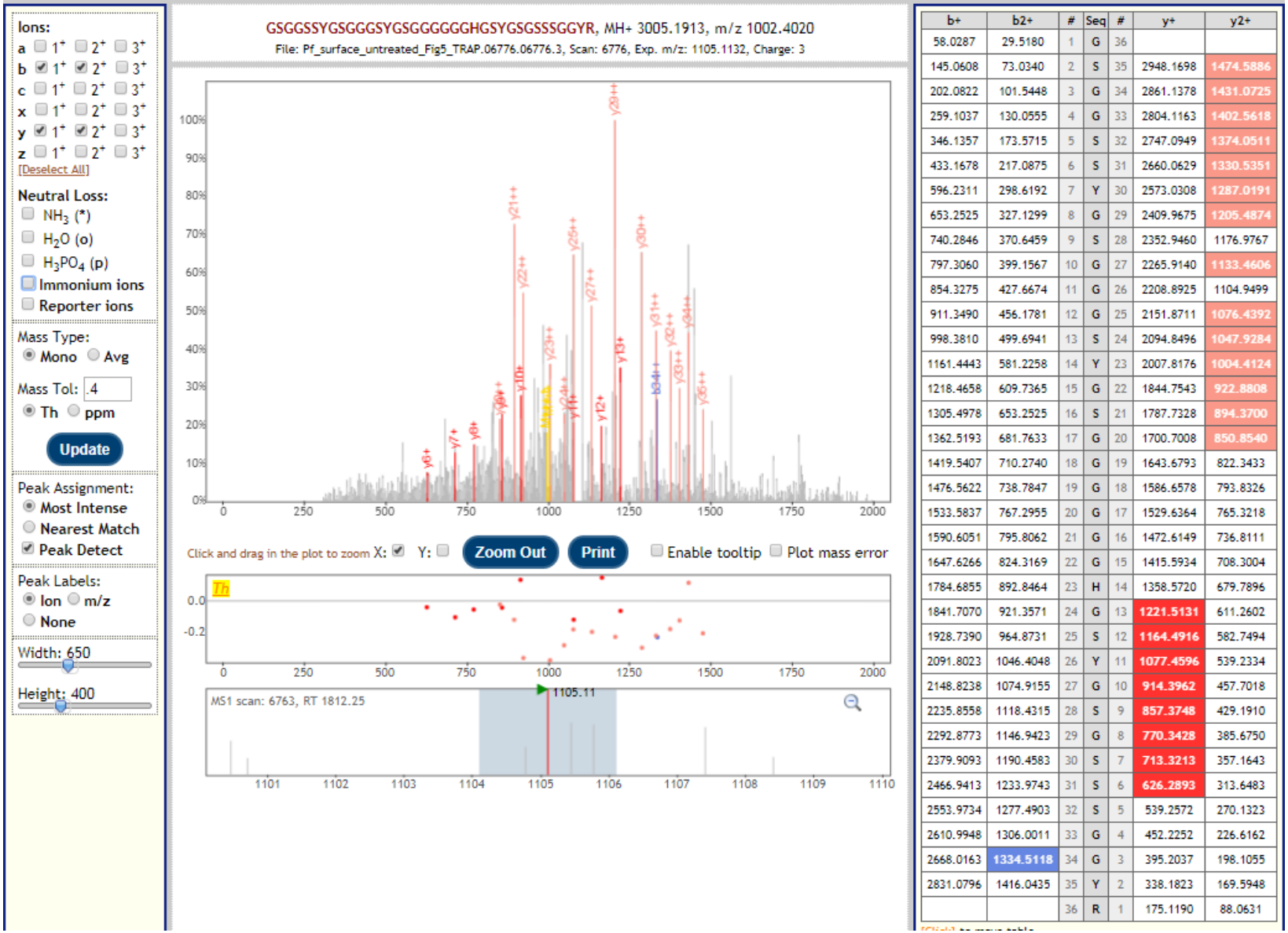
b+	b2+	#	Seq	#	y+	y2+
164.0706	82.5389	1	Y	37		
221.0921	111.0497	2	G	36	3005.1913	1503.0993
308.1241	154.5657	3	S	35	2948.1698	1474.5886
365.1456	183.0764	4	G	34	2861.1378	1431.0725
422.1670	211.5872	5	G	33	2804.1163	1402.5618
509.1991	255.1032	6	S	32	2747.0949	1374.0511
596.2311	298.6192	7	S	31	2660.0629	1330.5351
759.2944	380.1508	8	Y	30	2573.0308	1287.0191
816.3159	408.6616	9	G	29	2409.9675	1205.4874
903.3479	452.1776	10	S	28	2352.9460	1176.9767
960.3694	480.6883	11	G	27	2265.9140	1133.4606
1017.3908	509.1991	12	G	26	2208.8925	1104.9499
1074.4123	537.7098	13	G	25	2151.8711	1076.4392
1161.4443	581.2258	14	S	24	2094.8496	1047.9284
1324.5077	662.7575	15	Y	23	2007.8176	1004.4124
1381.5291	691.2682	16	G	22	1844.7543	922.8808
1468.5611	734.7842	17	S	21	1787.7328	894.3700
1525.5826	763.2949	18	G	20	1700.7008	850.8540
1582.6041	791.8057	19	G	19	1643.6793	822.3433
1639.6255	820.3164	20	G	18	1586.6578	793.8326
1696.6470	848.8271	21	G	17	1529.6364	765.3218
1753.6685	877.3379	22	G	16	1472.6149	736.8111
1810.6899	905.8486	23	G	15	1415.5934	708.3004
1947.7488	974.3781	24	H	14	1358.5720	679.7896
2004.7703	1002.8888	25	G	13	1221.5131	611.2602
