An in vitro one-pot synthetic biology approach to simulating diverging Golgi O-glycosylation of tumor-associated MUC1 from normal tissue MUC1

Abdullateef Nashed[1,2], Kyllen Dilsook[1,2], Tharindu Senapathi[1,3], and Kevin J. Naidoo[1,2*].

[1]Scientific Computing Research Unit Address, University of Cape Town, Rondebosch 7701, [2]Department of Chemistry, University of Cape Town, Rondebosch 7701. [3]Department of Chemistry, Faculty of Applied Sciences, University of Sri Jayewardenepura, Nugegoda 10250, Sri Lanka.

*To whom correspondence should be addressed

KEYWORDS: In vitro synthetic biology, Cancer, MUC1 antigens, systems chemical glycobiology, reaction dynamics.

ABSTRACT: Peptide O-glycosylation is a non-template-driven process that relies on the coordinated action of glycosyltransferases (GTs) within the endoplasmic reticulum (ER) and Golgi apparatus. An in vitro one-pot synthetic biology approach was developed to investigate the specificity and kinetics of GT O-GalNAc glycosylation that leads to tumor antigen glycoforms of mucin 1 (MUC1). The focus is to experimentally simulate the divergent glycosylation pathways that lead to the synthesis of cancer-associated antigens (Tn, T) and their sialylated derivatives. First, the biosynthetic details of the defining first step of GALNT re-localization from the ER to the Golgi was modeled using the one-pot method. Our findings reveal that an ER enriched with GALNTs results in complete Galnac (Tn) MUC1 site occupancy. This comes about as a function of two processes that are i) extended GALNT reaction time and ii) prevention of inhibition by subsequent glycosylation enzymes like C1GALT1. The modeling confirms that B3GNT6 has negligible specificity for MUC1 Tn, explaining the absence of core 3 and core 4 structures in MUC1 in both normal and cancerous breast cell lines. Moreover, ST6GALNAC1, and not ST6GALNAC2, is primarily responsible for α-2-6 sialylation of Tn and T antigens. Computer reaction dynamic simulations combined with kinetic experimental analysis show that ST6GALNAC1 prefers fully glycosylated MUC1 but moreover that its preference is the sialyation the S9 and T13 sites in the SAPDTR motif. This is especially the case when MUC1 concentration is great (i.e., highly expressed), suggesting that sTn upregulation on MUC1 in cancer is linked to the occupancy status of S9 and T13 glycosylated sites, that were previously found to be cancer-associated. The results from the one-pot synthesis approach presented here demonstrate its ability to simulate cellular glycosylation within the Golgi-ER. This systems modelling unpacks the molecular details of enzyme localization and substrate glycan occupancy that is fundamental to the regulatory mechanisms that gives rise to tumor-associated MUC1 antigens.

1. INTRODUCTION

Predicting the concentration of glyconjugates or the extent to which and how they are glycosylated is not possible with current experimental and computational tools. Simply put, the post-translational event of peptide or protein glycosylation is a non-template-driven process that relies on more than just glycoenzyme gene expression data or even the glycoenzyme expression levels themselves.1 A case in point is that while glycosyltransferase gene expression can be used to classify cancer,2 this genomic level data cannot be used to directly infer the difference between cancerous and healthy glycoconjugate expression. Specifically, the characteristically high degree of sialylation observed in tumor tissues3 and the associated structural modifications of glycans cannot be directly correlated with the genes that express the sialyltransferases. This is partly because the complex glycosylation pathways of a cell are intimately connected and intertwined with other critical metabolic and regulatory networks within the cell.4 Consequently, developing a systems biology model of glycoconjugate metabolic networks requires multiple components. However, a critical first step is constructing a developmental model that mimics the biosynthesis processes within the ER and Golgi apparatus. Using MUC1 as an example, a systems model must explain the preferential construction of a normal glycosylated state over a tumor-associated (TA) MUC1 glycosylated state while assigning key drivers to the MUC1 and TA-MUC1 glycopeptide outcomes.

