Published in final edited form as:

J Proteome Res. 2019 February 01; 18(2): 652–663. doi:10.1021/acs.jproteome.8b00638.

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

Kristian E. Swearingen^{1,*}, Jimmy K. Eng², David Shteynberg¹, Vladimir Vigdorovich³, Timothy A. Springer⁴, Luis Mendoza¹, Noah D. Sather³, Eric W. Deutsch¹, Stefan H. I. Kappe³, and Robert L. Moritz^{1,*}

¹Institute for Systems Biology, Seattle, WA, USA

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

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

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

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- β 1,3-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 re-analyze published proteomics data from *Plasmodium* parasites and identified multiple glycoforms of TSR-containing proteins.

Keywords

O-fucosylation; C-mannosylation; thrombospondin type 1 repeat; *Plasmodium*

*To whom correspondence should be addressed. kristian.swearingen@systemsbiology.org, robert.moritz@systemsbiology.org.

INTRODUCTION

Thrombospondin type 1 repeats (TSRs) are small (~60 amino acid residues) adhesive protein domains with a wide range of functions¹. TSRs are often modified with two unique glycans, O-linked fucose (O-Fuc) and C-linked mannose (C-Man). The motif CX₂₋₃(S/T)CX₂G is recognized by the O-fucosyltransferase POFUT2² and can be modified with O-Fuc at the serine/threonine (Ser/Thr or S/T) residue³. This fucose can be further extended to a fucose- β 1,3-glucose disaccharide (O-Fuc-Glc) by the β 1,3-glucosyltransferase B3GLCT⁴⁻⁵. 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) residues⁶⁻⁷. In *C. elegans*, where the first C-mannosyltransferase was discovered⁸, the enzyme Dpy19 only C-mannosylates the first Trp of the WXXW motif. However, it has recently been shown that humans possess multiple homologs of Dpy19, and that Dpy19L1 preferentially modifies the first Trp of WXXW while Dpy19L3 preferentially modifies the Trp of WXXC⁹. In TSRs, all of these motifs often appear in tandem as [WXX]₁₋₃CXX(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 protein³. 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 discernable defect from loss of glycosylation⁹⁻¹⁰.

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 modification is attached can usually be directly inferred from the MS² 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 phase³ and are present at very low abundance or absent from the collision induced dissociation (CID) and higher-energy collision (HCD) MS² 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 MS² 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 intact¹¹⁻¹². 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 ETD¹³. 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 re-analyzed 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 elsewhere¹⁵. 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 cut-off 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 elsewhere¹⁶. Six µg of *P*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 *P/CSP* tryptic digest was infused directly at 500 nL/min through an empty Pico frit column (New Objective) with a 75 μ m internal diameter (ID) and a 15 μ m 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 *PfTRAP* 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 μ g of peptides were loaded onto a 150 μ m ID fused silica trap column made in-house with a 10 mm bed of C18 (Dr. Maisch ReproSil Pur C18 AQ, 120 Å, 3 μ m) and washed with LC load buffer (2% (v/v) ACN/0.2 % (v/v) trifluoroacetic acid (TFA)) prior to separation on a column packed in-house with a 20 cm bed of the same stationary phase in a Pico frit column with a 75 μ m ID and a 10 μ m 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 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 Comet¹⁷ 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 Syzygy¹⁸. 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 dataset 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 non-zero 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.org¹⁹ 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 toolkit²³ and searched with Comet version 2017.01 rev.1¹⁷. The precursor mass tolerance was set to ± 10 ppm. For the recombinant *PT*TRAP LC/MS data, for which the MS² 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 to 0 (“use flanking peaks”). For the sporozoite global proteome data, for which the MS² 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”). Semi-trypsin 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* 3D7²⁴ (PlasmoDB v.35, www.plasmodb.org²⁵, 5548 entries), *P. yoelii yoelii* 17X²⁶ (PlasmoDB v.35, 6092 entries), or *P. vivax* P01²⁷ (PlasmoDB v.31) appended with sequence polymorphisms observed in other field isolates²² (6670 entries), as appropriate for the sample. Each of these *Plasmodium* databases was appended with the mosquito protein database *Anopheles stephensi* Indian AsteI2.3²⁸ (VectorBase, www.vectorbase.org²⁹, 11789 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 MS² spectra were analyzed using the TPP version 5.1.0 Syzygy. PSMs were assigned scores in PeptideProphet³⁰ 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 re-analysis of sporozoite proteome datasets, the DECOYPROBS flag was used to assign possible non-zero 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 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 nano-flow 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. Based on signal intensity of the parent ions, ~92% of glycosite-bearing peptide was modified with a mass equal to hexose plus deoxyhexose ($[M+2H]^{2+} = 1451.17 \text{ m/z}$), 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 (Fig 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 ($[M+2H]^{2+} = 1297.11 \text{ m/z}$) as well as lower-intensity doubly charged species matching the mass of the peptide modified with a single deoxyhexose (i.e., fucose, $[M+2H]^{2+} = 1370.14 \text{ m/z}$). MS²,