2091.8023	1046.4048	26	S	12	1164.4916	582.7494
2254.8657	1127.9365	27	Y	11	1077.4596	539.2334
2311.8871	1156.4472	28	G	10	914.3962	457.7018
2398.9191	1199.9632	29	S	9	857.3748	429.1910
2455.9406	1228.4739	30	G	8	770.3428	385.6750
2542.9726	1271.9900	31	S	7	713.3213	357.1643
2630.0047	1315.5060	32	S	6	626.2893	313.6483
2717.0367	1359.0220	33	S	5	539.2572	270.1323
2774.0582	1387.5327	34	G	4	452.2252	226.6162
2831.0796	1416.0435	35	G	3	395.2037	198.1055
2994.1430	1497.5751	36	Y	2	338.1823	169.5948
		37	R	1	175.1190	88.0631

Figure S15b. A representative PSM identifying the same keratin peptide seen in Figure S15a, but spuriously identifying it as a semi-tryptic fragment with neutral loss of 146.06. The semi-tryptic fragment GS.YGSGGSSYGSGGGSYGSGGGGGGHGSYGSGSSSGGYR has a mass loss of Gly-Ser = 144.05 Da. In addition, the +2H isotope peak 1105.44 *m/z* was incorrectly selected as the monoisotopic peak for this peptide. The total mass loss is then 144.05 +2*1.0078 = 146.07, which is within the matching tolerance of the searched-for neutral loss of 146.06 Da for putative loss of O-Fuc. Because of the extensive y-ion coverage, the PSM still received a high score. The more likely correct annotation is shown in Figure S15c.



