Published in final edited form as:

Semin Cell Dev Biol. 2015 May ; 41: 121–128. doi:10.1016/j.semcdb.2014.11.008.

Effects of *N*-glycan precursor length diversity on quality control of protein folding and on protein glycosylation

John Samuelson* and Phillips W. Robbins

Department of Molecular and Cell Biology, Boston University Goldman School of Dental Medicine, 72 East Concord St, Evans 425, Boston MA, 02118, USA

John Samuelson: jsamuels@bu.edu; Phillips W. Robbins: robbinspw2@gmail.com

Abstract

Asparagine-linked glycans (*N*-glycans) of medically important protists have much to tell us about the evolution of *N*-glycosylation and of *N*-glycan-dependent quality control (*N*-glycan QC) of protein folding in the endoplasmic reticulum. While host *N*-glycans are built upon a dolichol-pyrophosphate-linked precursor with 14 sugars (Glc₃Man₉GlcNAc₂), protist *N*-glycan precursors vary from Glc₃Man₉GlcNAc₂ (*Acanthamoeba*) to Man₉GlcNAc₂ (*Trypanosoma*) to Glc₃Man₅GlcNAc₂ (*Toxoplasma*) to Man₅GlcNAc₂ (*Entamoeba*, *Trichomonas*, and *Eimeria*) to GlcNAc₂ (*Plasmodium* and *Giardia*) to zero (*Theileria*). As related organisms have differing *N*-glycan lengths (e.g. *Toxoplasma*, *Eimeria*, *Plasmodium*, and *Theileria*), the present *N*-glycan variation is based upon secondary loss of Alg genes, which encode enzymes that add sugars to the *N*-glycan precursor. An *N*-glycan precursor with Man₅GlcNAc₂ is necessary but not sufficient for *N*-glycan QC, which is predicted by the presence of the UDP-glucose:glucosyltransferase (UGGT) plus calreticulin and/or calnexin. As many parasites lack glucose in their *N*-glycan precursor, UGGT product may be identified by inhibition of glucosidase II. The presence of an armless calnexin in *Toxoplasma* suggests secondary loss of *N*-glycan QC from coccidia. Positive selection for *N*-glycan sites occurs in secreted proteins of organisms with NG-QC and is based upon an increased likelihood of threonine but not serine in the second position versus asparagine. In contrast, there appears to be selection against *N*-glycan length in *Plasmodium* and *N*-glycan site density in *Toxoplasma*. Finally, there is suggestive evidence for *N*-glycan-dependent ERAD in *Trichomonas*, which glycosylates and degrades the exogenous reporter mutant carboxypeptidase Y (CPY*).

Keywords

endoplasmic reticulum; ER-associated degradation; evolution; *N*-glycan precursors; quality control of protein folding; parasites

© 2014 Elsevier Ltd. All rights reserved.

*Corresponding author at: Boston University Goldman School of Dental Medicine, 72 East Concord St, Evans 425, Boston MA, 02118, USA. Tel: 1 617 414 1054.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

1. Introduction

Parasitic protists cause malaria (*Plasmodium falciparum*), diarrhea (*Giardia lamblia* and *Cryptosporidium parvum*), dysentery (*Entamoeba histolytica*), neurological illness (*Trypanosoma brucei* and *Toxoplasma gondii*), heart disease (*Trypanosoma cruzi*), mucocutaneous lesions (*Leishmania sp.*), keratitis (*Acanthamoeba castellanii*), and sexual transmitted infections (*Trichomonas vaginalis*). Whole genome sequences of these medically important protists predict the “parts list” of proteins present in each organism, as well as those proteins that are not present [1–8]. The prediction of absence, which is important for analysis of secondary loss, is based upon the paucity of pseudogenes in these parasites, with the exception of *Trichomonas* [9]. Phylogenetic trees of small subunit rRNA genes or of housekeeping proteins allow us to identify protists with common ancestry, which correlates with their appearance (e.g. *Acanthamoeba* and *Entamoeba*) [10]. These trees also show that protists differ more from each other than do metazoans, fungi, and plants, but they fail to identify the oldest extant eukaryote. Instead they identify a set of deeply divergent eukaryotes including *Giardia*, *Trichomonas*, and *Trypanosoma*, which are called “excavates” [11]. In addition, subsequent to bottle-necking, the common ancestor of all eukaryotes has most certainly been lost [12].

Because of secondary loss of sets of genes (e.g. Alg genes that encode enzymes that make precursors to asparagine-linked glycans that are a focus here), it is possible to identify protists with a simpler, “ancestor-like” set of enzymes that make *N*-glycan precursors [13–15]. These protists with short or very short *N*-glycans show us that eukaryotes can grow and cause disease without *N*-glycan-dependent quality control of protein folding (*N*-glycan QC) or *N*-glycan-dependent ER-associated degradation (*N*-glycan ERAD) [16–18]. In turn, the protists without *N*-glycan QC demonstrate its impact on *N*-glycan site density in secreted proteins, by comparison to the vast majority of eukaryotes that have this system (see Section 4).

2. The present diversity of *N*-glycan precursors

2.1 Alg enzymes make *N*-glycan precursors

When we began this line of research ten years ago, we expected that protist *N*-glycans would resemble those of vast majority of metazoans, fungi, and plants that are composed of 14 sugars ($\text{Glc}_3\text{Man}_9\text{GlcNAc}_2$) (Fig. 1). The exception to this rule, of course, was *Trypanosoma cruzi*, which is missing the three glucose residues [19]. This absence was exploited by Armando Parodi, who characterized the UDP-glucose:glucosyltransferase (UGGT), the key enzyme in *N*-glycan QC, in the absence of confounding glucose residues on the *N*-glycan precursor [16]. In addition, there was a great deal of confusion as to whether *Plasmodium* is or is not able to make *N*-glycans [20].

