Single-subunit oligosaccharyltransferases of Trypanosoma brucei display different and predictable peptide acceptor specificities

Jinnelov, Anders A. J.; Ali, Liaqat; Tinti, Michele; Ferguson, Michael A. J.

Published in:
Journal of Biological Chemistry

DOI:
10.1074/jbc.M117.810945

Publication date:
2017

Document Version
Peer reviewed version

JBC Papers in Press. Published on September 19, 2017 as Manuscript M117.810945
The latest version is at http://www.jbc.org/cgi/doi/10.1074/jbc.M117.810945

Single-subunit oligosaccharyltransferases of *Trypanosoma brucei* display different and predictable peptide acceptor specificities.

Anders Jinnelov, Liaqat Ali, Michele Tinti and Michael A.J. Ferguson*
Wellcome Centre for Anti-Infectives Research, School of Life Sciences, University of Dundee, Dundee DD1 5EH, U.K.

*To whom correspondence should be addressed: Michael A.J. Ferguson, Wellcome Centre for Anti-Infectives Research, School of Life Sciences, University of Dundee, Dundee DD1 5EH, U.K. +44 1382 384219, m.a.j.ferguson@dundee.ac.uk

Running title: *N*-glycosylation in *Trypanosoma brucei*

Key words: Trypanosoma, oligosaccharyltransferase, *N*-linked glycosylation, STT3, glycoproteomics, parasitology, glycoprotein biosynthesis, machine learning

ABSTRACT

Trypanosoma brucei **causes African trypanosomiasis and contains three full-length oligosaccharyltransferase (OST) genes; two of which, *Tb*STT3A and *Tb*STT3B, are expressed in the bloodstream form of the parasite. These OSTs have different peptide acceptor and lipid-linked oligosaccharide donor specificities and trypanosomes do not follow many of the canonical rules developed for other eukaryotic *N*-glycosylation pathways, raising questions as to the basic architecture and detailed function of trypanosome OSTs. Here, we show by blue-native gel electrophoresis and stable isotope labelling in cell culture proteomics that the *Tb*STT3A and *Tb*STT3B proteins associate with each other in large complexes that contain no other detectable protein subunits. We probed the peptide acceptor specificities of the OSTs *in vivo* using a transgenic glycoprotein reporter system and performed glycoproteomics on endogenous parasite glycoproteins using sequential endoglycosidase-H and peptide-*N*-glycosidase-F digestions. This allowed us to assess the relative occupancies of numerous *N*-glycosylation sites by endoglycosidase-H resistant *N*-glycans originating from $Man_5GlcNAc_2$-PP-dolichol transferred by *Tb*STT3A, and endoglycosidase-H sensitive *N*-glycans originating from $Man_9GlcNAc_2$-PP-dolichol transferred by *Tb*STT3B. Using machine learning we assessed the features that best define *Tb*STT3A and *Tb*STT3B substrates *in vivo* and built an algorithm to predict the types of *N*-glycan most likely to predominate at all the putative *N*-glycosylation sites in the parasite proteome. Lastly, molecular modelling was used to suggest why *Tb*STT3A has a distinct preference for sequons containing and/or flanked by acidic amino acid residues. Together, these studies provide insights into how a highly divergent eukaryote has re-wired protein *N*-glycosylation to provide protein sequence-specific *N*-glycan modifications. Data are available via ProteomeXchange with identifiers PXD007236, PXD007267 and PXD007268.**

INTRODUCTION

The tsetse-fly transmitted protozoan parasite *Trypanosoma brucei* and its close relatives are responsible for human and animal African trypanosomiasis. The animal-infecting bloodstream forms of these organisms depend on surface coats made of glycosylphosphatidylinositol (GPI) anchored and *N*-glycosylated variant surface glycoprotein (VSG) to evade the innate host immune system (1) and the acquired immune system through antigenic variation (2). Further, they express many less abundant glycoproteins such as their novel transferrin receptors (3-5), a novel lysosomal/endosomal protein called p67 (6), the so-called invariant surface glycoproteins (ISGs) (7) and invariant endoplasmic reticulum glycoproteins (IGPs) (8), the Golgi/lysosmal glycoprotein tGLP-1 (9), the membrane-bound

This research was originally published in Journal of Biological Chemistry. Jinnelov, A., Liaqat, A., Tinti, M., Ferguson, M. 'Single-subunit oligosaccharyltransferases of Trypanosoma brucei display different and predictable peptide acceptor specificities', *Journal of Biological Chemistry* (2017). © the American Society for Biochemistry and Molecular Biology.

histidine acid phosphatase TbMBAP1 (10), the flagellar adhesion zone glycoproteins Fla1 and Fla2 (11) and others. While some of these are bloodstream form specific glycoproteins (VSGs, ISGs, TbMAP1, transferrin receptors), others are common to the tsetse midgut-dwelling procyclic form of the parasite. Further, procyclic form parasites also express unique glycoproteins, notably the abundant GPI-anchored procyclins some of which are N-glycosylated (12,13) , and the partially characterised high-molecular weight glycoconjugate (14,15). Many of the *N*-glycan structures expressed by *T. brucei* have been solved and these include conventional oligomannose and biantennary complex structures as well as paucimannose and extremely unusual 'giant' poly-*N*-acetyl-lactosamine (poly-LacNAc) containing complex structures in the bloodstream form of the parasite (16-19) . In contrast, only oligomannose *N*-glycans have been structurally described in wild type procyclic trypanosomes(12,20).

The unusual repertoire of *T. brucei* bloodstream form *N*-glycans and the original observation by Bangs and colleagues that EndoH-resistant *N*-glycans appear immediately following protein synthesis (21) and not following transport to the Golgi apparatus, has stimulated our group to study the fundamentals of protein *N*-glycosylation in this divergent eukaryotic pathogen.

Protein *N*-glycosylation is believed to be a ubiquitous post-translational modification among the eukaryotes, with the canonical model based primarily on extensive studies in mammalian cells and the yeast *Saccharomyces cerevisiae*, reviewed in (22,23). In this canonical model, there are a number of tenets that include: (i) The mature lipid-linked oligosaccharide (LLO) donor and preferred substrate for oligosaccharyltransferases (OSTs) is $Glc_3Man_9GlcNA_2$-PP-dolichol. (ii) The OSTs are hetero-oligomers of 8 or 9 distinct subunits. (iii) The OSTs may fall into two classes (A and B) according to their subunit composition with different peptide acceptor specificities. (iv) The ER enzyme UDP-glucose: glycoprotein glucosyltransferase (UGGT) operates on tri-antennary $Man_{9-7}GlcNA_2$ structures with a complete a-branch, but not on bi-antennary structures. (v) The action of Golgi mannosidase II is a pre-requisite for the conversion of oligomannose to complex *N*-glycans. (vi) The enzymes GnTI and GnTII have strict acceptor substrate specificities, operating on Man_5GlcNA_2 and $GlcNAcMan_3GlcNA_2$, respectively. However, none of these tenets apply to *N*-glycosylation in the protozoan parasite *Trypanosoma brucei* where: (i) The largest LLO is Man_9GlcNA_2 (24,25) and where different OSTs preferentially transfer either this structure or bi-antennary Man_5GlcNA_2 (25-28) (ii) There is no evidence for OST subunits other than the catalytic SST3 subunits in *T. brucei* or the related parasites *Trypanosoma cruzi* and the Leishmania (26,28-32). (iii) OST sub-classes and their peptide acceptor specificities (which are more disparate in *T. brucei* than for other eukaryotes) are defined only by their STT3 components (28). (iv) The parasite UGGT works efficiently on all structures (bi- and tri-antennary) with an intact A-branch (33). (v) Golgi-mannosidase II is absent, preventing the conversion form the oligomannose series to the complex series of *N*-glycans (25). (vi) The parasite GnTI and GnTII βGlcNAc-transferases have different specificities to canonical GnTIs and GnTIIs and belong to a different GT family (34,35).

Here, we: (i) Directly address the oligomeric states of *T. brucei* STT3 subunits. (ii) Look for evidence for any non-canonical OST subunits in addition to the *Tb*STT3s. (iii) Probe the peptide acceptor specificities of *Tb*STT3A (Tb927.5.890) and *Tb*STT3B (Tb927.5.900) using a reporter glycoprotein expression system and by glycoproteomics. (iv) Use machine learning to predict which putative *N*-glycosylation sites in bloodstream form *T. brucei* will be modified by *Tb*STT3A or *Tb*STT3A.

RESULTS

Blue native gel electrophoresis of *in situ* tagged *Tb*STT3A suggests it is present in high molecular weight complexes. To enable immunoprecipitation of *Tb*STT3A, the 3'-end of the endogenous gene in a heterozygote cell line (*Tb*STT3A/B/C$^{+/-}$) was *in situ* tagged with a sequence encoding a C-terminal HA_3 epitope. Transfected cells were cloned and analysed by Southern blotting to confirm correct insertion of the tag (Figure S1A). To check that the tag did not impair the function of *Tb*STT3A, the glycosylation of VSG221 was analysed in these cells. VSG221 receives different types of glycan at the two *N*-glycosylation sequons in the protein; EndoH-resistant $Man_5GlcNAc_2$ at N263 and EndoH-sensitive $Man_9GlcNAc_2$ at N428 (17,28). Following PNGaseF and EndoH treatment the typical digestion pattern for wild-type VSG221 was seen also for the transgenic cells (Figure S1B), showing that C-terminally tagged *Tb*STT3A-HA$_3$ is functional. Subsequently, cells were lysed under mild conditions (0.5% digitonin on ice for 30 min) and the clarified cell lysates were incubated with anti-HA mouse antibody, followed by magnetic beads coupled to protein G. The pull-out eluates were analysed by SDS-PAGE and Western blotting using an anti-HA antibody. As expected, epitope tagged *Tb*STT3A-HA$_3$ was detected running just below the position of the 75 kDa molecular weight marker in the *Tb*STT3A-HA$_3$ pull-out, whereas no band was seen in the wild-type cell pull-out (Figure 1A). However, when same eluates were analysed by blue native gel electrophoresis and anti-HA Western blotting, a smear (specific for the *Tb*STT3A-HA$_3$ cell line) was detected between 700 and 1200 kDa, suggesting that *Tb*STT3A is present in large complexes (Figure 1B).

SILAC proteomics shows *Tb*STT3A and *Tb*STT3B form hetero-oligomeric complexes without other subunits. Since the results from the blue native gel electrophoresis suggested *Tb*STT3A is present in high molecular weight complexes, we carried out pull-out experiments using stable isotope labelling in cell culture (SILAC). For this experiment, wild-type and transgenic parasites (expressing *Tb*STT3A-HA$_3$) were grown under identical conditions for eight cell divisions, except that the transgenic *Tb*STT3A-HA$_3$ cell line was grown in "heavy medium" containing stable isotope-labelled Lys and Arg (R_6K_4), whereas the wild-type cells were grown in "light medium" containing unlabelled Lys and Arg (R_0K_0). The transgenic *Tb*STT3A-HA$_3$ and the wild-type cells were harvested, washed, counted, mixed together in a 1:1 ratio and lysed in 0.5 % digitonin buffer. Anti-HA antibodies and protein G magnetic beads were used to pull-out the *Tb*STT3A-HA$_3$ tagged protein, and any binding partners, and the bead eluate was processed to tryptic peptides for LC-MS/MS analysis. In this kind of SILAC experiment, *Tb*STT3A-HA$_3$ and any proteins specifically associated with it can be distinguished from nonspecific contaminant proteins by the isotope ratios of their tryptic peptides. Thus, *Tb*STT3A-HA$_3$ and true associated protein peptides will have high heavy/light isotope ratios, whereas contaminant proteins will have approximately equal heavy/light isotope ratios (Figure 2A). The data set from the experiment was used to search a *T. brucei* predicted protein database using MaxQuant software. Each protein was displayed on a plot of the Log$_{10}$ value of the intensities of the unique peptides of that protein (y axis) and the Log$_2$ value of the heavy to light isotope ratios of the same peptides (x axis) (Figure 2B). The *Tb*STT3A-HA$_3$ (bait) protein had the highest heavy/light ratio (14 : 1), closely followed by *Tb*STT3B (10 : 1). Only three other proteins were significantly enriched (orange crosses, Figure 2B). However, these were only marginally (1.5-fold) enriched hits that are not known to localise to the ER.

The *Tb*STT3A-HA$_3$ cell line was further modified by the *in-situ* tagging of the remaining *Tb*STT3B allele, to yield a cell line expressing C-terminally MYC$_3$-tagged *Tb*STT3B-MYC$_3$. A complementary SILAC experiment, using *in situ* TbSTT3B-MYC$_3$ tagged bait and an anti-MYC pull-out, produced similar result to the *Tb*STT3A-HA$_3$ pull-out, with *Tb*STT3A being the only obvious binding partner for *Tb*STT3B-MYC$_3$ (Figure 2C). In this case, there was one other significant protein hit (orange cross, Figure 2C), corresponding to a glucose transporter, but this was different from those seen in (Figure 2B) and also unlikely to be an ER component.

Taken together, these data suggest that *Tb*STT3A and *Tb*STT3A form hetero-oligomeric complexes, with no other candidate subunits, although we cannot rule out the presence of low-affinity subunits that might be lost during immunoprecipitation.

The data for the SILAC proteomics experiments can be found at ProteomeXchange under entry PXD007236.