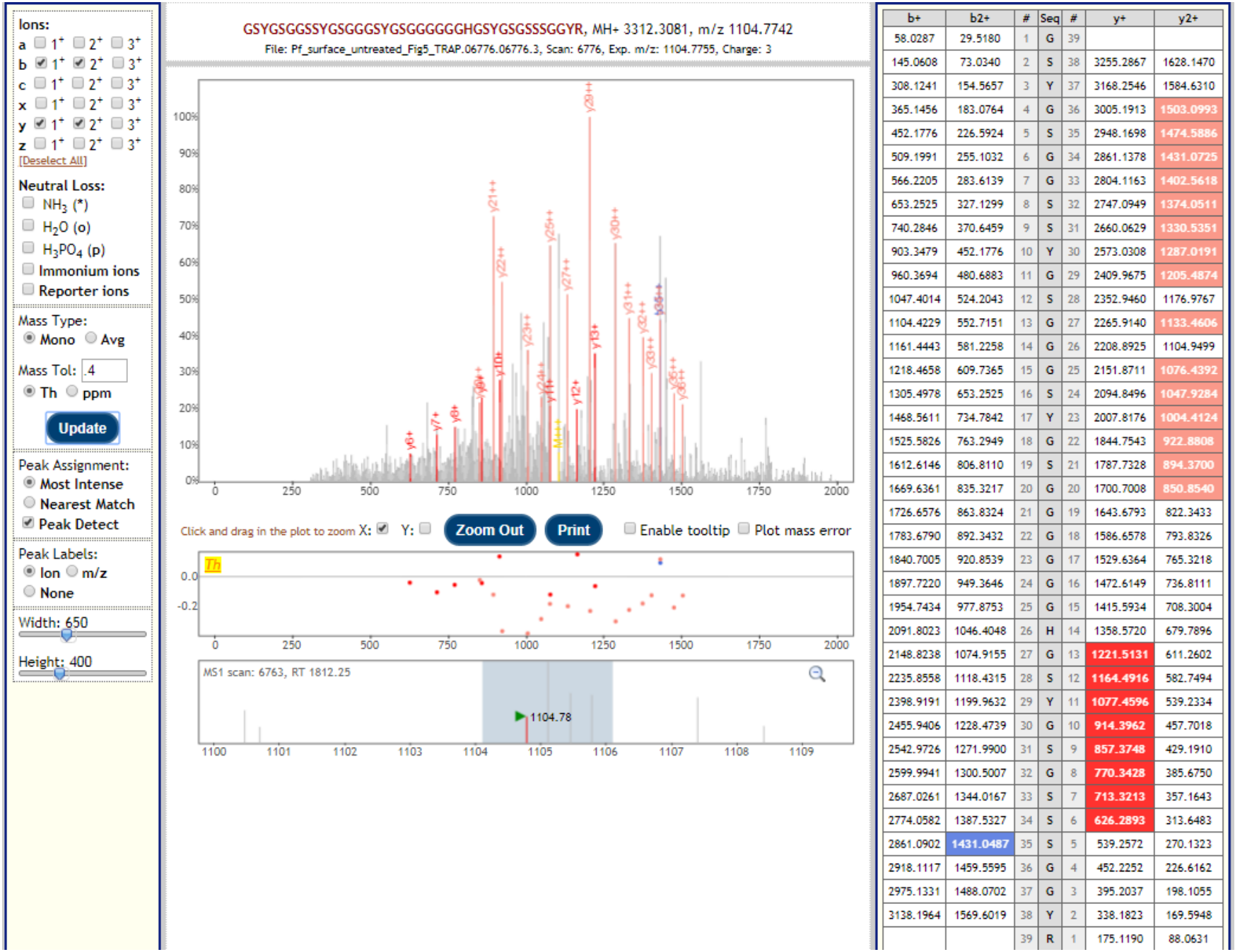
b+	b2+	#	Seq	#	y+	y2+
58.0287	29.5180	1	G	39		
145.0608	73.0340	2	S	38	3255.2867	1628.1470
308.1241	154.5657	3	Y	37	3168.2546	1584.6310
365.1456	183.0764	4	G	36	3005.1913	1503.0993
452.1776	226.5924	5	S	35	2948.1698	1474.5886
509.1991	255.1032	6	G	34	2861.1378	1431.0725
566.2205	283.6139	7	G	33	2804.1163	1402.5618
653.2525	327.1299	8	S	32	2747.0949	1374.0511
740.2846	370.6459	9	S	31	2660.0629	1330.5351
903.3479	452.1776	10	Y	30	2573.0308	1287.0191
960.3694	480.6883	11	G	29	2409.9675	1205.4874
1047.4014	524.2043	12	S	28	2352.9460	1176.9767
1104.4229	552.7151	13	G	27	2265.9140	1133.4606
1161.4443	581.2258	14	G	26	2208.8925	1104.9499
1218.4658	609.7365	15	G	25	2151.8711	1076.4392
1305.4978	653.2525	16	S	24	2094.8496	1047.9284
1468.5611	734.7842	17	Y	23	2007.8176	1004.4124
1525.5826	763.2949	18	G	22	1844.7543	922.8808
1612.6146	806.8110	19	S	21	1787.7328	894.3700
1669.6361	835.3217	20	G	20	1700.7008	850.8540
1726.6576	863.8324	21	G	19	1643.6793	822.3433
1783.6790	892.3432	22	G	18	1586.6578	793.8326
1840.7005	920.8539	23	G	17	1529.6364	765.3218
1897.7220	949.3646	24	G	16	1472.6149	736.8111
1954.7434	977.8753	25	G	15	1415.5934	708.3004
2091.8023	1046.4048	26	H	14	1358.5720	679.7896
2148.8238	1074.9155	27	G	13	1221.5131	611.2602
2235.8558	1118.4315	28	S	12	1164.4916	582.7494
2398.9191	1199.9632	29	Y	11	1077.4596	539.2334
2455.9406	1228.4739	30	G	10	914.3962	457.7018
2542.9726	1271.9900	31	S	9	857.3748	429.1910
2599.9941	1300.5007	32	G	8	770.3428	385.6750
2687.0261	1344.0167	33	S	7	713.3213	357.1643
2774.0582	1387.5327	34	S	6	626.2893	313.6483
2861.0902	1431.0487	35	S	5	539.2572	270.1323
2918.1117	1459.5595	36	G	4	452.2252	226.6162
2975.1331	1488.0702	37	G	3	395.2037	198.1055
3138.1964	1569.6019	38	Y	2	338.1823	169.5948
		39	R	1	175.1190	88.0631

Figure S15c. The more likely correct annotation of the peptide from Figure S15b, matched to the fully tryptic peptide with no neutral loss.



b+	b2+	#	Seq	#	y+	y2+
58.0287	29.5180	1	G	36		
145.0608	73.0340	2	S	35	2948.1698	1474.5886
202.0822	101.5448	3	G	34	2861.1378	1431.0725
259.1037	130.0555	4	G	33	2804.1163	1402.5618