The vast majority of Alg genes, which encode enzymes that make the *N*-glycan precursor that is transferred to Asn of proteins in the ER lumen, were discovered using a “mannose suicide” experiment [21]. Briefly, mutagenized yeast were labeled with tritiated mannose and then allowed to sit in a -70°C freezer for months. Upon thawing, those yeast knockouts that survived had retained less radioactive mannose, because synthesis of their *N*-glycan

precursors was blocked (e.g. Alg1 was unable to add the first mannose to dolichol-PP-GlcNAc₂). Alg genes, which are numbered based upon the order of their discovery, include those that make dolichol-PP-Man₅GlcNAc₂ (Fig. 1). These enzymes that use cytosolic UDP-GlcNAc and GDP-Man to add sugars to dolichol-phosphate are similar to those present in bacteria [14, 22]. In contrast, the enzymes that use dolichol-P-Man and dolichol-P-Glc to make dolichol-PP-Glc₃Man₉GlcNAc₂ in the lumen of the ER resemble enzymes that make the GPI-anchor precursor but do not resemble bacterial enzymes [23].

The extant eukaryotes appear to descend from a common ancestor that made Glc₃Man₉GlcNAc₂ (see argument below), and there are no alternative enzymes to catalyze the addition of a particular sugar to the *N*-glycan. A minor exception is the fusion of Alg13 and Alg14, which add the second GlcNAc to the *N*-glycan precursor, in *Entamoeba* and *Dictyostelium*. Further, there is no diversity in *N*-glycan precursors, as occurs for complex *N*-glycans made in the Golgi. Therefore, it is relatively easy to identify the Alg enzymes from the “parts list” of proteins of each parasite using BLASTP, as well as simple treeing methods to distinguish paralogs that result from gene duplication (e.g. Alg1, Alg2, and Alg11 that add the first five mannoses to the *N*-glycan precursor) [15].

2.2 Protists with short N-glycan precursors

Protists make all possible combinations of *N*-glycan precursors that are reasonable (e.g. it is not possible to add Glc to *N*-glycans that are missing a mannose arm) [15]. For example, coccidian parasites, which have oocyst walls and are spread by the fecal-oral route (*Toxoplasma* and *Cryptosporidium* that infect humans and *Eimeria* that infects chickens), are each missing enzymes that add four mannose to the *N*-glycan precursor (Alg3, Alg9, and Alg12). In some cases, they are also missing enzymes that add glucose (Alg6, Alg8, and Alg10), so that their *N*-glycan precursor then contains Glc_{0.3}Man₅GlcNAc₂ (Fig. 1).

Entamoeba and *Trichomonas*, which are unrelated to each other, are missing all the enzymes in the ER lumen that add mannose and glucose and so make a precursor that contains Man₅GlcNAc₂. These *N*-glycans were confirmed by metabolic labeling of parasites with tritiated mannose and analyzing *N*-glycans released with PNGaseF on a sizing column with yeast standards [15]. Alternatively, one can analyze released *N*-glycans or tryptic glycopeptides by mass spectrometry. Note that the parasite Man₅GlcNAc₂, which contains a single arm attached to 1,3-linked mannose that is the target of UGGT (Fig. 2), differs from mammalian or yeast Man₅GlcNAc₂ that results from ER mannosidase digestion of Man₉GlcNAc₂ [24]. Indeed the ER mannosidase converts the *Entamoeba* and *Trichomonas* *N*-glycans to biantennary Man₃ (Man₃GlcNAc₂), which is made by Golgi mannosidases in the human host. Interestingly, protists have the set of three cytosolic mannosyltransferases (Alg1, Alg2, and Alg11) that make Man₅GlcNAc₂ or have none at all [14,15]. In addition, Rft1, which has been genetically linked in yeast with transfer of Man₅GlcNAc₂ from the cytosol to the lumen of the ER but does not appear by itself to be the “flippase” for dolichol-PP-Man₅GlcNAc₂, shares the same phylogenetic profile as the cytosolic mannosyltransferases [15, 25–27].

2.3 Protists with very short N-glycan precursors

Plasmodium and *Giardia*, which are also unrelated to each other, are missing all of the Alg enzymes that add mannose or glucose (as well as Rft1) and so make a precursor that contains just GlcNAc₂ [15, 28, 29]. This result explains why tritiated mannose went into GPI-anchors but not N-glycans of *Plasmodium* [20]. Instead N-glycans of *Plasmodium* and *Giardia* are labeled with tritiated glucosamine, which is converted in the parasite to GlcNAc. Finally, *Theileria*, the cause of bovine malaria as well as lymphoma, has no N-glycans [30].

Because the coccidians (*Toxoplasma*, *Cryptococcus*, and *Eimeria*) and related apicomplexan parasites that infect red blood cells (RBCs) (*Plasmodium*, *Babesia*, and *Theileria*) share recent common ancestry, the present diversity of their N-glycans is based upon secondary loss of Alg genes from an ancestor that made Glc₃Man₅GlcNAc₂ [15, 29]. Selection against N-glycan length in parasites that infect RBCs is likely because secreted proteins targeted to the RBC cytosol must thread through a narrow channel called the PTEX [31]. *Acanthamoeba*, which shares common ancestry with *Entamoeba* that makes Man₅GlcNAc₂, has a complete set of Alg enzymes and so makes an N-glycan like that of the host (Glc₃Man₉GlcNAc₂) [4, 7, 15]. Secondary loss of N-glycans also occurs in fungi, if rarely. *Cryptococcus* is missing Alg enzymes that add glucose in the ER lumen and so makes an N-glycan precursor with Man₉GlcNAc₂ (like that of *Trypanosoma*), while *Encephalitozoon* (also known as microsporidium), a fungus with a markedly reduced genome, makes no N-glycans (Fig. 1) [15, 32].