Co-immunoprecipitation of *Tb*STT3A-HA$_3$ and *Tb*STT3B-MYC$_3$. The result from the SILAC pull-out experiments suggested that *Tb*STT3A is in a complex with *Tb*STT3B and to further test this hypothesis immunoprecipitation (IP) experiments were performed. Cells from the double-tagged cell line were harvested, washed and lysed in 0.5 % digitonin and *Tb*STT3A-HA$_3$ or *Tb*STT3B-MYC$_3$ was captured from the lysate using anti-HA or anti-MYC magnetic beads. Subsequently, the tagged proteins were detected by Western blotting using anti-HA and anti-MYC antibodies. Wild type cell lysates (containing no HA or MYC tagged genes) were used as a control. The results from the co-IP experiments are shown in (Figure

3). As expected, no bands in the region of *Tb*STT3A-HA$_3$ or *Tb*STT3B-MYC$_3$ were seen in the IPs from the control wild-type lysates (Figure 3, lanes 1, 3, 5 and 7). Also, as expected, *Tb*STT3A-HA$_3$ and *Tb*STT3B-MYC$_3$ were detected in the homologous anti-HA IP / anti-HA blot (Figure 3, lane 2) and anti-MYC IP / anti-MYC blot (Figure 3, lane 8). Significantly, *Tb*STT3B-MYC$_3$ can be seen to co-IP with *Tb*STT3A-HA$_3$ (Figure 3, lane 6) and *Tb*STT3A-HA$_3$ can be seen to co-IP with *Tb*STT3B-MYC$_3$ (Figure 3, lane 4), confirming their physical association predicted by the SILAC experiment.

In these experiments, the anti-HA IP / anti-HA Western blot signal for *Tb*STT3A-HA$_3$ is much stronger than the anti-MYC IP / anti-MYC Western blot signal for *Tb*STT3B-MYC$_3$. While some of this difference may be due to relative antibody affinities, it is also consistent with the higher expression of *Tb*STT3A in wild-type bloodstream form trypanosomes at both the mRNA and protein levels (28,36). The co-IP data suggest that a significant proportion of the total *Tb*STT3B-MYC$_3$ appears in the *Tb*STT3A-HA$_3$ IP (Figure 3, compare lanes 6 and 8), whereas only a minority of *Tb*STT3A-HA$_3$ appears in the *Tb*STT3B-MYC$_3$ IP (Figure 3, compare lanes 2 and 4). These data suggest that there may be some high molecular weight complexes made exclusively, or almost exclusively, of *Tb*STT3A whereas all, or most, of the *Tb*STT3B is present in complexes containing *Tb*STT3A.

Probing peptide acceptor substrate specificities of *Tb*STT3A and *Tb*STT3A using a reporter glycoprotein expression system. *Tb*STT3A is responsible for co-translational transfer of biantennary Man$_5$GlcNAc$_2$ predominantly to *N*-glycosylation sequons containing and/or flanked by acidic amino acids, whereas *Tb*STT3B catalyses post translational transfer of triantennary Man$_9$GlcNAc$_2$ to the remaining sterically accessible sequons (28). To improve our understanding of the acceptor peptide specificity in *T. brucei*, an *in vivo* assay was established using an artificial reporter glycoprotein, based on the *Tb*BiPN system described in (37). *Tb*BiPN is a non-glycosylated truncated version of *Tb*BiP, retaining its *N*-terminal signal peptide but lacking its C-terminal ER retention peptide, which enters the ER and is eventually secreted out of the cell *via* the Golgi apparatus. A pLEW82 expression plasmid (38) was modified to contain the *Tb*BiPN open reading frame fused to a 3'-sequence into which we could insert additional sequences, via AvrII and MfeI restriction sites, immediately upstream of a C-terminal HA$_3$ epitope tag. This construct was used to introduce sequences encoding a single reporter *N*-glycosylation sequon, flanked by five amino acid residues on each side. These TbBiP*N*-[XXXXXNXTXXXXX]-HA$_3$ constructs were transformed into bloodstream form trypanosomes to express the reporter glycoprotein.

We first validated the *in vivo* reporter assay by introducing TEGLLNATDEIAL and TILKSNYTAEPVR into the *Tb*BiPN construct and expressing them in *T. brucei*. The former sequence (with a pI of 3.42) is found in VSG MITat1.8 and is known, in that context, to receive exclusively biantennary Man$_5$GlcNAc$_2$ from *Tb*STT3A (39), whereas the latter (with a pI of 8.3) is found in the ESAG6 subunit of the transferrin receptor and is known not to be recognised and modified by *Tb*STT3A and, therefore, receives exclusively triantennary Man$_9$GlcNAc$_2$ from *Tb*STT3B (5). Aliquots of trypanosome lysates expressing these constructs were treated with and without EndoH or PNGaseF, followed by SDS-PAGE gel and Western blotting with anti-HA antibodies. The endoglycosidase EndoH can only digest triantannary Man$_9$GlcNAc$_2$ glycans transferred by *Tb*STT3B, whereas PNGaseF can digest both Man$_9$GlcNAc$_2$ and biantennary Man$_5$GlcNAc$_2$ transferred by *Tb*STT3A. Thus, distinct digestion patterns, depending on what type(s) of glycan(s) are bound to the sequon asparagine, can be visualised by Western blotting. The *in vivo* reporter assay faithfully recapitulated the experimental data for the VSG MITat1.8 and ESAG6 glycosylation sites, with TEGLLNATDEIAL- and TILKSNYTAEPVR-containing *Tb*BipN glycoproteins occupied predominantly by EndoH-resistant and EndoH-sensitive glycans, respectively (Figure S2).

Next, we investigated how each position flanking and within the sequon affects *Tb*STT3A recognition and transfer. First, the neutral sequence AAAAANATAAAAA (pI 6.01) was introduced into the reporter glycoprotein. For this construct, the majority (about 93% as measured by quantitative Licor imaging of the upper and lower bands of the EndoH-digests) of the anti-HA binding signal was sensitive to EndoH (Figure 4A, lane 1; Table 1). One aspartic acid was then introduced in all eleven possible positions, yielding peptides with the same pI value (3.10), and the proportion of the reporter glycoprotein processed by *Tb*STT3A (and therefore resistant to EndoH) was measured (Figure 4A, lanes 2 and 3; Table 1). The quantitative data, derived from two technical replicates of three biological replicates (Table 1), are summarised in (Figure 4B). From these data, it can be seen that the aspartic acid scan across the different positions leads to variation in recognition and glycan transfer by *TbSTT3A*; with the two positions immediately flanking the Asn residue apparently having the greatest influence and with residues *N*-terminal to the glycosylation sequon having greater influence to those C-terminal to the sequon.

Probing endogenous peptide acceptor substrate specificities of *Tb*STT3A and TbSTT3B by glycoproteomics. The glycoproteomics data from (28) were reprocessed and significantly augmented by combing them with data derived from the experiments outlined in Experimental Procedures. These experiments assess whether the endogenous trypanosome *N*-glycosylation sites are occupied by EndoH-sensitive oligomannose glycans (originating from the action of *Tb*STT3B) or by EndoH-resistant paucimannose and/or complex glycans (originating from the action of *Tb*STT3A).

Parasites were osmotically lysed to release >90% of the VSG coat as soluble form VSG (sVSG), through the action of the endogenous GPI-specific phospholipase C (40,41). The recovered cell ghosts, containing majority of the non-VSG cellular glycoproteins, were solubilized, denatured and S-alkylated. This preparation was then processed in two ways. In one approach the intact glycoproteins were first affinity purified using immobilized ricin (RCA_{120}) and Concanavalin-A (ConA) lectins. The enriched glycoproteins were then sequentially digested with EndoH and PNGaseF (the latter in the presence of $H_2[^{18}O]$), and digested with Lys-C and trypsin. In the other approach the denatured and S-alkylated proteins were first digested with Lys-C and trypsin and the glycopeptides were trapped with ricin (RCA_{120}) and ConA and subsequently digested with EndoH and PNGaseF (the latter in the presence of $H_2[^{18}O]$). In both cases, the resulting peptides were analysed by LC-MS/MS and the data used to search the *T. brucei* predicted protein database allowing for the possible presence of Asn-*N*-GlcNAc residues, the product of EndoH cleavage, and/or for the conversion of Asn residues into $[^{18}O]$Asp residues, the product of PNGaseF cleavage. These data, and reprocessed data from Izquierdo et al. (28), are shown in (Supplementary Table S1). Peptides containing Asn-*N*-GlcNAc and/or $[^{18}O]$Asp within an Asn-X-Ser/Thr sequon that were detected >=3 times were assigned as being predominantly *Tb*STT3A or *Tb*STT3B substrates when the proportion of the $[^{18}O]$Asp feature was >= 0.8 and <=0.4, respectively (Supplementary Table S2). We then analysed the amino acid frequencies immediately adjacent to *N*-glycosylation sequons of the assigned *Tb*TT3A and *Tb*TT3B substrates using WebLogo (42). The enrichment of negatively charged residues is the most striking feature of the *Tb*STT3A substrates (Figure 5A) along with a bias towards Thr over Ser in the sequon +2 position. Conversely, the *Tb*STT3B substrates are relatively enriched for positively charged residues upstream of the sequon but show no preference for Thr over Ser in the +2 position (Figure 5B). The data were also analysed by the two-sample logo visualization method (43), which compares two input peptide sequence lists against each other, highlighting features that predominate in each. This suggests enrichment for negatively charged amino acids, especially at positions -6, - 3, +6 and +7 for the *Tb*STT3A substrates and a preference for hydrophobic and positively charge amino acids in the -1 to -5 positions of the TbSTT3B substrates (Figure 5C).

The reprocessed data of (19) can be found at ProteomeXchange under entry PXD007237 and the new glycoproteomics data under PXD007238.

Building a glycosylation site predictor for *T. brucei* using machine learning. Machine learning has been successfully used in biological research to infer the peptide recognition specificities of, for example, protein kinases, phosphatases and SH2 domains (44). The first step of machine learning consists in transforming peptide sequences into biochemical features such as charge, hydrophobicity and relative positions. These features, organized in a machine-readable template, are than evaluated by artificial intelligence algorithms to highlight which amino acid properties of a peptide sequences are the most important in determining substrate recognition. We therefore decided to apply a machine learning approach to further leverage our glycoproteomics and glycoprotein reporter data and build an ensemble of prediction algorithms to assign putative *N*-glycosylation sites in the predicted *T. brucei* proteome. This ensemble algorithm (Voting Classifier) averages the outputs of a Random Forest classifier (RF), an Extra Tree Classifier (ETC) and a Support Vector Machine (SVM) classifier to predict which putative *N*-glycosylation sites will more likely be modified by *Tb*STT3A or *Tb*STT3B. The RF, ETC and SVM classifiers all weigh the features derived from the upstream (amino-terminal) side of the glycosylation sequon more than the features extracted from the downstream (carboxy-terminal) side (Supplementary Figure S3A-C). Moreover, the RF, ETC and SVM classifiers all preferentially use the cumulative negative charge upstream of the glycosylated asparagine, and the hydrophobicity of the peptide upstream and downstream, to discriminate between *Tb*STT3A and *Tb*STT3B substrates (Figure S4). We could detect a core of 5 important features shared by the three classifiers, namely the charge at pH 7.3 and the isoelectric point for the sequon ± 10 amino acid residues, the cumulative charge upstream the modified asparagine with a window of 13 and 16 amino acids and the bonus score derived from the aspartic acid scanning experiment described in (Figure 4), see Experimental Procedures. A list of putative *N*-glycosylation sites and the Voting Classifier predictions are shown in (Supplementary Table S3). We performed a two-sample logo visualization on the output (Figure 5D). As expected, the trends in plot are similar to those generated from the glycoproteomics data alone (Figure 5C).

Molecular modelling. Although *Tb*STT3C (Tb927.5.910) has not been studied in this paper because it is not expressed at detectable levels in bloodstream form *T. brucei*, data from its heterologous expression in yeast suggests that its peptide acceptor specificity is much more similar to *Tb*STT3A than *Tb*STT3B (28,45). To try to rationalise the preferences of *Tb*STT3A and *Tb*STT3C for acceptor sequons containing and/or flanked by negatively charged amino acid residues, we built molecular models of *Tb*STT3A, *Tb*STT3B and *Tb*STT3C using Phyre2 (46) based on the *Campylobacter lari* PglB structure (47). The predicted models were aligned with the PglB structure using PDBeFold (48) and the binding pockets of *Tb*STT3A and *Tb*STT3B were visualized with Chimera (49). Next, we looked for basic amino acid residues that were conserved in *Tb*STT3A and *Tb*STT3C that were different in *Tb*STT3B (Figure S6). Of these, the active site proximal residue 397 is particularly interesting as it contains a His residue in *Tb*STT3B but an Arg residue in *Tb*STT3A and *Tb*STT3C. Arginine has a flexible and strongly basic guanidinium cation side chain that could conceivably interact with acidic amino acid residues at or close to the NXS/T glycosylation sequon in the active site. Position 406 contains an Arg residue in *Tb*STT3A and *Tb*STT3C (in place of a neutral Gly residues in *Tb*STT3B) that could also conceivably interact with sequon-adjacent anionic residues in the acceptor peptide. Such ionic interactions between the acceptor peptide and the enzyme surface might increase the efficiency of substrate recognition and glycosylation of sequons containing and/or flanked by acidic amino acids.