MS³ and MS⁴ of these species confirmed the sequential loss of Glc and Fuc as well as the sequence of the peptide (Fig 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 MS² spectrum effectively identifying the unmodified peptide. 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 ($[M+2H]^{2+} = 1329.03 \text{ m/z}$). The precursor ion was fragmented at incrementally increasing CEs as above (Fig 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 (Fig 2). Unlike O-linked glycans, C-Man withstands CID and is 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 peaks³. Notably, two isobaric 1329.03 *m/z* species were present that had distinct LC retention times (Fig 2A), and had the C-Man attached at the first or the second Trp residue of the WXXW motif (Fig 2B and 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 MS² spectra of such peptides can confidently identify the sequence of the unmodified peptide, we reasoned that these peptides could be identified by a sequence 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 Comet¹⁷ 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 ($[M+2H]^{2+} = 1451.17 \text{ m/z}$). By 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 ($[M+2H]^{2+} = 1297.11 \text{ m/z}$), matching the mass of the peptide identified by the unmodified peptide fragment ions in the MS² 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 Fig 3.

To demonstrate this modified search approach, the recombinant *PTRAP* 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. MS² 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 (Fig 4). For example, each of the two positional isoforms of the singly C-mannosylated peptide ($[M + 2H]^{2+} = 1174.98 \text{ m/z}$) co-eluted with its higher-abundance, higher-mass counterpart ($[M + 2H]^{2+} = 1329.03 \text{ m/z}$) 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 MS². 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 MS² and retention times (Table S2, Fig 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 seven-fold greater than any other species. The glycopeptide was identified from multiple high-quality spectra, as was the co-eluting peptide that had lost the disaccharide due to in-source fragmentation. Interestingly, another co-eluting pair of peptides with identical respective masses was identified at 57.8 min. The MS² spectra for these isobars were indistinguishable from their more abundant, later-eluting counterparts (Fig 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, MS² spectra, and co-eluting 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 MS² 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 MS² from the same *m/z* and retention time correctly identified the singly mannosylated glycoform (Fig 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 re-searched for evidence of O-fucosylation, even if the MS² data were collected at low resolution (e.g. using an ion trap). To demonstrate this, we re-analyzed 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 POFUT2¹⁰. It has also been confirmed that *P. falciparum* possesses a version of the Dpy19 C-mannosyltransferase capable of modifying the TSR domain of *PTTRAP* with C-Man³⁵. 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 re-analyzed those datasets 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 MS¹ and low resolution MS². The *P. falciparum* data, which includes a global proteome analysis²⁰ and a biotinylated surface-enriched sample²¹, were part of the original evidence presented for O-fucosylation of TSRs in plasmodium²¹, and the *P. vivax* dataset subsequently demonstrated TSR glycosylation in that species as well²². In those analyses, select MS² spectra from parent ions matching predicted masses were manually annotated to provide evidence of O-fucosylation and C-mannosylation. Re-analyzing 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 (Fig 5, Table 2, Table S3, Fig S6-S14). Re-analyzing the *P. yoelii* data resulted in the first reported evidence of glycosylation of TSRs in that species (Fig 4, Fig 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 (Fig S12, Fig 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 *PT*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 motif³⁵.