Site specific chemoenzymatic synthesis is the standard method for producing model glycopeptides, where the initial sugar moiety of the core peptide is chemically bonded to a targeted residue. This is followed by enzymatic synthesis of the glycan, one sugar moiety at a time.5 Advances in understanding enzyme specificities and mechanisms have enabled the synthesis of more complex glycopeptides,5 expanding the dimensions of glycopeptide arrays. For example, Yoshimura et al.6 produced an array of 20 MUC1 glycopeptides that encompassed the Tn and T antigens and their sialylated forms, STn and ST, respectively. These glycans were synthesized at each of the five possible glycosylation sites of the MUC1 tandem repeat. The initial glycosylation was done by a chemical addition of the GalNAc residue at the selected site, followed by enzymatic glycosylation to complete the glycan structures. Good yields for all the single-site glycosylations were achieved in this way. Alternatively, a synthetic biology approach using genetically engineered human embryonic kidney (HEK) cells has been developed.7 This was achieved by rationally modifying the endogenous glycosylation pathways through knock-in or knock-out of specific GTs. The result are cells that can synthesize specific glycan structures that can either be displayed on the cell surface or on a probe protein designed for secretion. The advantage of the synthetic biology approach over in vitro methods is that a range of glycoconjugate structures biologically possible can be produced from a disease-specific genotype and as a consequence the engineered glycosylation pathway can be inferred from this. Furthermore, the glycoconjugate binding and interactions can be studied in the context of their cell-displayed (and cellularly generated) form. However, shortcomings of this approach include: i) the produced glycoconjugates rely on the endogenous machinery of the cells and cannot be completely controlled and optimized to avoid side reactions and incomplete glycosylation; ii) the pathway modifications are limited to the minimal cell survival needs of the selected glycosylation pathways. These drawbacks result in a heterogeneity of expressed structures, that obscure the synthesis of only one unique glycosylated structure necessary for epitope functional studies. In vitro methods that produce single, purifiable, and spectroscopically verifiable structures are more suitable than synthetic biology methods since the intention is to measure the kinetics of GTs as well as map out their selectivity and mechanistic action. The localization of GTs, such as GALNTs, has been found to be a regulatory mechanism involved in cancer phenotypes by altering O-GalNAc glycan structures and levels.8, 9 Consequently, in vitro methods must have the capability to mimic the alterations of in vivo glycosylation resulting from the spatial-temporal rearrangement of the distribution and presentation of GTs to the substrate in the ER-Golgi system.10, 11

Mucin 1 (MUC1), a highly glycosylated transmembrane protein overexpressed in epithelial cancers, contributes to tumorigenesis, immune evasion, metastasis, and ultimately poor prognosis. Its extracellular O-glycosylation weakens drug sensitivity by acting as a barrier and modulates signaling pathways, leading to decreased drug permeation and increased cancer cell survival. MUC1 O-glycosylation has been proposed as a target to enhance drug sensitivity and efficacy.12 Here, we present an in vitro method demonstrated on the MUC1 glycosylation process that simulates the ER-Golgi spatial and temporal regulation of glycosyltransferases (GTs). Specifically, we illustrate: i) the kinetic parameters governing GT activities and substrate specificities, ii) the molecular mechanisms govern the GTs site specificities, and iii) the effect of their expression and distribution along the ER-Golgi axis. Essential to this method is a tool that can accurately and consistently measure the kinetics across all GTs involved in biosynthesis. Previously, we reported the UGC assay13 as such a universal tool. We show that the multiple glycosylations needed to produce normal and tumor associated MUC1 glycoconjugate forms are achieved through an in vitro one-pot synthetic biology approach using a peptide fusion protein expression system. Contrived ER-Golgi conditions are created to model the O-GalNAc glycosylation of the MUC1 peptide, producing the cancer-associated antigens Tn, T, and their respective sialylated forms, sTn and sT. Following this the experimental model along with advanced computer reaction dynamics simulations, were used to assess the specificity of different GTs involved in the synthesis of these antigens at the five unique MUC1 tandem repeat glycosylation sites.

2. RESULTS AND DISCUSSION

The first step in mucin-type O-linked glycosylation is the addition of GalNAc to serine or threonine residues facilitated by several N-acetylgalactosaminyltransferases (GALNTs), forming the Thomsen-nouvelle (Tn) antigen (Figure 1 A). Following this the addition of galactose to the Tn antigen through T synthase (C1GALT1) is modified to form the T antigen (core 1). Alternative to this, the core 3 can be made by the addition of GlcNAc via β-1,3-N-acetylglucosaminyltransferase 6 (B3GNT6).