DISCUSSION

Although the *T. brucei* genome does not encode for any indentifiable OST subunits, other than three intact and one truncated *Tb*STT3, we decided to investigate whether there might be novel non-canonical *T. brucei* OST subunits. Precedents for kinetoplastid-specific subunits of otherwise conserved cellular machineries include clathrin-associating proteins and endocytic components (50,51), exocyst components (52), nuclear pore complex and nuclear lamina components (53,54) and subunits of the GPI transamidase (55). Blue-native gel electrophoresis of gently solubilised epitope-tagged endogenous *Tb*STT3A showed that it is present in high-molecular weight complexes, but quantitative SILAC proteomics of tagged *Tb*STT3A and *Tb*STT3B pull-outs showed that while these are mutual binding partners (confirmed by co-immunoprecipitation), no other subunits could be found by these methods. The lack of non-canonical or canonical *T. brucei* OST subunits (other than the STT3 catalytic subunits) is consistent with the ease with which *T. brucei* and other kinetoplastid STT3s can be functionally expressed in other eukaryotes, like *S. cerevisiae* and *Pichia pastoris* (29,31,32,56). Nevertheless, the blue-native gel and co-immunoprecipitation experiments show that, at least in the native environment of a bloodstream form trypanosome, *Tb*STT3A associates with itself and with the less-abundant *Tb*STT3B to form complexes with apparent molecular weights of between 600 kDa and 1.2 MDa. The nature of these complexes remains to be determined but has implications for how and whether *Tb*STT3s associate with the parasite translocon complex and how they access nascent glycoprotein sequons during and/or following protein translocation. The dimer/oligomer nature of yeast OST subunits, including Stt3, has been previously described (57).

To probe *Tb*STT3 peptide acceptor substrate specificities, we developed an artificial glycoprotein reporter system, based on a truncated version of *Tb*BiP (37) fused to a single glycosylation sequon flanked by five variable residues on either side. Constructs were expressed in bloodstream form *T. brucei* and their products assayed for the relative proportions of *N*-glycosylation by *Tb*STT3A and *Tb*STT3B. With this we able to recapitulate the preferential *N*-glycosylations of native peptide acceptor sequences. We then applied the system to analyse the *N*-glycosylation of an artificial 13-mer sequence (AAAAANATAAAAA) into which we sequentially introduced a single D residue in all 11 possible A sites. The data clearly confirmed that the presence of an acidic amino acid proximal to the sequon significantly increased its *N*-glycosylation by *Tb*STT3A, with the -1 and +1 positions relative to the *N*-glycosylated Asn residue having the greatest effect and the positions N-terminal to the sequon having a greater effect than those C-terminal to the sequon.

We then created a richer glycoproteomics dataset than we previously reported (28) by combining two alternative approaches: (i) Glycoprotein enrichment by lectin affinity chromatography, followed by trypsin digestion and sequential EndoH and PNGaseF digestion and (ii) Tryptic glycopeptide enrichment by lectin affinity chromatography, followed by trypsin digestion and sequential EndoH and PNGaseF digestion. In both cases, the PNGaseF digestion step was performed in $H_2[^{18}O]$ to distinguish between PNGaseF-mediated Asn deamidation and non-enzymatic deamidation during sample preparation and handling. These data were combined with reprocessed raw data from (28) to provide quantitative data on *Tb*STT3A and *Tb*STT3B *N*-glycosylation of 141 unique *N*-glycosylation sites. Logo plots confirmed the enrichment of acidic amino acid residues (Asp and Glu) surrounding *Tb*STT3A *N*-glycosylated sequons, the general depletion of hydrophobic residues and the selective depletion of basic residues (Arg and Lys) N-terminal to the sequon.

We also used hypothesis-free machine learning techniques to identify features that predispose sequons to be preferentially modified by *Tb*STT3A or *Tb*STT3B and, finally, combined these features and parameters derived from the experimental reporter glycoprotein data to develop a Voting Classifier prediction algorithm. This predictor was then applied to all the putative *N*-glycosylation sequons in the *T. brucei* proteome to predict those sites preferentially modified by *Tb*STT3A (leading to paucimannose and/or complex *N*-glycan occupancy) or *Tb*STT3B (leading to oligomannose *N*-glycan occupancy) in bloodstream form trypanosomes. Two-sample logo plot analysis of the output (a total of 1,291 predicted occupied *N*-glycosylation sites) largely echoes the experimental glycoproteomic and reporter glycoprotein data and implies that the *Tb*STT3A *N*-glycosylation sites are enriched for acidic residues and depleted of basic and hydrophobic residues, with the effects of these features more profound to the N-terminal side and within the sequon than to the C-terminal side of the sequon.

It is important to note that available pulse-chase data (21,58) suggest that *Tb*STT3A modifies VSG glycoproteins co-translationally, whereas *Tb*STT3B can act post-translationally (25) and that *Tb*STT3A and *Tb*STT3B knockdown data suggest that *Tb*STT3B is able to modify *Tb*STT3A sites, but not *vice versa* (28). Thus, all or most the aforementioned *Tb*STT3A versus *Tb*STT3B sequon selectivity features are dictated by the peptide/sequon acceptor specificity of *Tb*STT3A and not *Tb*STT3B, which appears to be able to utilise sequons in almost any amino acid sequence context. This property of *Tb*STT3B could be particularly useful from a biotechnological point of view to boost the efficient *N*-glycosylation of recombinant glycoproteins in eukaryotic expression systems. It also nicely explains why trypanosomes can transition from expressing a rich mixture of oligomannose *and* paucimannose/complex *N*-glycans in the bloodstream form parasite (16,18,19) to predominantly oligomannose *N*-glycans in the procyclic form of the parasite (12,20) by simply down-regulating the expression of *Tb*STT3A, as observed at both the mRNA (28,59) and protein-levels (36,60,61).

The results reported here and in (45) are consistent with the mechanisms of resistance in *T. brucei* to certain toxic lectins and carbohydrate-binding small molecules reported in an interesting series of studies (62-64). These workers demonstrated that the parasites could escape the effects of these trypanocidal agents, all of which bind principally to oligomannose *N*-glycans, by either switching to the expression of a VSG type that naturally does not carry oligomannose *N*-glycans or by supressing the expression of *Tb*STT3B. In this way, the parasites effectively exchange oligomannose for puacimannose and complex *N*-glycans that are poor ligands for the trypanocides.

Interestingly, *Tb*STT3A and *Tb*STT3B are found in tandem array in the trypanosome genome, together with a preceding truncated *Tb*STT3 pseudogene and followed by a full-length *Tb*STT3C gene that is more similar to *Tb*STT3B than to *Tb*STT3A. However, *Tb*STT3C is not significantly expressed in bloodstream form or procyclic form of the parasite (28,36,59-61). Nevertheless, transgenic expression of *Tb*STT3C in *S. cerevisiae* clearly shows that it is a functional OST with a similar preference for sequons flanked by acidic amino acids to *Tb*STT3A but an LLO donor specificity like *Tb*STT3B (28) and (45).Amino acid sequence alignment of *Tb*STT3A, B and C and molecular modelling building, based on the *Campylobacter lari* PglB structure (47) was performed. The models suggest that the presence of a large, flexible and highly positively charged Arg residue sidechain (R397) very close to the active site of the enzyme in *Tb*STT3A and *Tb*STT3C, compared to a His residue in *Tb*STT3B, may play a role in the selectivity of *Tb*STT3A and *Tb*STT3C for sequons containing and flanked by acidic Asp and Glu residues. *Tb*STT3A and *Tb*STT3C also contain Arg residues in place of neutral Q567 and G406 residues in *Tb*STT3B, locations close enough to the active site to potentially interact with sequon-flanking anionic residues. The accompanying paper (45) elegantly addresses the issues of peptide acceptor and LLO donor specificities of all three *Tb*STT3s by heterologous expression of each and chimeras thereof in various yeast mutants. That paper concludes that the region containing R397 and R406 in *Tb*STT3A and *Tb*STT3C controls peptide acceptor specificity, and this is consistent with our suggestions from molecular modelling.

There are similarities and differences between the multi-subunit mammalian STT3A- and STT3B-based OSTs and the single subunit *Tb*STT3A and *Tb*STT3B OSTs of *T. brucei*: (i) In both, the STT3A OSTs operate co-translationally and get the first option to glycosylate a given sequon, whereas the STT3B OSTs can operate post-translationally on what is left (25,65). (ii) In both, the OSTs show differences in peptide acceptor substrate specificity. However, in *T. brucei* this is controlled by the physicochemical properties of the amino acids surrounding the acceptor sequon, whereas in mammalian cells this is controlled by the position of the sequon relative to the C-terminus of the protein (66) or proximity to the signal-peptide cleaved N-terminus of the protein and to cysteine residues (67-69). The latter appears to relate to the presence of the mutually redundant MagT1 or TUSC3 thioredoxin-like oxidoreductase subunits (equivalent to the yeast Ost3 and Ost6 subunits) in the STT3B OST that may form mixed disulphides with the sequon proximal cysteine residues and thus increase residence time with the STT3B OST (69,70). A role for oxidoreductase activities of Ost3 and Ost6 in yeast OST acceptor site specificity was also previously (71,72). In this regard, it is worth noting that *Tb*STT3B and *Tb*STT3C contain a CXC motif (absent in *Tb*STT3A) that are predicted from the *C. lari* OST structure to be proximal to the acceptor peptide (47). Such CXC sequences can have a disulphide isomerase activity (73) that might conceivably increase the acceptor substrate range of *Tb*STT3B and *Tb*STT3C. (iii) Whereas both STT3A and STT3B OSTs prefer the mature $Glc_3Man_9GlcNAc_2$-PP-dolichol LLO donor, the *T. brucei* OSTs have distinct LLO donor specificities such that the presence of the ALG12-dependent c-branch of the conventional (but glucose-free) triantennary $Man_9GlcNAc_2$-PP-dolichol LLO is required by *Tb*STT3B but not tolerated by *Tb*STT3A (27) and (45). An important consequence of this differential LLO specificity is that, because Golgi mannosidase II activity is also absent in *T. brucei*, *N*-glycans derived from *Tb*STT3B glycosylation cannot be processed to paucimannose or complex structures, which must instead be derived exclusively from *Tb*STT3A glycosylation.

In summary, the two simultaneously operating acceptor substrate- and donor substrate-specific *N*-glycosylation systems of bloodstream form *T. brucei* have been further characterized in this paper. While no canonical OST subunits, other than catalytic STT3 subunits, could be found the parasite genome, we can now confirm that there are no non-canonical subunits either. Instead, *Tb*STT3A and *Tb*STT3B appear form multimeric high-molecular weight complexes containing either *Tb*STT3A alone or *Tb*STT3A and

*Tb*STT3B. Further insights into the peptide acceptor specificity of *Tb*STT3A have been provided and an algorithm has been generated to predict, proteome-wide, which OST will likely operate on which putative *N*-glycosylation site. Taken together with the unusual specificities of *T. brucei* UGGT, GnTI and GnTII enzymes described in the introduction (19,34,35), and the apparent absence of a regulated ER unfolded protein response (19,74), we may conclude that protein *N*-glycosylation and downstream processing in this divergent eukaryote is worthy of note, and that its unusual features may provide therapeutic possibilities.

EXPERIMENTAL PROCEDURES

Cultivation of *Trypanosomes*. Bloodstream form *Trypanosoma brucei*, genetically modified to express T7 polymerase and the tetracycline repressor protein (38), were cultured in HMI-9T medium (75) supplemented with 10 % fetal calf serum, 2 mM Glutamax^{TM} (Invitrogen) and 56 µM 1-thioglycerol (in place of 2-mercaptoethanol) and 2.5 µg/ml G418 antibiotic at 37 ºC in a 5% CO_2 incubator. Other antibiotics used, as appropriate, were hygromycin (4 µg/ml), puromycin (2.5 µg/ml), phleomycin (0.1 µg/ml) and tetracycline (0.5 µg/ml). SILAC labelling, using dialysed fetal calf serum, was performed in HMI11-SILAC media, as described in (36). L-Arginine U–$^{13}C_6$ and L-Lysine 4,4,5,5-$^2H_4$ (R_6K_4) were purchased from Cambridge Isotope Labs.

Generation of genetically modified trypanosomes with *in situ* tagged TbSTT3A-HA_3 and TbSTT3B-MYC_3. The *Tb*STT3A,B,C$^{-/+}$ heterozygote described in (28) was used for C-terminal HA_3 *in-situ* tagging of the remaining *Tb*STT3A allele using a pMOTagH4 plasmid and C-terminal MYC_3 *in-situ* tagging of the remaining *Tb*STT3B allele using a pMOTag4M4 plasmid (76). For the pMOTagH4 plasmid, the 1328-bp from the C-terminus of *Tb*STT3A and the 1057-bp 3' UTR downstream of the gene orf were PCR-amplified from genomic DNA using Kod Hot Start polymerase with primers 5'-ataagtatctcgagcaagtttgcttgccccgttcg-3' and 5'-ataagtaactcgagctc*gctctgaaaatacaggttttc*gacttcgtaa tggaaccgcttcgct-3' and 5'-ataagtatggatccccacatcgtttcaatcgccgc-3' and 5'-ataagtaaggatccactcacaatcgtgcttacagcc-3' as forward and reverse primers, respectively. The PCR products were cloned into the plasmid using the XhoI and BamHI (underlined). A TEV restriction site was included downstream of the orf of *Tb*STT3A (italics). The construct was linearized before being transfected into the *Tb*STT3A,B,C$^{-/+}$ heterozygote cell line and transfected cells were selected by addition of hygromycin. The pMOTag4M4 plasmid was ordered from Genescript. It included 1032-bp from the C-terminus of *Tb*STT3B, located upstream of the MYC_3 epitope in the plasmid. The plasmid also included 835-bp of the 3' UTR of *Tb*STT3B which were located downstream of the blasticidin resistance gene in the plasmid. The construct was linearized before being transfected into the *Tb*STT3A,B,C$^{-/+}$; *Tb*STT3A-HA_3 heterozygote cell line and transfected cells were selected by addition of blasticidin.