For all of the *Plasmodium* salivary gland sporozoite datasets 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 datasets ranged from 0.26 % to 0.56 %. Among the PSMs identified above this cut-off 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 datasets, were semi-tryptic 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 (Fig 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 MS² 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 datasets 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 (Fig 5). Further work will be required to determine whether the failure to detect the disaccharide in the *P.*

vivax and *P. yoelii* datasets is due to instrumental detection limit or the true absence of the glycan. In other organisms this glycan extension is performed by β -1,3-glucosyltransferase (B3GLCT). Although there is a *P. falciparum* gene with some sequence homology to B3GLCT²⁰ (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 Fig 5). The TSR domains of both *PI*CSP³⁶ and *PI*TRAP³⁷ are part of subunit malaria vaccine candidates³⁸. Since glycosylation effectively changes antibody recognition epitopes³⁹, 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 MS² spectra, including neutral losses from cross-ring cleavage of C-Man. We used this search strategy to re-analyze published proteomics datasets, 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 re-analyze existing datasets for evidence of glycosylation that were not previously searched for.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

FUNDING SOURCES

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 R01CA031798 (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 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	Beta-1,3-Glucosyltransferase

REFERENCES

1. Tucker RP, The thrombospondin type 1 repeat superfamily. *Int J Biochem Cell Biol* 2004, 36 (6), 969–74. [PubMed: 15094110]
2. Luo Y; Koles K; Vorndam W; Haltiwanger RS; Panin VM, Protein O-fucosyltransferase 2 adds O-fucose to thrombospondin type 1 repeats. *J Biol Chem* 2006, 281 (14), 9393–9. [PubMed: 16464857]
3. Hofsteenge J; Huwiler KG; Macek B; Hess D; Lawler J; Mosher DF; Peter-Katalinic J, C-mannosylation and O-fucosylation of the thrombospondin type 1 module. *J Biol Chem* 2001, 276 (9), 6485–98. [PubMed: 11067851]
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. [PubMed: 16899492]
5. Kozma K; Keusch JJ; Hegemann B; Luther KB; Klein D; Hess D; Haltiwanger RS; 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. [PubMed: 17032646]
6. Hofsteenge J; Muller DR; de Beer T; Löffler A; Richter WJ; Vliegenthart JF, 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. [PubMed: 7947762]
7. Julenius K, NetCGlyc 1.0: prediction of mammalian C-mannosylation sites. *Glycobiology* 2007, 17 (8), 868–76. [PubMed: 17494086]
8. Buettner FF; 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. [PubMed: 23562325]
9. Shcherbakova A; Tiemann B; Buettner FF; 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. [PubMed: 28202721]
10. Lopatnicki S; Yang ASP; John A; Scott NE; Lingford JP; O'Neill MT; Erickson SM; McKenzie NC; Jennison C; Whitehead LW; Douglas DN; Kneteman NM; Goddard-Borger ED; Boddey JA, Protein O-fucosylation in Plasmodium falciparum ensures efficient infection of mosquito and vertebrate hosts. *Nat Commun* 2017, 8 (1), 561. [PubMed: 28916755]
11. Bandini G; Haserick JR; Motari E; Ouologuem DT; Lourido S; Roos DS; Costello CE; Robbins PW; 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. [PubMed: 27663739]
12. Sorvillo N; Kaijen PH; Matsumoto M; Fujimura Y; van der Zwaan C; Verbij FC; Pos W; Fijnheer R; Voorberg J; Meijer AB, Identification of N-linked glycosylation and putative O-fucosylation, C-mannosylation sites in plasma derived ADAMTS13. *J Thromb Haemost* 2014, 12 (5), 670–9. [PubMed: 24977290]
13. Swaney DL; McAlister GC; Coon JJ, Decision tree-driven tandem mass spectrometry for shotgun proteomics. *Nat Methods* 2008, 5 (11), 959–64. [PubMed: 18931669]