2.4 Effects on the OST, complex N-glycans, and GPI-anchors

The oligosaccharyltransferase (OST) that moves the N-glycan from the precursor is composed of a catalytic peptide (STT3), which shares common ancestry with the bacterial enzyme, as well as multiple other peptides in metazoans, fungi, and plants [33]. The OST of *Giardia* and *Trypanosoma* is composed of a single catalytic peptide, while other protists have fewer non-catalytic peptides than higher eukaryotes. The OST of a particular eukaryote often prefers the endogenous N-glycan (e.g. Glc₃Man₉GlcNAc₂ for yeast and man or Man₅GlcNAc₂ for *Entamoeba* and *Trichomonas*) but this is not the case for *Trypanosoma* that transfer either Glc₃Man₉GlcNAc₂ or Man₉GlcNAc₂ [34]. While the *Giardia* OST transfers endogenous GlcNAc₂, it is also transfers with no particular preference exogenous N-glycans varying in size from Man₃GlcNAc₂ to Glc₃Man₉GlcNAc₂. Despite its short N-glycan, the occupied N-glycan sites of *Giardia* predominantly contain threonine (NxT) rather than serine (NxS) [28]. This result is consistent with previous observations that OSTs preferentially glycosylate NxT and that glycans are not part of the recognition site for STT3 [35–38]. An exceptional case is *Trypanosoma brucei*, in which a first STT3 transfers a truncated N-glycan precursor (Man₅GlcNAc₂) to N-glycan sites with an acidic amino acid in the -2 position, and then a second STT3 transfers the full-length N-glycan Man₉GlcNAc₂ to the canonical N-glycan site [39]. This transfer has important consequences, as *Trypanosoma* are missing the Golgi mannosidase and so cannot convert Man₉GlcNAc₂ to Man₃GlcNAc₂ (the building block for complex N-glycans).

Some parasites modify their *N*-glycans in the Golgi (e.g. *Trypanosoma* and *Trichomonas* add LacNAc, while *Entamoeba* adds Gal and Glc) [40, 41]. Many *N*-glycans, however, remain unmodified by ER mannosidases and Golgi glycosyltransferases and so are recognized by lectins that bind to the single 1,3-linked mannose arm of Man₅GlcNAc₂ (e.g. cyanovirin-*N* binding to *Entamoeba*, *Trichomonas*, *Toxoplasma*, and *Cryptosporidium*) [29, 42, 43]. Similarly, wheat germ agglutinin or *Griffonia simplicifolia* lectin II binds to unmodified GlcNAc₂ of *Giardia* and *Plasmodium* [28, 29]. Lectins that bind unmodified *N*-glycans are useful reagents for affinity purification of parasite glycoproteins from mass spectrometry. Lectins (e.g. cyanovirin-*N* and griffithsin), which bind high mannose *N*-glycans on gp120 of HIV and are candidate compounds to prevent heterosexual spread of HIV, also bind unmodified *N*-glycans on the surface of *Trichomonas* (our unpublished data) [44].

Secondary loss of enzymes (designated GPI in yeast or PIG in metazoans) that make GPI-anchors also occurs in protists but is independent of *N*-glycan status [45, 46]. For example, *Giardia* and *Entamoeba* are missing the mannosyltransferases that add the 3rd mannose to the GPI-precursor, while *Trichomonas* is missing the entire set of GPI-synthetic enzymes [2, 8, 47]. Interestingly, *Trichomonas* retains the Alg5 enzyme that makes Dol-P-Glc despite the absence of glucose in its *N*-glycan precursor. This result, which supports the idea that an ancestor of *Trichomonas* once glucosylated its *N*-glycan precursor, led us to the discovery that Dol-P-Glc is used to make *O*-linked glycans [48]. Finally, *Trypanosoma cruzi* and some *Leishmania* have a predicted glucosidase I that removes the terminal Glc on Glc₃Man₉GlcNAc₂. Since these parasites make an *N*-glycan precursor with Man₉GlcNAc₂, the presence of the glucosidase I gene is evidence for the secondary loss of these Alg6, Alg8, and Alg10 genes [14, 15]. The present function of *T. cruzi* glucosidase I is unclear, as no glucosidase I activity was detected in membrane extracts of the parasite [49].

In summary, secondary loss of Alg genes encoding enzymes that make the *N*-glycan precursor explains the present diversity of *N*-glycans among medically important protists. While *N*-glycan precursor length has profound effects on *N*-glycan QC, *N*-glycan ERAD, and *N*-glycan site density in secreted proteins (next sections), it appears to have no effect on the ability of these parasites to grow and cause a wide range of human illnesses. In contrast, decreased function or loss of function of the Alg enzymes that make *N*-glycan precursors have debilitating effects on human development, resulting in an array of phenotypes that are referred to as type 1 congenital disorders of glycosylation (CDG-I) [50].

3. Effect of *N*-glycan diversity on quality control of protein folding

3.1 *N*-glycan QC

Because the UGGT in the ER lumen adds glucose to a single arm attached to 1,3-linked mannose of the *N*-glycan bound to protein, protists that make *N*-glycan precursors with at least seven sugars (Man₅GlcNAc₂) (e.g. *Entamoeba*, *Trichomonas*, *Toxoplasma*, *Cryptosporidium*, *Trypanosoma*, and *Acanthamoeba*) are theoretically capable of *N*-glycan QC (Fig. 2) [15, 16, 18, 19]. Conversely, those organisms with very short *N*-glycan precursors (e.g. *Plasmodium* and *Giardia*) or no *N*-glycans (*Theileria* and *Encephalitozoon*) are incapable of *N*-glycan QC (or *N*-glycan ERAD, as discussed in Section 3) [17].

Consistent with these predictions, UGGT, glucosidase II that removes glucose from GlcMan₅GlcNAc₂ or GlcMan₉GlcNAc₂, and lectins that bind glucosylated *N*-glycans (calnexin and/or calreticulin) are present in *Entamoeba*, *Trichomonas*, *Trypanosoma*, and *Acanthamoeba*, while they are absent in *Plasmodium*, *Giardia*, and *Theileria*. It does not appear to matter whether organisms with *N*-glycan QC have calreticulin only (*Entamoeba* and *Trypanosoma*), calnexin only (*Saccharomyces*), or both lectins (*Trichomonas*) [18]. Organisms with predicted *N*-glycan QC also have the *N*-glycan-binding lectin that binds well-folded glycoproteins and directs them to the Golgi (ERGIC-53 and related VIP proteins), as well as a UDP-Glc transporter that moves the nucleotide sugar from the cytosol to the ER-lumen [51, 52]. The UDP-Glc transporter was first identified by expressing *Entamoeba* protein in *Giardia* that does not have NG-QC and does not transport UDP-Glc [53].