SDS-PAGE and Western blotting. Reducing SDS-PAGE was run using pre-cast Novex Bis-Tris gels with MOPS running buffer (Invitrogen). Proteins were transferred to nitrocellulose using an iBlot system (Invitrogen) and stained with Ponceau S (Sigma) before being blocked in 50 mM TrisHCl, 0.15 M NaCl, 0.05 % Tween 20, 0.25 % BSA, 0.05 % Na and 2 % fish skin gelatin pH 7.4 for 20 min. The membrane was then incubated for 30 min with primary antibody in a 50 ml Falcon tube followed by washing using a SnapID system (Millipore). Subsequently, the labelled secondary antibody was incubated for 30 min followed by a washing step. The blots were imaged using an ODYSSEY® SA near infrared imager (LI-COR Biosciences). Secondary LI-COR antibodies (IRDye-800CW goat anti-mouse 1:15000 or IRDye-680RD donkey anti mouse 1:20000) were used to bind the primary mouse anti-HA and anti-MYC antibodies.

Blue native gels and Western blotting. Blue Native gel electrophoresis was run using components from the Native Page kit (Invitrogen). The protocol was followed to the manufacturer's instructions except that no G-250 was added to the sample buffer and the 1x NativePage Light Cathode buffer was diluted 1:4 in 1x running buffer to reduce Coomassie interference of the post-blotting LiCor imaging.

Immunoprecipitation. Cells cultures (100 ml) were grown to log phase (approximately 2.5x10^6 cells/ml) and lysed for 30 min on ice in 0.5%

digitonin, 50 mM Tris-HCl pH6.8, 20 mM EDTA plus the protease inhibitors 0.8 mM PMSF, 0.1 mM TLCK, and 1 x EDTA-free protease inhibitor cocktail (Roche). Subsequently, the lysate was centrifuged (4°C, 11000g, 15 min) and the supernatant was moved to a new tube. Protein G magnetic beads, pre-washed in lysis buffer, were added to the lysate (30 min, 4°C) and captured to absorb nonspecific binding components. The lysate was moved to a new tube followed by incubation for 1 h with anti-HA or anti-MYC antibody (1 µg/ml) at 4°C followed by fresh pre-washed magnetic beads. The beads were captured and washed twice with 1 ml of lysis buffer and once with 10 mM Tris-HCl pH6.8, 4 mM EDTA, 0.1 % digitonin containing the same protease inhibitors. Proteins were eluted from the beads in 30 µl reducing SDS-sample buffer with heating (100°C for 10 min). The eluted proteins where subsequently separated by SDS-PAGE.

SILAC proteomics. Heavy and light labelled cells were harvested separately (15 min, 800 g, 4 °C) and washed and resuspended in trypanosome dilution buffer (20 mM Na_2HPO_4, 2 mM NaH_2PO_4, 80 mM NaCl, 5 mM KCl, 20 mM glucose) for cell counting. The cells were mixed 1 : 1 before undergoing immunoprecipitation, as described above. The eluted proteins in 25 µl reducing SDS sample buffer were S-alkylated with 5 µl 300 mM iodoacetamide (30 min, dark) and loaded on Novex NUPAGE 4-12% Bis Tris gel and run at 200 V using MOPS buffer until the proteins had migrated about 2 cm into the gel (visualised by Simply Blue Safe Stain, Thermo Fisher). The protein-containing region of the gel was excised and subjected to in-gel trypsin digestion and aliquots of the extracted peptides were analysed on an LTQ-Orbitrap Velos Pro mass spectrometer coupled with a Dionex Ultimate 3000 RS HPLC system (Thermo Fisher). The sample peptides were loaded at 5 µL/min onto a trap column (100 µm × 2 cm, PepMap nanoViper C18 column, 5 µm, 100 Å, Thermo Scientific) equilibrated in 98% buffer A (2% acetonitrile and 0.1% formic acid (v/v)) and 2% buffer B (80% acetonitrile and 0.08% formic acid (v/v)). The trap column was washed for 3 min at the same flow rate and then switched in-line with a Thermo Scientific resolving C18 column (75 µm × 50 cm, PepMap RSLC C18 column, 2 µm, 100 Å). The peptides were eluted from the column at a constant flow rate of 300 nl/min with a linear gradient from 98% buffer A to 40% buffer B in 128 min, and then to 98% buffer B by 130 min. LTQ-Orbitap Velos Pro was used in data dependent mode. A scan cycle comprised an MS1 scan (*m/z* range from 335-1800) in the Orbitrap (resolution 60,000) followed by 15 sequential data-dependant collision induced dissociation MS2 scans (the threshold value was set at 5000 and the minimum injection time was set at 200 ms).

Glycoproteomics. The protein-based approach was based on the methodiology described in (28). Cells were harvested by centrifugation and osmotically lysed at 3.5 x10^8 cells/ml for 5 min at 37 °C in the presence of 0.1 µM 1-chloro-3-tosylamide-7-amino-2-heptone (TLCK), 1mM benzamidine, 1mM phenyl-methyl sulfonyl fluoride (PMSF), 1 µg/ml leupeptin, 1 µg/ml aprotinin and Phosphatase Inhibitor Mixture II (Calbiochem). Cell ghosts from a total of 3.5 x 10^9 trypanosomes, enriched for non-VSG cellular glycoproteins, were collected by centrifugation (16,000 g, 15 min, 4°C) and solubilised in 250 µl of detergent buffer (4% SDS, 0.1 M DTT, 0.1 M Tris-HCl, pH 7.5) using probe sonication for 30 s before and after heating to 85°C for 20 min. S-alkylation was performed by mixing with 250 µl of 8 M urea, 0.5 M iodoacetamide (IAA), 0.1 M Tris-HCl, pH 8.5, for 1 h, room temperature, in the dark. Unreacted IAA was quenched by the addition of 10 µl of the detergent buffer. After centrifugation (16,000 x g, 15 min) the supernatant was transferred to a filtration device with a 30 kDa molecular cut-off (Sartorius) and the majority of the detergent was removed by diafiltration with 8 M urea, 0.1 M Tris-HCl, pH 8.5 followed by lectin binding buffer (1 mM $CaCl_2$, 1 mM $MnCl_2$, 150 mM NaCl in 40 mM Tris-HCl, pH 7.4). An aliquot (400 µg protein) was mixed with 150 µl packed volume of ricin (RCA_{120})-agarose beads (Vector Laboratories) and rotated gently for 2 h. The beads were washed with lectin binding buffer and eluted with 300 µl of the same buffer containing 30 mg/ml lactose and 30 mg/ml galactose (Sigma-Aldrich) in 40 mM Tris-HCl, pH 7.4. After overnight incubation, the supernatant containing the eluted glycoproteins were collected by centrifugation (10,000 x g for 10 min). ConA-coupled agarose beads (0.15 ml packed volume, Vector Laboratories) was added to the recovered supernatant from the RCA_{120} pull-down and incubated at 4 °C for overnight. The beads were washed with lectin binding buffer and eluted with 300 µl of 0.5 M methyl-alpha-D-mannopyranoside (Sigma) in 40 mM Tris-HCl, pH 7.4. After gentle rotation for 2 h the supernatant containing glycoproteins were recovered by centrifugation.

The RCA_{120} and ConA eluted glycoproteins were mixed, transferred to a 30 kDa filter, the buffer exchanged with 25 mM ammonium acetate, pH 5.5, and subjected to digestion with 180 mU endoglycosidase H (EndoH, Roche Applied Sciences) overnight at 37 ºC. The EndoH released glycans were removed by centrifugation and the remaining material was exchanged into 40 mM ammonium bicarbonate buffer in $H_2[^{18}O]$ (Sigma), and subsequently digested with 100 units of *N*-glycosidase F (PNGaseF, Roche) dissolved in $H_2[^{18}O]$. PNGaseF in the presence of $H_2[^{18}O]$ converts Asn to $[^{18}O]$Asp with a mass increment of 2.9890 Da that can be readily distinguished from spontaneous deamidation (mass increment of 0.9858 Da). After overnight incubation at 37 ºC, the PNGaseF released glycans were removed by centrifugation and the deglycosylated proteins were diafiltered into 40 mM ammonium bicarbonate and subsequently digested with a mixture of 1:100 (enzyme: substrate) Lys-C and 1:20 trypsin (Roche) at 37 ºC for 48 h. The peptides were collected by centrifugation through the filter, dried in Speedvac and desalted using Zip Tip C18 micro column (10 µl, Merck Millipore) prior liquid chromatography tandem mass spectrometry (LC-MS/MS).

An alternative (glyco)peptide-based approach was also performed, based on the method of (77). An aliquot of the denatured and S-alkylated sample (400 µg) was first digested on a 10 kDa filter with a mix of Lys-C and trypsin, as described above. The (glyco)peptides were collected by centrifugation through the filter in lectin binding buffer and were mixed with lectin solution containing a mixture of ConA and RCA_{120} resulting in mixtures of (glyco)peptides and lectins with a mass proportion of 1:2. After gentle rotation for 2 h the mixtures were transferred to a 30 kDa filter, the lectin-captured glycopeptides were diafiltered into 25 mM ammonium acetate, pH 5.5, and subjected to EndoH digestion. After overnight incubation at 37ºC, the EndoH released peptides (Asn *N*-GlcNAc residue) were collected by gentle centrifugation. The peptides containing EndoH resistant glycans still bound to lectins were diafiltered into 40 mM ammonium bicarbonate buffer in $H_2[^{18}O]$ and digested with PNGaseF overnight at 37ºC. The de-glycosylated peptides were collected by centrifugation, mixed with the EndoH released fraction, dried in a Speedvac and desalted using ZipTip C18 prior LC-MS/MS, which was performed as described above.

LC-MS/MS analysis and data processing. For glycoproteomics, the LC was performed on a fully automated Ultimate U3000 Nano LC System (Dionex) fitted with a C18 trap- (PepMap nanoViper, Thermo Scientific) and resolving columns (PepMap RSLC) with inner diameters of 100 and 75 µm and lengths of 2 and 50 cm, respectively. Mobile phases consisted of 0.1% formic acid (Sigma) in 2% acetonitrile (Merck, Darmstadt Germany) for solvent A and 0.08% formic acid in 80% acetonitrile for solvent B. Samples were loaded in solvent A. A linear gradient was set as follows: 0% B for 5 min, then a gradient up to 40% B in 122 min and to 98% B in 10 min. A 20 min wash at 98% B is used to keep the column sensitive and prevent carryover, and a 20-min equilibration with 2% B completed the gradient. The LC system was coupled to an LTQ-Orbitrap Velos mass spectrometer (Thermo Scientific) equipped with an Easy spray ion source and operated in positive ion mode. The spray voltage was set to 2 kV, and the ion transfer tube at 250 ºC. The full scans were acquired in a Fourier transform MS mass analyser that covered an *m/z* range of 335-1800 at a resolution of 60 000. The MS/MS analysis were performed under data-dependent mode to fragment the top 15 precursors using collision induced dissociation (CID). A normalised collision energy of -35 eV, an isolation width of *m/z* 2.0, an activation Q value of 0.250, and a time of 100 ms were used. The raw files were converted to mgf format by MSConvert software from ProteoWizard (proteowizard.sourceforge.net). The searches were carried out against the *T. brucei* 927 annotated proteins database (v.8.0, downloaded from TriTrypDB (78), www.tritrypdb.org/) using Mascot software (v.2.4.0, Matrix Science Inc., Boston, MA). The search parameters for Mascot software were set as follows: peptide tolerance, 5 ppm; MS/MS tolerance, 0.5 Da; enzyme, trypsin; one missed cleavage allowed; and fixed carbamidomethyl modifications of cysteines. Oxidation of methionine, *N*-acetylglucosamine modification of Asn and deamidation of Asn to Asp containing a single ^{18}O atom (2.9890 Da mass increase) are used as variable modifications.

Dataset Extraction. We extracted from the mascot result files all the peptides with an ion-score >20. From this we selected 350 glycosylated sites with a deamidation (Asn to $[^{18}O]$Asp conversion) or Asn-*N*-HexNAc modification embedded in the N.^P[ST] consensus. The list was used to count the number of times (>=3) that these changes were detected for each peptide. To increase the number of modified peptides we decided to re-process a previously published work in our laboratory (28). This dataset was re-searched with the same mascot parameters and database used for this publication. This made it possible to include 14 new peptides preferentially HexNac modified and compile a list of 186 peptides used for the next phase of machine learning implemented in python with the scikit-learn package (79).