14. Kakuda S; Haltiwanger RS, Analyzing the posttranslational modification status of Notch using mass spectrometry. *Methods Mol Biol* 2014, 1187, 209–21. [PubMed: 25053492]
15. Doud MB; Koksai AC; Mi LZ; Song G; Lu C; Springer TA, Unexpected fold in the circumsporozoite protein target of malaria vaccines. *Proc Natl Acad Sci U S A* 2012, 109 (20), 7817–22. [PubMed: 22547819]
16. Carbonetti S; Oliver BG; Vigdorovich V; Dambrauskas N; Sack B; Bergl E; Kappe SHI; Sather DN, A method for the isolation and characterization of functional murine monoclonal antibodies by single B cell cloning. *J Immunol Methods* 2017, 448, 66–73. [PubMed: 28554543]
17. Eng JK; Jahan TA; Hoopmann MR, Comet: an open-source MS/MS sequence database search tool. *Proteomics* 2013, 13 (1), 22–4. [PubMed: 23148064]
18. Deutsch EW; Mendoza L; Shteynberg D; Slagel J; Sun Z; Moritz RL, Trans-Proteomic Pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. *Proteomics Clin Appl* 2015, 9 (7–8), 745–54. [PubMed: 25631240]
19. Kusebauch U; Deutsch EW; Campbell DS; Sun Z; Farrah T; Moritz RL, Using PeptideAtlas, SRMATlas, and PASSEL: Comprehensive Resources for Discovery and Targeted Proteomics. *Curr Protoc Bioinformatics* 2014, 46, 13 25 1–28.
20. Lindner SE; Swearingen KE; Harupa A; Vaughan AM; Sinnis P; Moritz RL; Kappe SH, Total and putative surface proteomics of malaria parasite salivary gland sporozoites. *Mol Cell Proteomics* 2013, 12 (5), 1127–43. [PubMed: 23325771]
21. Swearingen KE; Lindner SE; Shi L; Shears MJ; Harupa A; Hopp CS; Vaughan AM; Springer TA; Moritz RL; Kappe SH; 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), e1005606. [PubMed: 27128092]
22. Swearingen KE; Lindner SE; Flannery EL; Vaughan AM; Morrison RD; Patrapuvich R; Koepfli C; Muller I; Jex A; Moritz RL; Kappe SHI; Sattabongkot J; Mikolajczak SA, Proteogenomic analysis of the total and surface-exposed proteomes of Plasmodium vivax salivary gland sporozoites. *PLoS Negl Trop Dis* 2017, 11 (7), e0005791. [PubMed: 28759593]
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. [PubMed: 18606607]
24. Gardner MJ; Hall N; Fung E; White O; Berriman M; Hyman RW; Carlton JM; Pain A; Nelson KE; Bowman S; Paulsen IT; James K; Eisen JA; Rutherford K; Salzberg SL; Craig A; Kyes S; Chan MS; Nene V; Shallom SJ; Suh B; Peterson J; Angiuoli S; Pertea M; Allen J; Selengut J; Haft D; Mather MW; Vaidya AB; Martin DM; Fairlamb AH; Fraunholz MJ; Roos DS; Ralph SA; McFadden GI; Cummings LM; Subramanian GM; Mungall C; Venter JC; Carucci DJ; Hoffman SL; Newbold C; Davis RW; Fraser CM; Barrell B, Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* 2002, 419 (6906), 498–511. [PubMed: 12368864]
25. Aureocoechea C; Brestelli J; Brunk BP; Dommer J; Fischer S; Gajria B; Gao X; Gingle A; Grant G; Harb OS; Heiges M; Innamorato F; Iodice J; Kissinger JC; Kraemer E; Li W; Miller JA; Nayak V; Pennington C; Pinney DF; Roos DS; Ross C; Stoeckert CJ, Jr.; Treatman C; Wang H, PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids Res* 2009, 37 (Database issue), D539–43. [PubMed: 18957442]
26. Otto TD; Bohme U; Jackson AP; Hunt M; Franke-Fayard B; Hoeijmakers WA; Religa AA; Robertson L; Sanders M; Ogun SA; Cunningham D; Erhart A; Billker O; Khan SM; Stunnenberg HG; Langhorne J; Holder AA; Waters AP; Newbold CI; Pain A; Berriman M; Janse CJ, A comprehensive evaluation of rodent malaria parasite genomes and gene expression. *BMC Biol* 2014, 12, 86. [PubMed: 25359557]
27. Auburn S; Bohme U; Steinbiss S; Trimarsanto H; Hostetler J; Sanders M; Gao Q; Nosten F; Newbold CI; Berriman M; Price RN; Otto TD, A new Plasmodium vivax reference sequence with improved assembly of the subtelomeres reveals an abundance of pir genes. *Wellcome Open Res* 2016, 1, 4. [PubMed: 28008421]
28. Jiang X; Peery A; Hall AB; Sharma A; Chen XG; Waterhouse RM; Komissarov A; Riehle MM; Shouche Y; Sharakhova MV; Lawson D; Pakpour N; Arensburger P; Davidson VL; Eglmeier K; Emrich S; George P; Kennedy RC; Mane SP; Maslen G; Oranganje C; Qi Y; Settlege R; Tojo M; Tubio JM; Unger MF; Wang B; Vernick KD; Ribeiro JM; James AA; Michel K; Riehle MA;