UGGT activity was demonstrated in *Entamoeba* and *Trichomonas*, which do not have Glc in their *N*-glycan precursors, by inhibiting glucosidase II with castanospermine and seeing a large increase in the amount of GlcMan₅GlcNAc₂ present in lysed parasites [16–18]. Alternatively, membrane extracts of *Entamoeba* were shown to use UDP-Glc to glucosylate denatured thyroglobulin or to add Glc to Man₅GlcNAc₂ attached to an iodinated NYT peptide. The latter result suggests that the *Entamoeba* UGGT is active even when attached to a short peptide, which is not the case for the *S. pombe* or *Trypanosoma* UGGT.

To our surprise, *Toxoplasma* and *Cryptosporidium*, which make *N*-glycan precursors with Glc₂₋₃Man₅GlcNAc₂, do not have UGGT, ERGIC-53, or intact calnexin and/or calreticulin (Fig. 2) [18]. While *Toxoplasma* has a calnexin that is missing the arm that binds the protein disulfide isomerase (PDI), it does not appear to have *N*-glycan QC, as suggested by the absence of positive selection for *N*-glycan sites in its secreted proteins (see Section 3) [54]. In contrast, positive selection for *N*-glycan sites is found in all other eukaryotes with *N*-glycan QC. This result suggests that the product of *Toxoplasma* glucosidases I and II (GlcMan₅GlcNAc₂) is not bound and refolded by the armless calnexin. In contrast, the product of *Saccharomyces* glucosidases I and II (GlcMan₉GlcNAc₂) appears to be bound and refolded by calnexin, as there is positive selection for *N*-glycan sites in its proteins despite the fact that the *Saccharomyces* UGGT ortholog (Kre5) does not glucosylate *N*-glycans [55].

3.2 *N*-glycan ERAD

Whether protists might have *N*-glycan ERAD is more difficult to judge from the parts list of their predicted proteins, because the set of effectors is not as well-defined as those for *N*-glycan QC (e.g. the dislocon is not molecularly identified) [17, 18, 56]. Further numerous components of *N*-glycan ERAD (e.g. BIP, PDI, peptidylprolyl isomerase (PPI), HRD1, Der1, and CDC48) are not specific to this pathway and are present in organisms such as *Giardia* and *Plasmodium* that do not have *N*-glycan ERAD. In turn, *N*-glycan ERAD activity is judged by overexpression of an exogenous misfolded protein, which may be secreted or membrane bound, rather than by inhibition of glucosidase II or by a direct assay of membrane extracts, as for *N*-glycan QC. With these caveats, *Trichomonas* and *Trypanosoma*, which have *N*-glycan QC, have orthologs of Mns1 and EDEM, respectively,

and each has mannosidase activity as a recombinant protein (Fig. 2) [18]. *Trichomonas* also has a cytosolic PNGaseF, which is active as a recombinant protein. *Trichomonas* adds *N*-glycans to wild-type carboxypeptidase Y (CPY) of *Saccharomyces*, as well as to mutant CPY*, which misfolds and is degraded in a proteasome-dependent manner (our unpublished data). *Trichomonas* has a distant homolog to OS-9, the lectin that binds 1,6-linked mannose. The 1,6-linked mannose (marked by an asterisk in Fig. 2) is revealed by host ER mannosidases prior to dislocation of the misfolded protein into the cytosol. The unprocessed *N*-glycan of *Trichomonas* (Man₅GlcNAc₂) contains the 1,2-linked mannose recognized by UGGT, as well as the 1,6-linked mannose recognized by OS-9, suggesting a possible “tug of war” between *N*-glycan QC and *N*-glycan ERAD.

The safest conclusion then is that some but not all protists have *N*-glycan QC, while the evidence for *N*-glycan ERAD is unclear. In contrast, all protists have proteins involved in *N*-glycan-independent QC of proteins folding (BIP and other chaperones, PDI, and PPI), although most parasites are missing the transmembrane kinase/RNase (IRE1) that triggers the unfolded protein response in higher eukaryotes [57]. All protists also have proteins involved in *N*-glycan-independent ERAD, although studies with misfolded reporter proteins lacking *N*-glycan sites have not been performed [18]. Remarkably, *Toxoplasma* has a second set of ERAD proteins (Der1 and Cdc48) that appear to be essential for transporting nuclear-encoded proteins from the ER lumen to a chloroplast-derived organelle called the apicoplast [58].

4. Effect of *N*-glycan QC on protein evolution

4.1 Hypotheses tested

Addition of *N*-glycans may affect protein folding in four ways. First, addition of *N*-glycans affects the thermodynamic stability of the protein [59]. Second, by association with the Sec61 translocon, the OST affects protein folding when it binds the nascent peptide and adds *N*-glycans [14]. Third, *N*-glycans are the target for UGGT, calnexin/calreticulin, and ERGIC53 [16]. Fourth, they are part of *N*-glycan ERAD [17, 56]. To judge the relative importance of these processes, we determined whether there is a positive selection for *N*-glycan site density (NxT or NxS) in secreted proteins versus cytosolic proteins, where no *N*-glycans are added. We also compared *N*-glycans site densities in the secreted proteins of the vast majority of eukaryotes that have *N*-glycan QC with those of protists (*Plasmodium*, *Giardia*, *Toxoplasma*, *Cryptosporidium*, and *Theileria*) that do not have *N*-glycan QC. To do this experiment, we collected protein sequences from as many dissimilar organisms as were available in 2009, removed redundant sequences that fill the NR database at the NCBI, and separated the remaining unique sequences into secreted proteins (with an *N*-terminal signal peptide) and nucleocytosolic proteins with no signal peptide or transmembrane helices [60, 61]. We then compared the expected density of *N*-glycan sites based upon the amino acid composition of the proteins of each organism with the actual *N*-glycan site density (Fig. 3).