The data set reported here and that from (28) identified 170 common and 180 and 155 unique glycosylation sites, respectively (Fig S1A). To check the consistency of the two datasets, we selected 92 peptides that were observed >= 4 times in both datasets and plotted the frequencies of the Asn-*N*-HexNAc and Asn to [^{18}O]Asp modifications. The two datasets had a good correlation (r^2 = 0.83) (Fig S1B). However, the experimental procedures used to generate the 2009 dataset (i.e., without the use of $H_2{}^{18}O$) cannot discriminate between the spontaneous non-enzymatic deamidation of asparagine to aspartic acid versus the PNGaseF-mediated deamidation produced during the cleavage of the *N*-glycan. From the linear regression, we could deduce that non-enzymatic deamidation contributed a significant amount (about 18%) of the total deamidation seen in the 2009 dataset (Fig S1B). For this reason, we only used Asn-*N*-HexNAc containing (TbSTT3B substrate) peptides (with a frequency of >= 0.6) from this dataset to augment our machine learning training set.

Machine Learning. The deamidation proportion (DP) was computed for each peptide as (DC / DC + HC) where DC is the Deamidation Count (i.e., the number of times a peptide with [^{18}O]Asp is detected) and HC is the HexNAc Count (i.e., the number of times a peptide with Asn-*N*-HexNAc is detected). This score was used to classify each peptide as preferentially deamidated (score>0.8 n=70) or preferentially HexNac modified (score<0.3% n=56). This dataset was used to extract sequence based feature from 10 amino acids before and after the glycosylated asparagine with the ASAP package in Python (80). We also added some in-house features derived from the knowledge of the *Tb*STT3A recognition and transfer experiment reported in (Figure 4B and Table 1). To this end we created 'Bonus Features' for each residue position probed in that experiment based on the increased transfer efficiency observed when that site is occupied by an Asp (or, by inference, a Glu residue). We also included a 'Bonus All' feature that summed all of the bonus scores for the peptide when it contained more than one Asp and/or Glu residue and a 'Bonus Max' feature that selected only the highest bonus score in such cases. Finally, we also created 'Bonus Presence D' and 'Bonus Presence E' features that simply recorded the presence or absence of Asp or Glu, respectively, at each residue location. We then developed three machine learning algorithms: a random forest classifier (RFC), an extra tree classifier (ETC) and a support vector machine classifier (SVM). The predictors were used to extract the importance of all the features and to rank the features with recursive feature elimination and cross-validated selection of the best number of features (RFECV methodology). The selected features were used to train the three machine learning algorithms (RFC, ETC, SVM) that were further optimized using a Bayesian global optimization with Gaussian processes (https://github.com/fmfn/BayesianOptimization). The optimized machine learning algorithms were combined in a voting classifier to produce our final predictor. The ability of the developed classifiers (RFC, ETC, SVM and Voting Classifier) to discriminate between the deamidated or HexNac modified peptides was assessed with the area under the curve (AUC) of the receiver operating characteristic (ROC) curve. The AUC score was computed 100 times with a five-fold cross validation, using each time a different random split of the original dataset (Figure S5).

CONFLICTS OF INTEREST: The authors declare that they have no conflicts of interest with the contents of this article.

AUTHOR CONTRIBUTIONS: M.A.J.F. conceived and coordinated the study. A.J., L.A., and M.A.J.F. designed the experiments. A.J. performed the biochemical, SILAC proteomic and molecular modelling experiments. L.A. performed the glycoproteomic experiments. M.T. performed the bioinformatic analyses and designed the machine–learning experiments.

ACKNOWLEDGEMENTS: This work was supported by a Wellcome Trust Investigator Award to M.A.J.F. (101842). Wellcome Trust grant 097045/B/11/Z provided mass spectrometry and proteomics infrastructure support. A.J. thanks the School of Life Sciences, University of Dundee, for a PhD studentship. We thank the Dundee Proteomics facility team for their help and Amy Tavendale for assistance with processing the SILAC data. We thank Angela Mehlert, Lucia Güther, Mick Urbaniak and Jim Proctor for helpful advice and comments and Markus Aebi for helpful discussions and kindly sharing information prior to publication.

REFERENCES

1. Schwede, A., Macleod, O. J. S., MacGregor, P., and Carrington, M. (2015) How Does the VSG Coat of Bloodstream Form African Trypanosomes Interact with External Proteins? *PLoS Pathog* **11**, e1005259
2. Horn, D. (2014) Antigenic variation in African trypanosomes. *Mol Biochem Parasitol* **195**, 123-129
3. Salmon, D., Geuskens, M., Hanocq, F., Hanocq-Quertier, J., Nolan, D., Ruben, L., and Pays, E. (1994) A novel heterodimeric transferrin receptor encoded by a pair of VSG expression site-associated genes in T. brucei. *Cell* **78**, 75-86
4. Steverding, D. (2000) The transferrin receptor of Trypanosoma brucei. *Parasitol Int* **48**, 191-198
5. Mehlert, A., Wormald, M. R., and Ferguson, M. A. J. (2012) Modeling of the N-glycosylated transferrin receptor suggests how transferrin binding can occur within the surface coat of Trypanosoma brucei. *PLoS Pathog* **8**, e1002618
6. Peck, R. F., Shiflett, A. M., Schwartz, K. J., McCann, A., Hajduk, S. L., and Bangs, J. D. (2008) The LAMP-like protein p67 plays an essential role in the lysosome of African trypanosomes. *Mol Microbiol* **68**, 933-946
7. Jackson, A. P., Allison, H. C., Barry, J. D., Field, M. C., Hertz-Fowler, C., and Berriman, M. (2013) A cell-surface phylome for African trypanosomes. *PLoS Negl Trop Dis* **7**, e2121
8. Allison, H., O'Reilly, A. J., Sternberg, J., and Field, M. C. (2014) An extensive endoplasmic reticulum-localised glycoprotein family in trypanosomatids. *Microb Cell* **1**, 325-345
9. Lingnau, A., Zufferey, R., Lingnau, M., and Russell, D. G. (1999) Characterization of tGLP-1, a Golgi and lysosome-associated, transmembrane glycoprotein of African trypanosomes. *J Cell Sci* **112 Pt 18**, 3061-3070
10. Engstler, M., Weise, F., Bopp, K., Grunfelder, C. G., Gunzel, M., Heddergott, N., and Overath, P. (2005) The membrane-bound histidine acid phosphatase TbMBAP1 is essential for endocytosis and membrane recycling in Trypanosoma brucei. *J Cell Sci* **118**, 2105-2118
11. LaCount, D. J., Barrett, B., and Donelson, J. E. (2002) Trypanosoma brucei FLA1 is required for flagellum attachment and cytokinesis. *J Biol Chem* **277**, 17580-17588
12. Treumann, A., Zitzmann, N., Hulsmeier, A., Prescott, A. R., Almond, A., Sheehan, J., and Ferguson, M. A. (1997) Structural characterisation of two forms of procyclic acidic repetitive protein expressed by procyclic forms of Trypanosoma brucei. *J Mol Biol* **269**, 529-547
13. Acosta-Serrano, A., Cole, R. N., Mehlert, A., Lee, M. G., Ferguson, M. A., and Englund, P. T. (1999) The procyclin repertoire of Trypanosoma brucei. Identification and structural characterization of the Glu-Pro-rich polypeptides. *J Biol Chem* **274**, 29763-29771
14. Guther, M. L., Lee, S., Tetley, L., Acosta-Serrano, A., and Ferguson, M. A. (2006) GPI-anchored proteins and free GPI glycolipids of procyclic form Trypanosoma brucei are nonessential for growth, are required for colonization of the tsetse fly, and are not the only components of the surface coat. *Mol Biol Cell* **17**, 5265-5274
15. Guther, M. L., Beattie, K., Lamont, D. J., James, J., Prescott, A. R., and Ferguson, M. A. (2009) Fate of glycosylphosphatidylinositol (GPI)-less procyclin and characterization of sialylated non-GPI-anchored surface coat molecules of procyclic-form Trypanosoma brucei. *Eukaryot Cell* **8**, 1407-1417
16. Zamze, S. E., Wooten, E. W., Ashford, D. A., Ferguson, M. A., Dwek, R. A., and Rademacher, T. W. (1990) Characterisation of the asparagine-linked oligosaccharides from Trypanosoma brucei type-I variant surface glycoproteins. *Eur J Biochem* **187**, 657-663
17. Zamze, S. E., Ashford, D. A., Wooten, E. W., Rademacher, T. W., and Dwek, R. A. (1991) Structural characterization of the asparagine-linked oligosaccharides from Trypanosoma brucei type II and type III variant surface glycoproteins. *J Biol Chem* **266**, 20244-20261

18. Atrih, A., Richardson, J. M., Prescott, A. R., and Ferguson, M. A. J. (2005) Trypanosoma brucei glycoproteins contain novel giant poly-N-acetyllactosamine carbohydrate chains. *J Biol Chem* **280**, 865-871
19. Izquierdo, L., Atrih, A., Rodrigues, J. A., Jones, D. C., and Ferguson, M. A. J. (2009) Trypanosoma brucei UDP-glucose:glycoprotein glucosyltransferase has unusual substrate specificity and protects the parasite from stress. *Eukaryotic cell* **8**, 230-240
20. Bandini, G., Marino, K., Guther, M. L., Wernimont, A. K., Kuettel, S., Qiu, W., Afzal, S., Kelner, A., Hui, R., and Ferguson, M. A. (2012) Phosphoglucomutase is absent in Trypanosoma brucei and redundantly substituted by phosphomannomutase and phospho-N-acetylglucosamine mutase. *Mol Microbiol* **85**, 513-534
21. Bangs, J. D., Doering, T. L., Englund, P. T., and Hart, G. W. (1988) Biosynthesis of a variant surface glycoprotein of Trypanosoma brucei. Processing of the glycolipid membrane anchor and N-linked oligosaccharides. *J Biol Chem* **263**, 17697-17705
22. Aebi, M. (2013) N-linked protein glycosylation in the ER. *Biochim Biophys Acta* **1833**, 2430-2437
23. Cherepanova, N., Shrimal, S., and Gilmore, R. (2016) N-linked glycosylation and homeostasis of the endoplasmic reticulum. *Curr Opin Cell Biol* **41**, 57-65
24. Acosta-Serrano, A., O'Rear, J., Quellhorst, G., Lee, S. H., Hwa, K. Y., Krag, S. S., and Englund, P. T. (2004) Defects in the N-linked oligosaccharide biosynthetic pathway in a Trypanosoma brucei glycosylation mutant. *Eukaryot Cell* **3**, 255-263
25. Manthri, S., Guther, M. L., Izquierdo, L., Acosta-Serrano, A., and Ferguson, M. A. (2008) Deletion of the TbALG3 gene demonstrates site-specific N-glycosylation and N-glycan processing in Trypanosoma brucei. *Glycobiology* **18**, 367-383
26. Jones, D. C., Mehlert, A., Guther, M. L., and Ferguson, M. A. (2005) Deletion of the glucosidase II gene in Trypanosoma brucei reveals novel N-glycosylation mechanisms in the biosynthesis of variant surface glycoprotein. *J Biol Chem* **280**, 35929-35942
27. Izquierdo, L., Mehlert, A., and Ferguson, M. A. J. (2012) The lipid-linked oligosaccharide donor specificities of Trypanosoma brucei oligosaccharyltransferases. *Glycobiology* **22**, 696-703
28. Izquierdo, L., Schulz, B. L., Rodrigues, J. A., Guther, M. L., Procter, J. B., Barton, G. J., Aebi, M., and Ferguson, M. A. (2009) Distinct donor and acceptor specificities of Trypanosoma brucei oligosaccharyltransferases. *EMBO J* **28**, 2650-2661
29. Castro, O., Movsichoff, F., and Parodi, A. J. (2006) Preferential transfer of the complete glycan is determined by the oligosaccharyltransferase complex and not by the catalytic subunit. *Proc Natl Acad Sci U S A* **103**, 14756-14760
30. Kelleher, D. J., Banerjee, S., Cura, A. J., Samuelson, J., and Gilmore, R. (2007) Dolichol-linked oligosaccharide selection by the oligosaccharyltransferase in protist and fungal organisms. *J Cell Biol* **177**, 29-37
31. Nasab, F. P., Schulz, B. L., Gamarro, F., Parodi, A. J., and Aebi, M. (2008) All in one: Leishmania major STT3 proteins substitute for the whole oligosaccharyltransferase complex in Saccharomyces cerevisiae. *Molecular biology of the cell* **19**, 3758-3768
32. Hese, K., Otto, C., Routier, F. H., and Lehle, L. (2009) The yeast oligosaccharyltransferase complex can be replaced by STT3 from Leishmania major. *Glycobiology* **19**, 160-171
33. Izquierdo, L., Atrih, A., Rodrigues, J. A., Jones, D. C., and Ferguson, M. A. (2009) Trypanosoma brucei UDP-glucose:glycoprotein glucosyltransferase has unusual substrate specificity and protects the parasite from stress. *Eukaryot Cell* **8**, 230-240
34. Damerow, M., Rodrigues, J. A., Wu, D., Guther, M. L., Mehlert, A., and Ferguson, M. A. (2014) Identification and functional characterization of a highly divergent N-acetylglucosaminyltransferase I (TbGnTI) in Trypanosoma brucei. *J Biol Chem* **289**, 9328-9339
35. Damerow, M., Graalfs, F., Guther, M. L., Mehlert, A., Izquierdo, L., and Ferguson, M. A. (2016) A Gene of the beta3-Glycosyltransferase Family Encodes N-

Acetylglucosaminyltransferase II Function in Trypanosoma brucei. *J Biol Chem* **291**, 13834-13845