- Luckhart S; Sharakhov IV; Tu Z, Genome analysis of a major urban malaria vector mosquito, *Anopheles stephensi*. *Genome Biol* 2014, 15 (9), 459. [PubMed: 25244985]
29. Giraldo-Calderon GI; Emrich SJ; MacCallum RM; Maslen G; Dialynas E; Topalis P; Ho N; Gesing S; VectorBase C; Madey G; Collins FH; Lawson D, VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. *Nucleic Acids Res* 2015, 43 (Database issue), D707–13. [PubMed: 25510499]
30. Keller A; Nesvizhskii AI; 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. [PubMed: 12403597]
31. Song G; Koksall AC; Lu C; Springer TA, Shape change in the receptor for gliding motility in *Plasmodium* sporozoites. *Proc Natl Acad Sci U S A* 2012, 109 (52), 21420–5. [PubMed: 23236185]
32. Song G; Springer TA, Structures of the *Toxoplasma* gliding motility adhesin. *Proc Natl Acad Sci U S A* 2014, 111 (13), 4862–7. [PubMed: 24639528]
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. [PubMed: 21472528]
34. Hahne H; Gholami AM; Kuster B, Discovery of O-GlcNAc-modified proteins in published large-scale proteome data. *Mol Cell Proteomics* 2012, 11 (10), 843–50. [PubMed: 22661428]
35. Hoppe CM; Albuquerque-Wendt A; Bandini G; Leon DR; Shcherbakova A; Buettner FFR; Izquierdo L; Costello CE; Bakker H; Routier FH, Apicomplexan C-Mannosyltransferases Modify Thrombospondin Type I-containing Adhesins of the TRAP Family. *Glycobiology* 2018, 28 (5), 333–343. [PubMed: 29432542]
36. Olotu A; Fegan G; Wambua J; Nyangweso G; Leach A; Lievens M; Kaslow DC; 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. [PubMed: 27355532]
37. Moorthy VS; Imoukhuede EB; Milligan P; Bojang K; Keating S; Kaye P; Pinder M; Gilbert SC; Walraven G; Greenwood BM; Hill AS, A randomised, double-blind, controlled vaccine efficacy trial of DNA/MVA ME-TRAP against malaria infection in Gambian adults. *PLoS Med* 2004, 1 (2), e33. [PubMed: 15526058]
38. Kester KE; Gray Heppner D, Jr.; Moris P; Ofori-Anyinam O; Krzych U; Tornieporth N; McKinney D; Delchambre M; Ockenhouse CF; Voss G; Holland C; Beckey JP; Ballou WR; Cohen J; Rts STG, 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. [PubMed: 24950358]
39. Luca VC; Jude KM; Pierce NW; Nachury MV; Fischer S; Garcia KC, Structural biology. Structural basis for Notch1 engagement of Delta-like 4. *Science* 2015, 347 (6224), 847–53. [PubMed: 25700513]
40. Goddard-Borger ED; Boddey JA, Implications of *Plasmodium* glycosylation on vaccine efficacy and design. *Future Microbiol* 2018, 13, 609–612. [PubMed: 29624074]