4.2 Selection for *N*-glycan sites with Thr but not Ser

There were three prominent results (Fig. 3). First, independent of whether organisms have long or short *N*-glycans and whether they have or do not have *N*-glycan QC, those with AT-

rich genomes (e.g. *Plasmodium*, *Dictyostelium*, and *Entamoeba*) have a greater density of expected *N*-glycan sites, as asparagine is encoded by AT-rich codons (AAT/C). Conversely, *Toxoplasma* with a GC-rich genome has fewer expected *N*-glycan sites. Humans are for the most part AT-GC neutral and have an intermediate density of expected *N*-glycan sites, while *Saccharomyces* is slightly AT-rich and has more expected *N*-glycan sites. To our knowledge, this is the only effect of codon usage on protein glycosylation, as Ser and Thr that are *O*-glycosylated are encoded by a balance of AT- and GC-rich codons [18, 62].

Second, in organisms with *N*-glycan QC the actual density of *N*-glycans sites with Thr (NxT) in their secreted proteins is markedly increased versus the expected density of sites, and this increase is independent of AT-content. In contrast and to our surprise, there was no increase in density of *N*-glycan sites with Ser (NxS) in organisms with *N*-glycan QC. For these same organisms, the expected and actual *N*-glycans site densities with Thr or Ser were the same in nucleocytoplasmic proteins (negative controls), which are not *N*-glycosylated. The expected *N*-glycans site densities were the same for secreted and nucleocytoplasmic proteins of organisms with *N*-glycan QC, showing that increase in the actual *N*-glycan site density is not because of an increase in the density of Asn or Thr. Instead the increase in the actual density of *N*-glycan sites with Thr (NxT) is based upon an increased likelihood that Thr is in the +2 position with regards to Asn. The increase in the *N*-glycan site density in hemagglutinin of influenza virus as it mutates over a 20 year period in the human host is also based upon an increased likelihood that Thr or Ser is in the +2 position with regards to Asn rather than an increase in the density of Asn, Thr, or Ser [18, 63]. Similarly, the very high density of *N*-glycan sites on gp120 of the HIV is predominantly explained by an increased likelihood that Thr or Ser is in the +2 position with regards to Asn.

Third, in organisms without *N*-glycan QC, there is no difference between the actual *N*-glycan site density with Thr or Ser and the expected *N*-glycan site density for its secreted proteins. This result shows that *N*-glycan QC, not just the presence of *N*-glycans that are as many as 9 to 10 sugars long in *Cryptosporidium* and *Toxoplasma*, is the important determinant for positive selection for *N*-glycan sites in secreted proteins.

4.3 Possible explanations for these surprising results

Despite the preference of OSTs for *N*-glycan sites with Thr over those with Ser, ~1/3 of occupied *N*-glycan sites contains Ser [37]. There is then no easy explanation for why there is positive selection for *N*-glycan sites with Thr but not Ser in the secreted proteins of organisms with *N*-glycan QC. There must be some component of the *N*-glycan QC system that is sensitive to the site of *N*-glycans and not just to the status of the *N*-glycan or of glycoprotein folding. Regardless of the mechanism, these evolutionary studies of *N*-glycan site density in secreted proteins are a strong argument for the importance of *N*-glycan QC (and possibly *N*-glycan ERAD that is present in the vast majority of these organisms). Such evidence has been relatively hard to come by experimentally, where knockout of UGGT gives a phenotype only when yeast or trypanosomes are severely stressed [64, 65]. Finally, as an exception that perhaps proves the rule, there is negative selection against the sites of *N*-glycosylation in *Toxoplasma* proteins that pass through the ER and then thread through a pore into the apicoplast, so that many of these proteins have no *N*-glycan sites [29].

Conclusions and future studies

Studies of medically important protists has shown the importance of secondary loss of Alg enzymes to explain the evolution of *N*-glycans and to show that *N*-glycan length, *N*-glycan QC, and *N*-glycan ERAD are not as important for viability of single cell organisms as for higher eukaryotes [50]. These studies have defined what appears to be the minimum proteins involved in *N*-glycan QC and have shown the mechanism whereby *N*-glycan QC selects for *N*-glycan site density in secreted proteins. Unanswered questions revolve primarily around how *N*-glycan precursors are flipped into the ER lumen, whether there is *N*-glycan ERAD in some protists, and why there is selection for *N*-glycan sites with Thr but not Ser in secreted proteins of eukaryotes with *N*-glycan QC.

Acknowledgments

We thank a terrific set of students and post-doctoral fellows, who performed this work. In alphabetical order, they are Giulia Bandini, Sulagna Banerjee, Guy Bushkin, Andrea Carpentieri, Aparajita Chatterjee, John Cipollo, Jake Cui, Kariona Grabska, John Haserick, Paula Magnelli, Edwin Motari, and Dan Ratner. We thank collaborators including Cathy Costello, Reid Gilmore, Marc-Jan Gubbels, Carlos Hirschberg, Barry O'Keefe, and Temple Smith. This work was supported by the NIH grants GM031318, AI44070, and AI048082.

Abbreviations

CPY	<i>Saccharomyces</i> carboxypeptidase Y
CPY*	misfolded CPY mutant
N-glycan	Asn-linked glycan
PPI	peptidylprolyl isomerase
PDI	protein disulfide isomerase
QC	quality control
UGGT	UDP-glucose:glucosyltransferase

References

1. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, et al. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*. 2002; 419:498–511. [PubMed: 12368864]
2. Morrison HG, McArthur AG, Gillin FD, Aley SB, Adam RD, Olsen GJ, et al. Genomic minimalism in the early diverging intestinal parasite *Giardia lamblia*. *Science*. 2007; 317:1921–26. [PubMed: 17901334]
3. Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G, Lancto CA, et al. Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. *Science*. 2004; 304:441–5. [PubMed: 15044751]
4. Clark CG, Alsmark UC, Tazreiter M, Saito-Nakano Y, Ali V, Marion S, Weber C, et al. Structure and content of the *Entamoeba histolytica* genome. *Adv Parasitol*. 2007; 65:51–190. [PubMed: 18063096]
5. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, et al. The genome of the African trypanosome *Trypanosoma brucei*. *Science*. 2005; 309:416–22. [PubMed: 16020726]