36. Urbaniak, M. D., Martin, D. M., and Ferguson, M. A. (2013) Global quantitative SILAC phosphoproteomics reveals differential phosphorylation is widespread between the procyclic and bloodstream form lifecycle stages of Trypanosoma brucei. *Journal of proteome research* **12**, 2233-2244
37. Bangs, J. D., Brouch, E. M., Ransom, D. M., and Roggy, J. L. (1996) A soluble secretory reporter system in Trypanosoma brucei. Studies on endoplasmic reticulum targeting. *J Biol Chem* **271**, 18387-18393
38. Wirtz, E., Leal, S., Ochatt, C., and Cross, G. A. (1999) A tightly regulated inducible expression system for conditional gene knock-outs and dominant-negative genetics in Trypanosoma brucei. *Mol Biochem Parasitol* **99**, 89-101
39. Mehlert, A., Sullivan, L., and Ferguson, M. A. J. (2010) Glycotyping of Trypanosoma brucei variant surface glycoprotein MITat1.8. *Mol Biochem Parasitol* **174**, 74-77
40. Cardoso de Almeida, M. L., and Turner, M. J. (1983) The membrane form of variant surface glycoproteins of Trypanosoma brucei. *Nature* **302**, 349-352
41. Ferguson, M. A., Haldar, K., and Cross, G. A. (1985) Trypanosoma brucei variant surface glycoprotein has a sn-1,2-dimyristyl glycerol membrane anchor at its COOH terminus. *J Biol Chem* **260**, 4963-4968
42. Crooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E. (2004) WebLogo: a sequence logo generator. *Genome Res* **14**, 1188-1190
43. Vacic, V., Iakoucheva, L. M., and Radivojac, P. (2006) Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics* **22**, 1536-1537
44. Miller, M. L., Jensen, L. J., Diella, F., Jorgensen, C., Tinti, M., Li, L., Hsiung, M., Parker, S. A., Bordeaux, J., Sicheritz-Ponten, T., Olhovsky, M., Pasculescu, A., Alexander, J., Knapp, S., Blom, N., Bork, P., Li, S., Cesareni, G., Pawson, T., Turk, B. E., Yaffe, M. B., Brunak, S., and Linding, R. (2008) Linear motif atlas for phosphorylation-dependent signaling. *Science signaling* **1**, ra2
45. Poljak. (2017) Analysis of Substrate Specificity of Trypanosoma brucei OSTs by Functional Expression in Yeast, Submitted. *JBC*
46. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., and Sternberg, M. J. (2015) The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* **10**, 845-858
47. Lizak, C., Gerber, S., Numao, S., Aebi, M., and Locher, K. P. (2011) X-ray structure of a bacterial oligosaccharyltransferase. *Nature* **474**, 350-355
48. Krissinel, E., and Henrick, K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta crystallographica. Section D, Biological crystallography* **60**, 2256-2268
49. Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004) UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of computational chemistry* **25**, 1605-1612
50. Kalb, L. C., Frederico, Y. C., Boehm, C., Moreira, C. M., Soares, M. J., and Field, M. C. (2016) Conservation and divergence within the clathrin interactome of Trypanosoma cruzi. *Sci Rep* **6**, 31212
51. Manna, P. T., Obado, S. O., Boehm, C., Gadelha, C., Sali, A., Chait, B. T., Rout, M. P., and Field, M. C. (2017) Lineage-specific proteins essential for endocytosis in trypanosomes. *J Cell Sci* **130**, 1379-1392
52. Boehm, C. M., Obado, S., Gadelha, C., Kaupisch, A., Manna, P. T., Gould, G. W., Munson, M., Chait, B. T., Rout, M. P., and Field, M. C. (2017) The Trypanosome Exocyst: A Conserved Structure Revealing a New Role in Endocytosis. *PLoS Pathog* **13**, e1006063

53. Obado, S. O., Brillantes, M., Uryu, K., Zhang, W., Ketaren, N. E., Chait, B. T., Field, M. C., and Rout, M. P. (2016) Interactome Mapping Reveals the Evolutionary History of the Nuclear Pore Complex. *PLoS Biol* **14**, e1002365

54. Maishman, L., Obado, S. O., Alsford, S., Bart, J. M., Chen, W. M., Ratushny, A. V., Navarro, M., Horn, D., Aitchison, J. D., Chait, B. T., Rout, M. P., and Field, M. C. (2016) Co-dependence between trypanosome nuclear lamina components in nuclear stability and control of gene expression. *Nucleic Acids Res* **44**, 10554-10570

55. Nagamune, K., Ohishi, K., Ashida, H., Hong, Y., Hino, J., Kangawa, K., Inoue, N., Maeda, Y., and Kinoshita, T. (2003) GPI transamidase of Trypanosoma brucei has two previously uncharacterized (trypanosomatid transamidase 1 and 2) and three common subunits. *Proc Natl Acad Sci U S A* **100**, 10682-10687

56. Choi, B. K., Warburton, S., Lin, H., Patel, R., Boldogh, I., Meehl, M., d'Anjou, M., Pon, L., Stadheim, T. A., and Sethuraman, N. (2012) Improvement of N-glycan site occupancy of therapeutic glycoproteins produced in Pichia pastoris. *Appl Microbiol Biotechnol* **95**, 671-682

57. Yan, A., Wu, E., and Lennarz, W. J. (2005) Studies of yeast oligosaccharyl transferase subunits using the split-ubiquitin system: topological features and in vivo interactions. *Proc Natl Acad Sci U S A* **102**, 7121-7126

58. Ferguson, M. A., Duszenko, M., Lamont, G. S., Overath, P., and Cross, G. A. (1986) Biosynthesis of Trypanosoma brucei variant surface glycoproteins. N-glycosylation and addition of a phosphatidylinositol membrane anchor. *J Biol Chem* **261**, 356-362

59. Siegel, T. N., Hekstra, D. R., Wang, X., Dewell, S., and Cross, G. A. (2010) Genome-wide analysis of mRNA abundance in two life-cycle stages of Trypanosoma brucei and identification of splicing and polyadenylation sites. *Nucleic Acids Res* **38**, 4946-4957

60. Butter, F., Bucerius, F., Michel, M., Cicova, Z., Mann, M., and Janzen, C. J. (2013) Comparative proteomics of two life cycle stages of stable isotope-labeled Trypanosoma brucei reveals novel components of the parasite's host adaptation machinery. *Molecular & cellular proteomics : MCP* **12**, 172-179

61. Urbaniak, M. D., Guther, M. L., and Ferguson, M. A. (2012) Comparative SILAC proteomic analysis of Trypanosoma brucei bloodstream and procyclic lifecycle stages. *PloS one* **7**, e36619

62. Castillo-Acosta, V. M., Ruiz-Perez, L. M., Etxebarria, J., Reichardt, N. C., Navarro, M., Igarashi, Y., Liekens, S., Balzarini, J., and Gonzalez-Pacanowska, D. (2016) Carbohydrate-Binding Non-Peptidic Pradimicins for the Treatment of Acute Sleeping Sickness in Murine Models. *PLoS Pathog* **12**, e1005851

63. Castillo-Acosta, V. M., Ruiz-Perez, L. M., Van Damme, E. J., Balzarini, J., and Gonzalez-Pacanowska, D. (2015) Exposure of Trypanosoma brucei to an N-acetylglucosamine-binding lectin induces VSG switching and glycosylation defects resulting in reduced infectivity. *PLoS Negl Trop Dis* **9**, e0003612

64. Castillo-Acosta, V. M., Vidal, A. E., Ruiz-Perez, L. M., Van Damme, E. J., Igarashi, Y., Balzarini, J., and Gonzalez-Pacanowska, D. (2013) Carbohydrate-binding agents act as potent trypanocidals that elicit modifications in VSG glycosylation and reduced virulence in Trypanosoma brucei. *Mol Microbiol* **90**, 665-679

65. Ruiz-Canada, C., Kelleher, D. J., and Gilmore, R. (2009) Cotranslational and posttranslational N-glycosylation of polypeptides by distinct mammalian OST isoforms. *Cell* **136**, 272-283

66. Shrimal, S., Cherepanova, N. A., and Gilmore, R. (2015) Cotranslational and posttranslocational N-glycosylation of proteins in the endoplasmic reticulum. *Semin Cell Dev Biol* **41**, 71-78

67. Shrimal, S., Cherepanova, N. A., and Gilmore, R. (2015) Cotranslational and posttranslocational N-glycosylation of proteins in the endoplasmic reticulum. *Seminars in cell & developmental biology* **41**, 71-78

68. Cherepanova, N., Shrimal, S., and Gilmore, R. (2016) N-linked glycosylation and homeostasis of the endoplasmic reticulum. *Current opinion in cell biology* **41**, 57-65

69. Cherepanova, N. A., Shrimal, S., and Gilmore, R. (2014) Oxidoreductase activity is necessary for N-glycosylation of cysteine-proximal acceptor sites in glycoproteins. *J Cell Biol* **206**, 525-539

70. Mohorko, E., Owen, R. L., Malojcic, G., Brozzo, M. S., Aebi, M., and Glockshuber, R. (2014) Structural basis of substrate specificity of human oligosaccharyl transferase subunit N33/Tusc3 and its role in regulating protein N-glycosylation. *Structure* **22**, 590-601

71. Schulz, B. L., Stirnimann, C. U., Grimshaw, J. P., Brozzo, M. S., Fritsch, F., Mohorko, E., Capitani, G., Glockshuber, R., Grutter, M. G., and Aebi, M. (2009) Oxidoreductase activity of oligosaccharyltransferase subunits Ost3p and Ost6p defines site-specific glycosylation efficiency. *Proc Natl Acad Sci U S A* **106**, 11061-11066

72. Schulz, B. L., and Aebi, M. (2009) Analysis of glycosylation site occupancy reveals a role for Ost3p and Ost6p in site-specific N-glycosylation efficiency. *Molecular & cellular proteomics : MCP* **8**, 357-364

73. Woycechowsky, K. J., and Raines, R. T. (2003) The CXC motif: a functional mimic of protein disulfide isomerase. *Biochemistry* **42**, 5387-5394

74. Tiengwe, C., Muratore, K. A., and Bangs, J. D. (2016) Surface proteins, ERAD and antigenic variation in Trypanosoma brucei. *Cell Microbiol* **18**, 1673-1688

75. Hirumi, H., and Hirumi, K. (1989) Continuous cultivation of Trypanosoma brucei blood stream forms in a medium containing a low concentration of serum protein without feeder cell layers. *J Parasitol* **75**, 985-989

76. Oberholzer, M., Morand, S., Kunz, S., and Seebeck, T. (2006) A vector series for rapid PCR-mediated C-terminal in situ tagging of Trypanosoma brucei genes. *Mol Biochem Parasitol* **145**, 117-120

77. Zielinska, D. F., Gnad, F., Wisniewski, J. R., and Mann, M. (2010) Precision mapping of an in vivo N-glycoproteome reveals rigid topological and sequence constraints. *Cell* **141**, 897-907

78. Aslett, M., Aurrecoechea, C., Berriman, M., Brestelli, J., Brunk, B. P., Carrington, M., Depledge, D. P., Fischer, S., Gajria, B., Gao, X., Gardner, M. J., Gingle, A., Grant, G., Harb, O. S., Heiges, M., Hertz-Fowler, C., Houston, R., Innamorato, F., Iodice, J., Kissinger, J. C., Kraemer, E., Li, W., Logan, F. J., Miller, J. A., Mitra, S., Myler, P. J., Nayak, V., Pennington, C., Phan, I., Pinney, D. F., Ramasamy, G., Rogers, M. B., Roos, D. S., Ross, C., Sivam, D., Smith, D. F., Srinivasamoorthy, G., Stoeckert, C. J., Jr., Subramanian, S., Thibodeau, R., Tivey, A., Treatman, C., Velarde, G., and Wang, H. (2010) TriTrypDB: a functional genomic resource for the Trypanosomatidae. *Nucleic Acids Res* **38**, D457-462

79. Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., and Varoquaux, G. (2014) Machine learning for neuroimaging with scikit-learn. *Front Neuroinform* **8**, 14

80. Brandes, N., Ofer, D., and Linial, M. (2016) ASAP: a machine learning framework for local protein properties. *Database (Oxford)* **2016**

FIGURE LEGENDS:

Figure 1. Denaturing and blue native gel electrophoresis of *Tb*STT3A-HA$_3$. SDS-PAGE and anti-HA Western blotting (panel A) or blue native gel electrophoresis and anti-HA Western blotting (panel B) of anti-HA/protein G magnetic bead pull-outs from digitonin lysates of wild-type (lane 1) and *in situ* *Tb*STT3A-HA$_3$ tagged (lane 2) bloodstream form trypanosomes.

Figure 2. Overview of the SILAC pull-out experiment and plot of proteomics data. Panel A: Overview of the SILAC experiment. Wild-type bloodstream form cells were grown in light (R_0K_0) medium, and cells expressing i*n situ*-tagged *Tb*STT3A-HA$_3$ were labelled with heavy (R_6K_4) medium. The cells were mixed 1:1 and *Tb*STT3A-HA$_3$ was enriched by affinity-selection on anti-HA magnetic beads. Peptides from *Tb*STT3A-HA$_3$ and genuine binding proteins have high heavy/light isotope ratios, whereas those from contaminants will have ratios close to 1:1 ($Log_2 = 0$). Panel B: Results from the *Tb*STT3A SILAC pull-out experiment. The plot shows the Log_2 of the heavy-to-light isotope ratio (x axis) versus the Log_{10} value of the intensities of the peptides belonging to each protein that was detected (y axis). The black curves (marked sigma 3) represent three standard deviations from the mean. Proteins plotted in orange have a heavy-to-light ratio above the sigma 3 cut off and are significantly enriched. *Tb*STT3A-HA$_3$ (bait) and *Tb*STT3B (both annotated) were shown to be highly enriched and are highlighted in red. Panel C: Results from the *Tb*STT3B-MYC$_3$ SILAC pull-out experiment. The plot is the same as in panel B except that *in situ*-tagged *Tb*STT3B-MYC$_3$ was used as bait. Again, *Tb*STT3A and *Tb*STT3A (both annotated and highlighted in red) were significantly enriched.