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 four-fold, but the spectrum was still dominated by the intact peptide stripped of the glycan.

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 MS² 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.

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

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 MS² fragmentation could be seen as lower-intensity, co-eluting peaks. (B) A representative MS² 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 MS². 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.

Figure 5. *In vivo* glycosylation of *Plasmodium* TSR domains identified from re-analysis 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 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 Fig S6-13). Glycosylation of CSP and TRAP has been reported in salivary gland sporozoites of *P. falciparum* (Pf)²¹ and *P. vivax* (Pv)²². Our re-analysis of published data from *P. yoelii* (Py) salivary gland sporozoites²⁰ has provided the first evidence of TSR glycosylation in that species.

Table 1.
Glycopeptides identified from LC-MS/MS analysis of recombinant *Pf*TRAP.

A tryptic digest of recombinant *P. falciparum* TRAP expressed in mammalian cells was analyzed by LC-MS/MS and the MS² spectra were searched for evidence of O-linked fucose or glucosylfucose by our automated database searching method.

Modified Peptide ^a	O-linked Glycan ^b	[M+2H] ²⁺ <i>m/z</i> ^c	Retention time (min) ^d	Proportion in sample ^e	PSMs ^f
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

^aThe peptide from the TSR domain of *P. falciparum* TRAP containing potential glycosylation sites. The localization of C-Man was directly detected from the MS² fragment spectra. Any O-Fuc or O-Fuc-Glc is presumed to have been attached at the C-terminal Thr residue.

^bThe 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 MS² fragmentation.

^cThe *m/z* of the doubly charged peptide parent ion.

^dThe retention time of the peptide (see Fig 4).

^eThe relative proportion of the peptide in the sample based on chromatographic peak height (see Fig 4).

^fThe 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 MS² fragmentation.

Table 2.
Evidence for *in vivo* glycosylation of CSP and TRAP.

Previously published proteomics data from analyses of *Plasmodium* salivary gland sporozoites was 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).

Sample ^a	Total PSMs ^b	CSP PSMs ^c	TRAP PSMs ^d	Decoy PSMs ^e	Non-Decoy False Positives ^f	Fully Tryptic Non-Decoy False Positives ^g
<i>P. vivax</i> VK210 ²²	352	259	83	0	10	1
<i>P. vivax</i> VK247 ²²	250	171	33	0	46	0
<i>P. falciparum</i> global proteome ²⁰	57	54	0	0	3	1
<i>P. falciparum</i> surface-enriched ²¹	12	0	4	0	8	1
<i>P. yoelii</i> ²⁰	26	18	7	0	1	0

^aSources 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 dataset 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.

^bTotal number of peptide spectrum matches (PSMs) identified with a neutral loss putatively corresponding to an O-linked glycan.

^cTotal number of PSMs identifying the glycosite-containing peptide of CSP with evidence for O-fucosylation.

^dTotal number of PSMs identifying the glycosites-containing peptide of TRAP with evidence for O-fucosylation.

^eThe number of PSMs identified from decoy proteins.

^fThe number PSMs identifying likely incorrect non-decoy proteins.

^gThe number PSMs identifying fully tryptic peptides mapping to likely incorrect non-decoy proteins.