346.1357	173.5715	5	S	32	2747.0949	1374.0511
433.1678	217.0875	6	S	31	2660.0629	1330.5351
596.2311	298.6192	7	Y	30	2573.0308	1287.0191
653.2525	327.1299	8	G	29	2409.9675	1205.4874
740.2846	370.6459	9	S	28	2352.9460	1176.9767
797.3060	399.1567	10	G	27	2265.9140	1133.4606
854.3275	427.6674	11	G	26	2208.8925	1104.9499
911.3490	456.1781	12	G	25	2151.8711	1076.4392
998.3810	499.6941	13	S	24	2094.8496	1047.9284
1161.4443	581.2258	14	Y	23	2007.8176	1004.4124
1218.4658	609.7365	15	G	22	1844.7543	922.8808
1305.4978	653.2525	16	S	21	1787.7328	894.3700
1362.5193	681.7633	17	G	20	1700.7008	850.8540
1419.5407	710.2740	18	G	19	1643.6793	822.3433
1476.5622	738.7847	19	G	18	1586.6578	793.8326
1533.5837	767.2955	20	G	17	1529.6364	765.3218
1590.6051	795.8062	21	G	16	1472.6149	736.8111
1647.6266	824.3169	22	G	15	1415.5934	708.3004
1784.6855	892.8464	23	H	14	1358.5720	679.7896
1841.7070	921.3571	24	G	13	1221.5131	611.2602
1928.7390	964.8731	25	S	12	1164.4916	582.7494
2091.8023	1046.4048	26	Y	11	1077.4596	539.2334
2148.8238	1074.9155	27	G	10	914.3962	457.7018
2235.8558	1118.4315	28	S	9	857.3748	429.1910
2292.8773	1146.9423	29	G	8	770.3428	385.6750
2379.9093	1190.4583	30	S	7	713.3213	357.1643
2466.9413	1233.9743	31	S	6	626.2893	313.6483
2553.9734	1277.4903	32	S	5	539.2572	270.1323
2610.9948	1306.0011	33	G	4	452.2252	226.6162
2668.0163	1334.5118	34	G	3	395.2037	198.1055
2831.0796	1416.0435	35	Y	2	338.1823	169.5948
		36	R	1	175.1190	88.0631

Figure S15d. A representative PSM identifying the same keratin peptide seen in Figure S15a, but spuriously identifying it as a semi-tryptic fragment with neutral loss of 308.11. The semi-tryptic fragment GSY.GSGGSSYGSGGGSYGSGGGGGGHGSYGSGSSSGGYR has a mass loss of Gly-Ser-Tyr = 307.12 Da. In addition, the +1H isotope peak 1105.11 *m/z* was incorrectly selected as the monoisotopic peak for this peptide. The total mass loss is then 307.12 +1.0078 = 308.12, which is within the matching tolerance of the searched-for neutral loss of 308.11 Da for putative loss of O-Fuc-Glc. Because of the extensive y-ion coverage, the PSM still received a high score. The correct annotation is shown in Figure S15e.