Figure 3. Co-immunoprecipitation of *Tb*STT3A and *Tb*STT3B. Digitonin lysates from wild type cells (lanes 1, 3, 5 and 7) and *Tb*STT3A-HA$_3$ and *Tb*STT3B-MYC$_3$ double *in-situ* tagged cells (lanes 2, 4, 6 and 8) were subjected to IP and Western blotted with anti-HA or anti-MYC antibodies, as indicated. The red rectangles highlight the bands corresponding to *Tb*STT3A-HA$_3$ (lanes 2 and 4) and *Tb*STT3B-MYC$_3$ (lanes 6 and 8).

Figure 4. *In vivo* assay of TbSTT3A substrate specificity. Panel A: The constructs described in (Table 1) were expressed in bloodstream form trypanosomes and the resulting TbBiP*N*-[XXXXXNXTXXXXX]-HA$_3$ reporter glycoproteins were visualised in cell lysates by SDS-PAGE and anti-HA Western blotting. Representative examples are shown for the sequences indicated (lanes 1-3). The proportions of the anti-HA signals that were sensitive (lower bands)

and resistant (upper bands) to EndoH were quantified and are reported in (Table 1). Panel B: Summary of the quantitative data from (Table 1) showing the influence of replacing single neutral Ala residues with an acidic Asp residue at each possible position.

Figure 5. Glycoproteomic data logos and predictor. The amino acid frequencies of the preferentially deamidated (A) or preferentially HexNAc modified peptides (B) identified by mass spectrometry are visualized with the WebLogo web service. The Two Sample logo web service was used to visualize the amino acids enriched (upper part) or depleted (lower part) in the sequences of the preferentially deamidated peptides identified by mass spectrometry (C) or predicted by the machine learning algorithm (D) by using the preferentially HexNAc modified peptides as negative sample.

Figure 6. Molecular models of the active sites of *Tb*STT3A and *Tb*STT3B. Molecular models of the predicted active sites of *Tb*STT3A (panel A) and *Tb*STT3B (panel B) with an acceptor peptide (GDQNAT) based on the crystal structure of *C. lari* PglB (Lizak et al., 2011). Of note are the residues in red: Arg397 in *Tb*STT3A (His397 in *Tb*STT3B) and Arg406 in *Tb*STT3A (Gly406 in *Tb*STT3B), where the guanidinium cations of the Arg sidechains could interact with acidic residues at or close to the acceptor peptide sequon.

Table 1. Effect of aspartic acid on glycosite recognition by *Tb*STT3A.

Name	**Sequence**	**pI**	**% EndoH-resistant (% STT3A transfer)[a]**
Alanine control (Ala)	AAAAANATAAAAA	6.01	6.8±3.7
D -5	**D**AAAANATAAAAA	3.10	34.4±6.1
D -4	A**D**AAANATAAAAA	3.10	44.8±8.9
D -3	AA**D**AANATAAAAA	3.10	36.9±2.6
D -2	AAA**D**ANATAAAAA	3.10	36.5±7.2
D -1	AAAA**D**NATAAAAA	3.10	70.5±1.1
D +1	AAAAAN**D**TAAAAA	3.10	61.0±11.3
D +3	AAAAANAT**D**AAAA	3.10	24.0±2.7
D +4	AAAAANATA**D**AAA	3.10	16.4±4.8
D +5	AAAAANATAA**D**AA	3.10	21.6±4.4
D +6	AAAAANATAAA**D**A	3.10	27.2±5.8
D +7	AAAAANATAAAA**D**	3.10	10.7±3.1

[a]These mean and standard deviation of the mean figures are based on n=8 from two technical replicates of 3 biological replicates.

Fig. 1

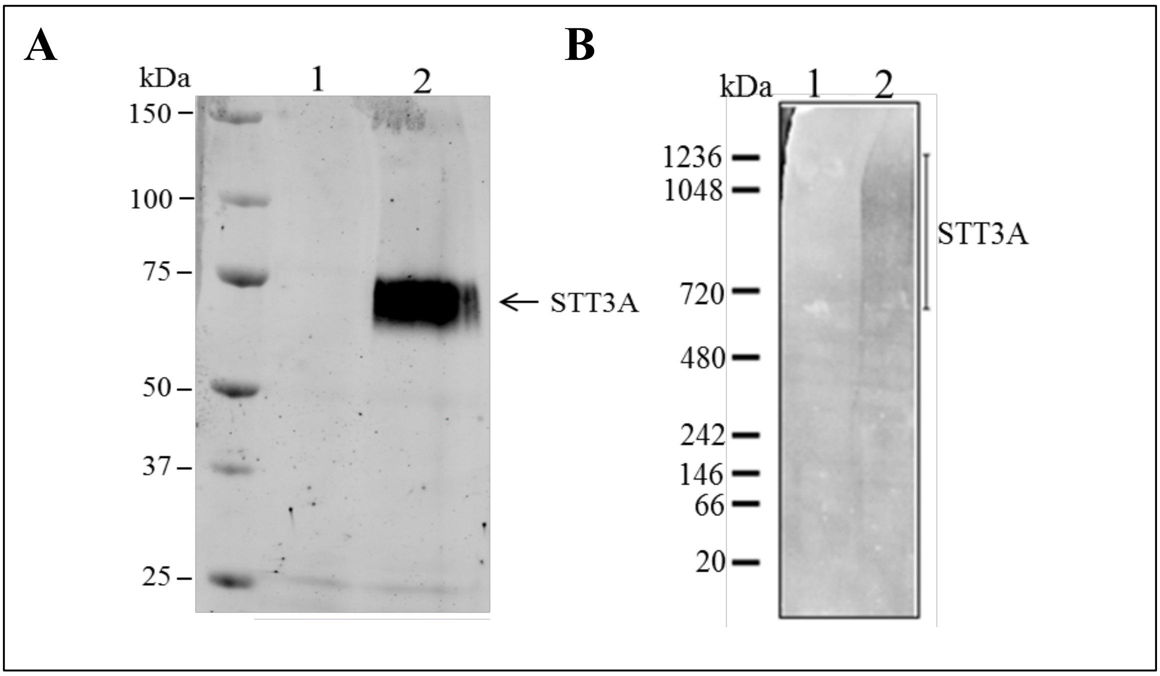

A
kDa
1
2
150
100
75
50
37
25
STT3A
B
kDa
1
2
1236
1048
720
480
242
146
66
20
STT3A

Fig. 2

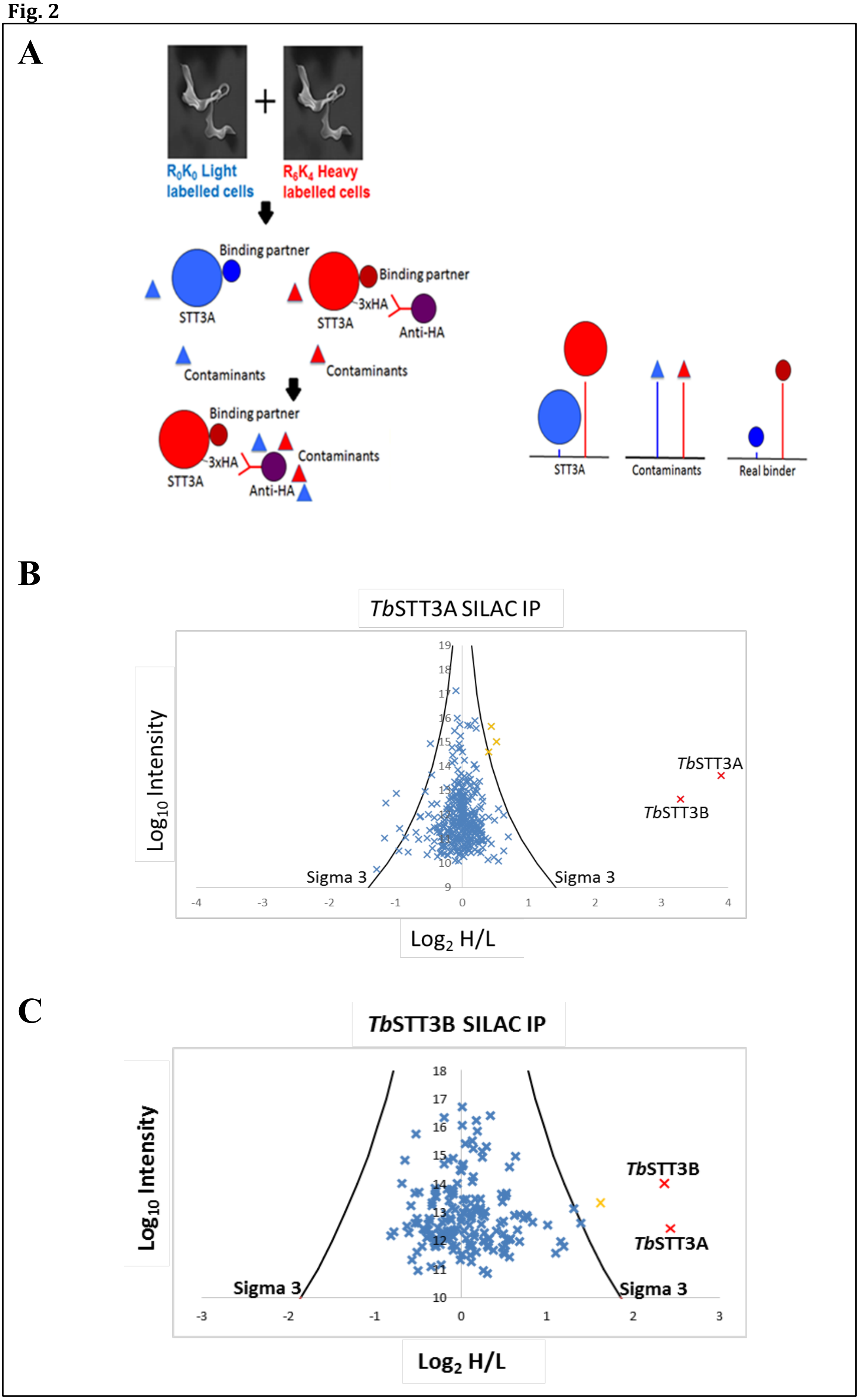

A
R_0K_0 Light labelled cells
R_6K_4 Heavy labelled cells
Binding partner
STT3A
Binding partner
3xHA
STT3A
Anti-HA
Contaminants
Contaminants
Binding partner
3xHA
STT3A
Anti-HA
Contaminants
STT3A
Contaminants
Real binder
B
*Tb*STT3A SILAC IP
Log_{10} Intensity
Log_2 H/L
*Tb*STT3A
*Tb*STT3B
Sigma 3
Sigma 3
19
18
17
16
15
14
13
12
11
10
9
-4
-3
-2
-1
0
1
2
3
4
C
*Tb*STT3B SILAC IP
Log_{10} Intensity
Log_2 H/L
*Tb*STT3B
*Tb*STT3A
Sigma 3
Sigma 3
18
17
16
15
14
13
12
11
10
-3
-2
-1
0
1
2
3

Fig 3

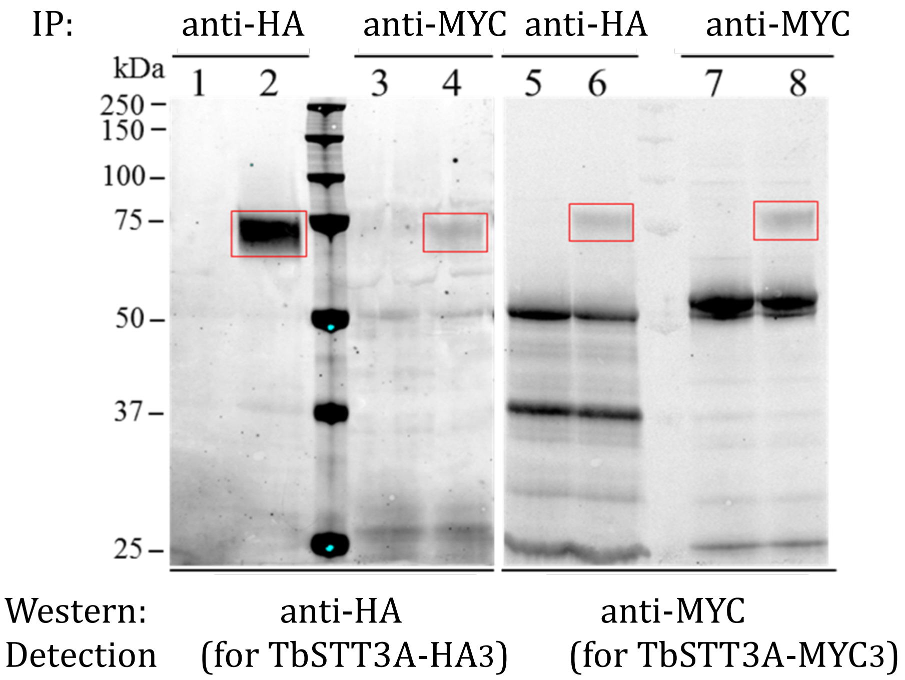

IP:
anti-HA
anti-MYC
anti-HA
anti-MYC
kDa
1
2
3
4
5
6
7
8
250
150
100
75
50
37
25
Western:
Detection
anti-HA
(for TbSTT3A-HA3)
anti-MYC
(for TbSTT3A-MYC3)