6. El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Tran AN, et al. The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. *Science*. 2005; 309:409–15. [PubMed: 16020725]
7. Clarke M, Lohan AJ, Liu B, Lagkouvardos I, Roy S, Zafar N, et al. Genome of *Acanthamoeba castellanii* highlights extensive lateral gene transfer and early evolution of tyrosine kinase signaling. *Genome Biol*. 2013; 14:R11. [PubMed: 23375108]
8. Carlton JM, Hirt RP, Silva JC, Delcher AL, Schatz M, et al. Draft genome sequence of the sexually transmitted pathogen *Trichomonas vaginalis*. *Science*. 2007; 315:207–12. [PubMed: 17218520]
9. Cui J, Smith T, Samuelson J. The large family of *Trichomonas* genes encoding transmembrane adenyl cyclases results from massive gene duplication and concomitant development of pseudogenes. *PLoS Negl Trop Dis*. 2010; 4:e782. [PubMed: 20689771]
10. Sogin ML, Silberman JD. Evolution of the protists and protistan parasites from the perspective of molecular systematics. *Int J Parasitol*. 1998; 28:11–20. [PubMed: 9504331]
11. Hampl V, Hug L, Leigh JW, Dacks JB, Lang BF, Simpson AG, Roger AJ. Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic “supergroups”. *Proc Natl Acad Sci USA*. 2009; 106:3859–64. [PubMed: 19237557]
12. Vishwanath P, Favaretto P, Hartman H, Mohr SC, Smith TF. Ribosomal protein-sequence block structure suggests complex prokaryotic evolution with implications for the origin of eukaryotes. *Mol Phylogenet Evol*. 2004; 33:615–25. [PubMed: 15522791]
13. Aebl M. N-linked protein glycosylation in the ER. *Biochim Biophys Acta*. 2013; 1833:2430–7. [PubMed: 23583305]
14. Schwarz F, Aebl M. Mechanisms and principles of N-linked protein glycosylation. *Curr Opin Struct Biol*. 2011; 21:576–82. [PubMed: 21978957]
15. Samuelson J, Banerjee S, Magnelli P, Cui J, Kelleher DJ, Gilmore R, et al. The diversity of dolichol-linked precursors to Asn-linked glycans likely results from secondary loss of sets of glycosyltransferases. *Proc Natl Acad Sci USA*. 2005; 102:1548–53. [PubMed: 15665075]
16. D'Alessio C, Caramelo JJ, Parodi AJ. UDP-Glc:glycoprotein glucosyltransferase-glucosidase II, the ying-yang of the ER quality control. *Semin Cell Dev Biol*. 2010; 21:491–9. [PubMed: 20045480]
17. Merulla J, Fasana E, Soldà T, Molinari M. Specificity and regulation of the endoplasmic reticulum-associated degradation machinery. *Traffic*. 2013; 14:767–77. [PubMed: 23521725]
18. Banerjee S, Vishwanath P, Cui J, Kelleher DJ, Gilmore R, Robbins PW, et al. The evolution of N-glycan-dependent endoplasmic reticulum quality control factors for glycoprotein folding and degradation. *Proc Natl Acad Sci USA*. 2007; 104:11676–81. [PubMed: 17606910]
19. Parodi AJ. N-glycosylation in trypanosomatid protozoa. *Glycobiology*. 1993; 3:193–9. [PubMed: 8358146]
20. Berhe S, Gerold P, Kedees MH, Holder AA, Schwarz RT. *Plasmodium falciparum*: merozoite surface proteins 1 and 2 are not posttranslationally modified by classical N- or O-glycans. *Exp Parasitol*. 2000; 94:194–7. [PubMed: 10831385]
21. Huffaker T, Robbins PW. Yeast mutants deficient in protein glycosylation. *Proc Natl Acad Sci USA*. 1983; 80:7466–7470. [PubMed: 6369318]
22. Wacker M, Linton D, Hitchen PG, Nita-Lazar M, Haslam SM, North SJ, et al. N-linked glycosylation in *Campylobacter jejuni* and its functional transfer into *E. coli*. *Science*. 2002; 298:1790–3. [PubMed: 12459590]
23. Oriol R, Martinez-Duncker I, Chantret I, Mollicone R, Codogno P. Common origin and evolution of glycosyltransferases using Dol-P-monosaccharides as donor substrate. *Mol Biol Evol*. 2002; 19:1451–63. [PubMed: 12200473]
24. Herscovics A. Importance of glycosidases in mammalian glycoprotein biosynthesis. *Biochim Biophys Acta*. 1999; 1473:96–107. [PubMed: 10580131]
25. Helenius J, Ng DT, Marolda CL, Walter P, Valvano MA, Aebl M. Translocation of lipid-linked oligosaccharides across the ER membrane requires Rft1 protein. *Nature*. 2002; 415:447–50. [PubMed: 11807558]
26. Frank CG, Sanyal S, Rush JS, Waechter CJ, Menon AK. Does Rft1 flip an N-glycan lipid precursor? *Nature*. 2008; 454:E3–5. [PubMed: 18668045]