b+	b2+	#	Seq	#	y+	y2+
58.0287	29.5180	1	G	39		
145.0608	73.0340	2	S	38	3255.2867	1628.1470
308.1241	154.5657	3	Y	37	3168.2546	1584.6310
365.1456	183.0764	4	G	36	3005.1913	1503.0993
452.1776	226.5924	5	S	35	2948.1698	1474.5886
509.1991	255.1032	6	G	34	2861.1378	1431.0725
566.2205	283.6139	7	G	33	2804.1163	1402.5618
653.2525	327.1299	8	S	32	2747.0949	1374.0511
740.2846	370.6459	9	S	31	2660.0629	1330.5351
903.3479	452.1776	10	Y	30	2573.0308	1287.0191
960.3694	480.6883	11	G	29	2409.9675	1205.4874
1047.4014	524.2043	12	S	28	2352.9460	1176.9767
1104.4229	552.7151	13	G	27	2265.9140	1133.4606
1161.4443	581.2258	14	G	26	2208.8925	1104.9499
1218.4658	609.7365	15	G	25	2151.8711	1076.4392
1305.4978	653.2525	16	S	24	2094.8496	1047.9284
1468.5611	734.7842	17	Y	23	2007.8176	1004.4124
1525.5826	763.2949	18	G	22	1844.7543	922.8808
1612.6146	806.8110	19	S	21	1787.7328	894.3700
1669.6361	835.3217	20	G	20	1700.7008	850.8540
1726.6576	863.8324	21	G	19	1643.6793	822.3433
1783.6790	892.3432	22	G	18	1586.6578	793.8326
1840.7005	920.8539	23	G	17	1529.6364	765.3218
1897.7220	949.3646	24	G	16	1472.6149	736.8111
1954.7434	977.8753	25	G	15	1415.5934	708.3004
2091.8023	1046.4048	26	H	14	1358.5720	679.7896
2148.8238	1074.9155	27	G	13	1221.5131	611.2602
2235.8558	1118.4315	28	S	12	1164.4916	582.7494
2398.9191	1199.9632	29	Y	11	1077.4596	539.2334
2455.9406	1228.4739	30	G	10	914.3962	457.7018
2542.9726	1271.9900	31	S	9	857.3748	429.1910
2599.9941	1300.5007	32	G	8	770.3428	385.6750
2687.0261	1344.0167	33	S	7	713.3213	357.1643
2774.0582	1387.5327	34	S	6	626.2893	313.6483
2861.0902	1431.0487	35	S	5	539.2572	270.1323
2918.1117	1459.5595	36	G	4	452.2252	226.6162
2975.1331	1488.0702	37	G	3	395.2037	198.1055
3138.1964	1569.6019	38	Y	2	338.1823	169.5948
		39	R	1	175.1190	88.0631

Figure S15e. The more likely correct annotation of the peptide from Figure S15d, matched to the fully tryptic peptide with no neutral loss