Fig 4

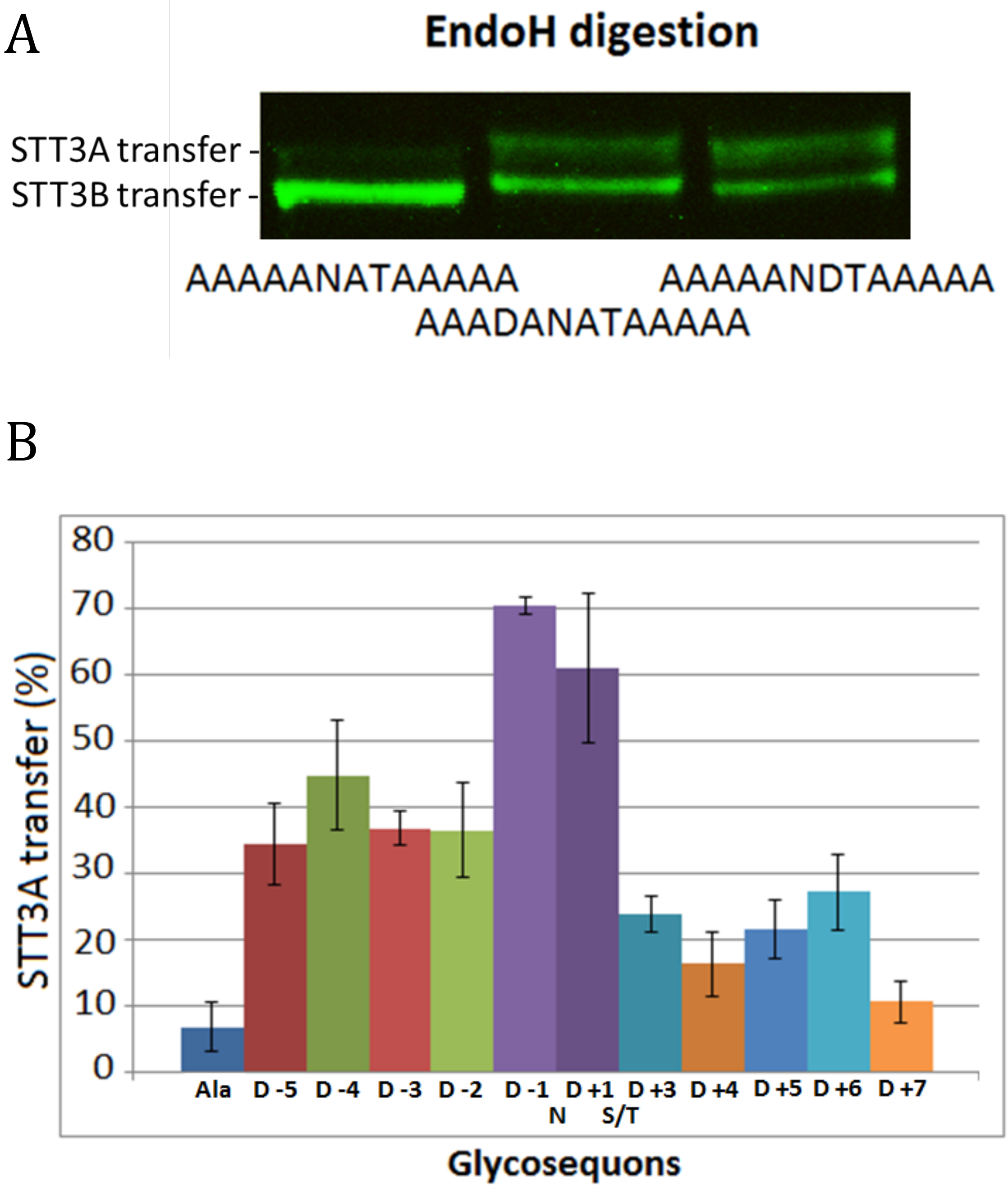

A
EndoH digestion
STT3A transfer -
STT3B transfer -
AAAAANATAAAAA
AAAADANATAAAAA
AAAAANDTAAAAA
B
STT3A transfer (%)
80
70
60
50
40
30
20
10
0
Ala
D -5
D -4
D -3
D -2
D -1
D +1
D +3
D +4
D +5
D +6
D +7
N
S/T
Glycosequons

Fig 5

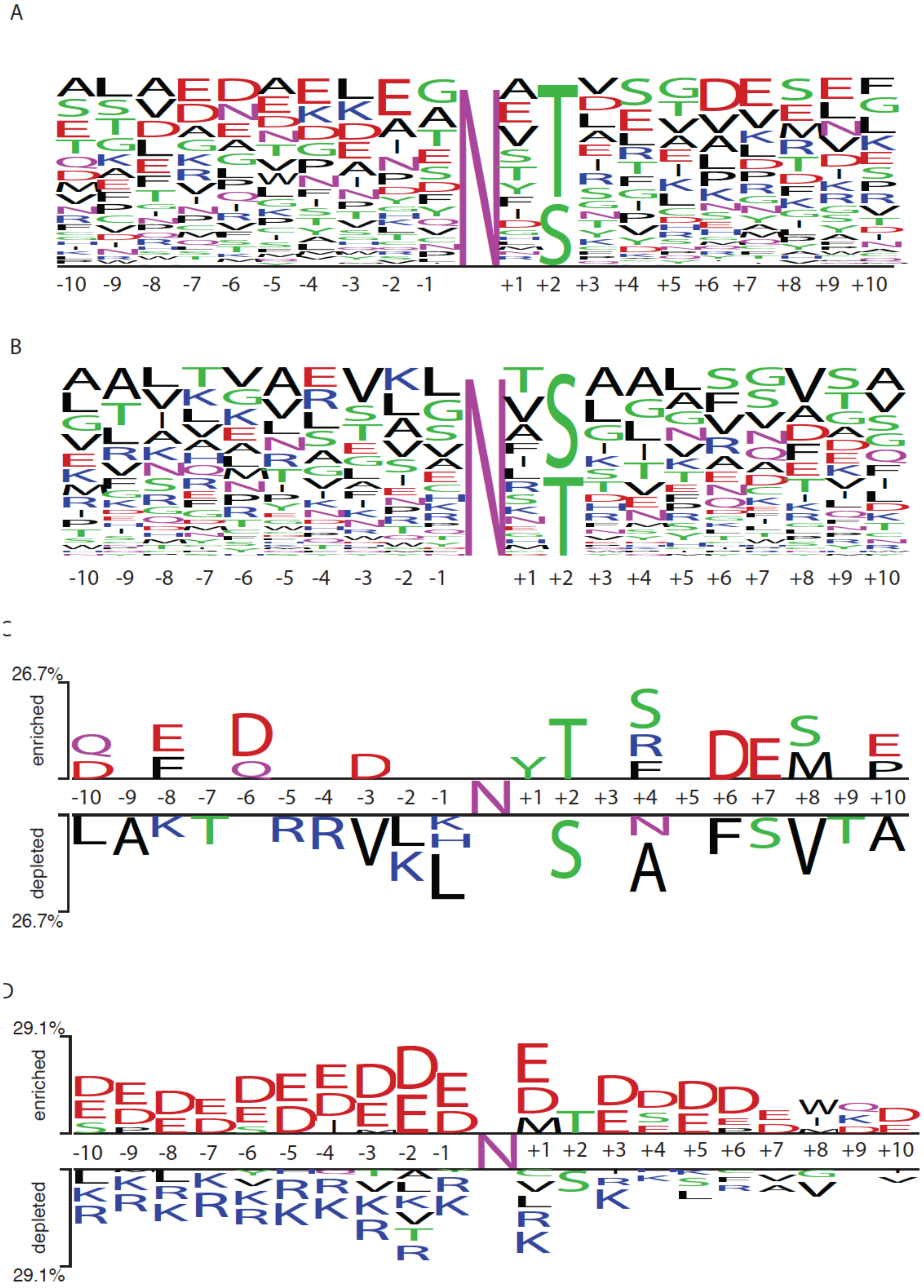
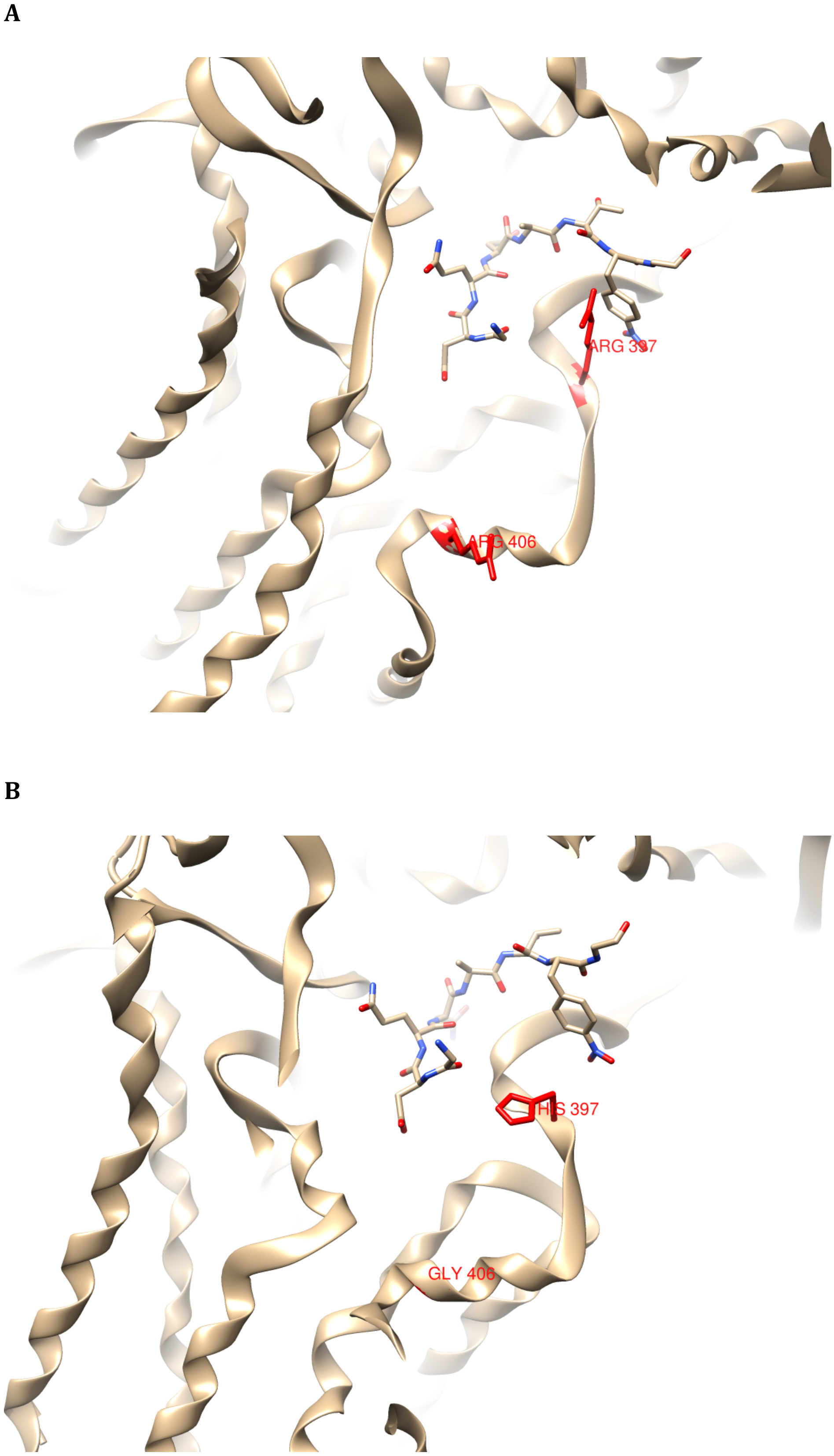

A
-10 -9 -8 -7 -6 -5 -4 -3 -2 -1 +1 +2 +3 +4 +5 +6 +7 +8 +9 +10
B
-10 -9 -8 -7 -6 -5 -4 -3 -2 -1 +1 +2 +3 +4 +5 +6 +7 +8 +9 +10
C
26.7%
enriched
depleted
26.7%
-10 -9 -8 -7 -6 -5 -4 -3 -2 -1 N +1 +2 +3 +4 +5 +6 +7 +8 +9 +10
D
29.1%
enriched
depleted
29.1%
-10 -9 -8 -7 -6 -5 -4 -3 -2 -1 N +1 +2 +3 +4 +5 +6 +7 +8 +9 +10

Fig 6



Single-subunit oligosaccharyltransferases of Trypanosoma brucei display different and predictable peptide acceptor specificities.

Anders A. J. Jinnelov, Liaqat Ali, Michele Tinti and Michael A. J. Ferguson

J. Biol. Chem. published online September 19, 2017