27. Jelk J, Gao N, Serricchio M, Signorell A, Schmidt RS, Bangs JD, et al. Glycoprotein biosynthesis in a eukaryote lacking the membrane protein Rft1. *J Biol Chem.* 2013; 288:20616–23. [PubMed: 23720757]
28. Ratner DM, Cui J, Steffen M, Moore LL, Robbins PW, Samuelson J. Changes in the N-glycome, glycoproteins with Asn-linked glycans, of *Giardia lamblia* with differentiation from trophozoites to cysts. *Eukaryot Cell.* 2008; 7:1930–40. [PubMed: 18820077]
29. Bushkin GG, Ratner DM, Cui J, Banerjee S, Duraisingh MT, Jennings CV, et al. Suggestive evidence for Darwinian selection against asparagine-linked glycans of *Plasmodium falciparum* and *Toxoplasma gondii*. *Eukaryot Cell.* 2010; 9:228–41. [PubMed: 19783771]
30. Gardner MJ, Bishop R, Shah T, de Villiers EP, Carlton JM, Hall N, et al. Genome sequence of *Theileria parva*, a bovine pathogen that transforms lymphocytes. *Science.* 2005; 309:134–7. [PubMed: 15994558]
31. Elsworth B, Matthews K, Nie CQ, Kalanon M, Charnaud SC, Sanders PR, et al. PTEX is an essential nexus for protein export in malaria parasites. *Nature.* 2014; 511:587–91. [PubMed: 25043043]
32. Katinka MD, Duprat S, Cornillot E, Méténier G, Thomarat F, Prensier G, et al. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature.* 2001; 414:450–3. [PubMed: 11719806]
33. Kelleher DJ, Gilmore R. An evolving view of the eukaryotic oligosaccharyltransferase. *Glycobiology.* 2006; 16:47R–62R.
34. Kelleher DJ, Banerjee S, Cura AJ, Samuelson J, Gilmore R. Dolichol-linked oligosaccharide selection by the oligosaccharyltransferase in protist and fungal organisms. *J Cell Biol.* 2007; 177:29–37. [PubMed: 17403929]
35. Marshall RD. Glycoproteins. *Annu Rev Biochem.* 1972; 41:673–702. [PubMed: 4563441]
36. Breuer W, Klein RA, Hardt B, Bartoschek A, Bause E. Oligosaccharyltransferase is highly specific for the hydroxy amino acid in Asn-Xaa-Thr/Ser. *FEBS Lett.* 2001; 501:106–10. [PubMed: 11470266]
37. Zielinska DF, Gnad F, Schropp K, Wi niewski JR, Mann M. Mapping N-glycosylation sites across seven evolutionarily distant species reveals a divergent substrate proteome despite a common core machinery. *Mol Cell.* 2012 May 25; 46(4):542–8. [PubMed: 22633491]
38. Lizak C, Gerber S, Numao S, Aebi M, Locher KP. X-ray structure of a bacterial oligosaccharyltransferase. *Nature.* 2011 Jun 15; 474(7351):350–5. [PubMed: 21677752]
39. Izquierdo L, Schulz BL, Rodrigues JA, Güther ML, Procter JB, Barton GJ, Aebi M, Ferguson MA. Distinct donor and acceptor specificities of *Trypanosoma brucei* oligosaccharyltransferases. *EMBO J.* 2009; 8:2650–61. [PubMed: 19629045]
40. Atrih A, Richardson JM, Prescott AR, Ferguson MA. *Trypanosoma brucei* glycoproteins contain novel giant poly-N-acetyllactosamine carbohydrate chains. *J Biol Chem.* 2005; 280:865–71. [PubMed: 15509560]
41. Magnelli P, Cipollo JF, Ratner DM, Cui J, Kelleher D, Gilmore R, Costello CE, Robbins PW, Samuelson J. Unique Asn-linked oligosaccharides of the human pathogen *Entamoeba histolytica*. *J. Biol. Chem.* 2008; 283:18355–18364.
42. Carpentieri A, Ratner DM, Ghosh SK, Banerjee S, Bushkin GG, Cui J, Lubrano M, Steffen M, Costello CE, O'Keefe B, Robbins PW, Samuelson J. The antiretroviral lectin cyanovirin-N targets well-known and novel targets on the surface of *Entamoeba histolytica* trophozoites. *Eukaryot Cell.* 2010; 9:1661–1668. [PubMed: 20852023]
43. Adams EW, Ratner DM, Bokesch HR, McMahon JB, O'Keefe BR, et al. Oligosaccharide and glycoprotein microarrays as tools in HIV glycobiology; glycan-dependent gp120/protein interactions. *Chem Biol.* 2004; 11:875–81. [PubMed: 15217620]
44. Kouokam JC, Huskens D, Schols D, Johannemann A, Riedell SK, Walter W, et al. Investigation of griffithsin's interactions with human cells confirms its outstanding safety and efficacy profile as a microbicide candidate. *PLoS One.* 2011; 6:e22635. [PubMed: 21829638]
45. Ferguson MA. The structure, biosynthesis and functions of glycosylphosphatidylinositol anchors, and the contributions of trypanosome research. *J Cell Sci.* 1999; 112:2799–809. [PubMed: 10444375]

46. Orlean P, Menon AK. Thematic review series: lipid posttranslational modifications. GPI anchoring of protein in yeast and mammalian cells, or: how we learned to stop worrying and love glycopospholipids. *J Lipid Res.* 2007; 48:993–1011. [PubMed: 17361015]
47. Moody-Haupt S, Patterson JH, Mirelman D, McConville MJ. The major surface antigens of *Entamoeba histolytica* trophozoites are GPI-anchored proteophosphoglycans. *J Mol Biol.* 2000; 297:409–20. [PubMed: 10715210]
48. Grabiska KA, Ghosh SK, Guan Z, Cui J, Raetz CR, Robbins PW, Samuelson J. Dolichyl-phosphate-glucose is used to make O-glycans on glycoproteins of *Trichomonas vaginalis*. *Eukaryot Cell.* 2008; 7:1344–51. [PubMed: 18552282]
49. Bosch M, Trombetta S, Engstrom U, Parodi AJ. Characterization of dolichol diphosphate oligosaccharide: protein oligosaccharyltransferase and glycoprotein-processing glucosidases occurring in trypanosomatid protozoa. *J Biol Chem.* 1988; 263:17360–5. [PubMed: 3053710]
50. Freeze HH, Chong JX, Bamshad MJ, Ng BG. Solving glycosylation disorders: fundamental approaches reveal complicated pathways. *Am J Hum Genet.* 2014; 94:161–75. [PubMed: 24507773]
51. Schrag JD, Procopio DO, Cygler M, Thomas DY, Bergeron JJ. Lectin control of protein folding and sorting in the secretory pathway. *Trends Biochem Sci.* 2003; 28:49–57. [PubMed: 12517452]
52. Caffaro CE, Hirschberg CB. Nucleotide sugar transporters of the Golgi apparatus: from basic science to diseases. *Acc Chem Res.* 2006; 39:805–12. [PubMed: 17115720]
53. Banerjee S, Cui J, Robbins PW, Samuelson J. Use of *Giardia*, which appears to have a single nucleotide-sugar transporter for UDP-GlcNAc, to identify the UDP-Glc transporter of *Entamoeba*. *Mol Biochem Parasitol.* 2008; 159:44–53. [PubMed: 18346800]
54. Cui J, Smith T, Robbins PW, Samuelson J. Darwinian selection for sites of Asn-linked glycosylation in phylogenetically disparate eukaryotes and viruses. *Proc Natl Acad Sci USA.* 2009; 106:13421–13426. [PubMed: 19666543]
55. Fernández FS, Trombetta SE, Hellman U, Parodi AJ. Purification to homogeneity of UDP-glucose:glycoprotein glucosyltransferase from *Schizosaccharomyces pombe* and apparent absence of the enzyme from *Saccharomyces cerevisiae*. *J Biol Chem.* 1994; 269:30701–6. [PubMed: 7982990]
56. Hebert DN, Molinari M. Flagging and docking: dual roles for N-glycans in protein quality control and cellular proteostasis. *Trends Biochem Sci.* 2012; 37:404–10. [PubMed: 22921611]
57. Chen Y, Brandizzi F. IRE1: ER stress sensor and cell fate executor. *Trends Cell Biol.* 2013; 23:547–55. [PubMed: 23880584]
58. Agrawal S, van Dooren GG, Beatty WL, Striepen B. Genetic evidence that an endosymbiont-derived endoplasmic reticulum-associated protein degradation (ERAD) system functions in import of apicoplast proteins. *J Biol Chem.* 2009; 284:33683–91. [PubMed: 19808683]
59. Shental-Bechor D, Levy Y. Effect of glycosylation on protein folding: a close look at thermodynamic stabilization. *Proc Natl Acad Sci USA.* 2008; 105:8256–61. [PubMed: 18550810]
60. Bendtsen JD, Nielsen H, von Heijne G, Brunak S. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol.* 2004; 340:783–95. [PubMed: 15223320]
61. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J Mol Biol.* 2001; 305:567–80. [PubMed: 11152613]
62. Singer GA, Hickey DA. Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. *Mol Biol Evol.* 2000; 17:1581–8. [PubMed: 11070046]
63. Zhang M, Gaschen B, Blay W, Foley B, Haigwood N, Kuiken C, et al. Tracking global patterns of N-linked glycosylation site variation in highly variable viral glycoproteins: HIV, SIV, and HCV envelopes and influenza hemagglutinin. *Glycobiology.* 2004; 14:1229–46. [PubMed: 15175256]
64. Fanchiotti S, Fernández F, D'Alessio C, Parodi AJ. The UDP-Glc:Glycoprotein glucosyltransferase is essential for *Schizosaccharomyces pombe* viability under conditions of extreme endoplasmic reticulum stress. *J Cell Biol.* 1998; 143:625–35. [PubMed: 9813085]
65. Izquierdo L, Atrih A, Rodrigues JA, Jones DC, Ferguson MA. *Trypanosoma brucei* UDP-glucose:glycoprotein glucosyltransferase has unusual substrate specificity and protects the parasite from stress. *Eukaryot Cell.* 2009; 8:230–40. [PubMed: 19114500]

Fig. 1.

Alg enzymes predict *N*-glycan precursors. Secondary loss of Alg enzymes from apicomplexan parasites (top) and from fungi (bottom) predicts the present diversity of *N*-glycan precursors. *Toxoplasma* has all the Alg enzymes except those that add mannose in the ER lumen and so makes a precursor with $\text{Glc}_3\text{Man}_5\text{GlcNAc}_2$. Boxes outline those Alg enzymes and the *N*-glycan precursors of the other parasites. Similarly, *Saccharomyces* has a complete set of Alg enzymes and makes $\text{Glc}_3\text{Man}_9\text{GlcNAc}_2$, while *Cryptococcus* is missing those that add glucose. *Encephalitozoon*, like *Theileria* has no Alg enzymes and so makes no *N*-glycans. Drawn after ref. [29].

Fig. 2. Predicted *N*-glycan QC and *N*-glycan ERAD in higher eukaryotes (left) and *Trichomonas* (right). Enzymes and substrates involved in *N*-glycan QC are marked in red, while those involved in *N*-glycan ERAD are marked in blue). *N*-glycan sugars are as in Fig. 1. Abbreviations for this figure only are glucosidases I and II (GlsI and GlcII), calreticulin (CRT), and calnexin (CNX). Asterisks mark 1,6-linked mannose recognized by the OS-9 lectin. *N*-glycan QC and *N*-glycan ERAD are absent in *Giardia* and *Plasmodium*. Drawn after ref. [18].

Fig. 3.

Positive selection for *N*-glycan sites in secreted proteins of eukaryotes with *N*-glycan QC. Densities of *N*-glycan sites with Thr (per 500 amino acids) in secreted proteins of eukaryotes with *N*-glycan QC (top) and those without *N*-glycan QC (bottom) are plotted versus the AT content of the genome. Note that for each organism there are two spots marked: the *N*-glycan density expected by amino acid composition of proteins (green) and the actual *N*-glycan density (magenta). For selected organisms, a vertical bar connects the two spots: *Homo* (Hs), *Saccharomyces* (Sc), *Dictyostelium* (Dd), *Entamoeba* (Eh), *Plasmodium* (Pf), and *Toxoplasma* (Tg). In both plots, the expected *N*-glycan density increases with AT content, as Asn is encoded by AAT/C. In organisms with *N*-glycan QC there is a difference between the actual density of *N*-glycan sites with Thr and the expected difference (positive selection). There is no difference between actual and expected *N*-glycan site density for organisms without *N*-glycan QC (no selection). Drawn after ref. [54].