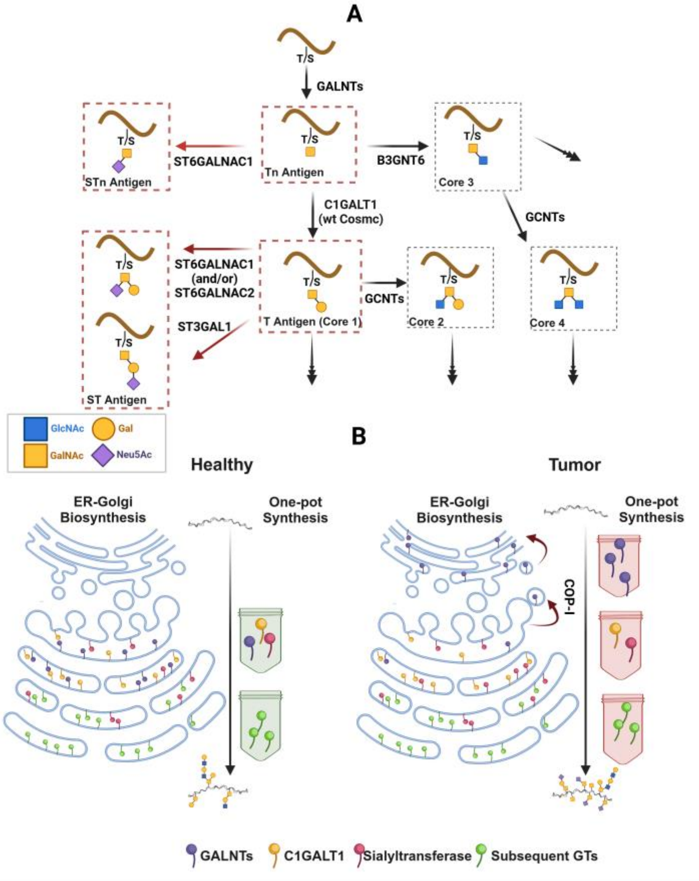


Figure 1. Mucin O-GalNAc glycosylation. A. O-GalNAc glycosylation pathway showing the reactions that lead to the formation of the cancer-associated Tn and T antigens and their sialylated forms sTn and sT, respectively. B. Differential localization of GALNTs between normal and epithelial cancer attributed to COP-1 mediated retrograde activation in cancer and the corresponding in vitro synthesis design.

These foundational structures undergo further branching and elongation with successive additions of monosaccharides such as GlcNAc and galactose, generating diverse glycan chains. Sialylation of Tn and T antigens produce their sialylated forms (sTn and sT), mediated by specific sialyltransferases (Figure 1 A). Clinically, Tn, T, sTn, and sT antigens are significant through their role in establishing the hallmarks of cancer such as tumor progression, immune evasion, and metastasis.[14, 15] This makes them central biomarkers and therapeutic targets in cancer care. The glycosylation of proteins takes place in the ER-Golgi system mostly through glycosyltransferases (GTs) and in some instances in combination with glycosidases. These glycoenzymes are distributed across specific cisternae.[10, 11, 16] The distribution of GTs is determined by the length of their transmembrane domains in correlation to the thickness and lipid content of the cisternae membrane. The localization of GTs along the ER-Golgi axis is dynamic, and these enzymes are constantly shuffled in both directions via a complex but tightly regulated vesicle system involving COP-I and COP-II vesicles.[16, 17] This localization across various cisternae has led to an assembly line of compartments performing sequential glycosylation to build glycans on target proteins. In the case of an organism disease state a protein's glycan is often altered when there is deregulation of this localization, such as the relocalization of GALNTs from the cis Golgi to the ER in tumour formation.8, 9 (Figure 1, B).

2.1 The Sequential One-pot synthesis method:

To capture the spatial-temporal segregation of glycosylation pathways and model the effect of the altered localization of GTs on a target protein glycosylation, we constructed a sequential glycosylation of MUC1 using a one-pot glycosyltransferase glycan synthesis assembly line as a construction platform. A fusion tag protein carrying the MUC1 peptide was designed as the assembly vehicle (Figure 2 A). The carrier vehicle (fusion protein and tags) was tested for biosynthesis interference (Figure 2 B). In the assembly line design for a glycan biosynthesis, the kinetics of the GT-catalyzed reactions and the associated intermediate glycan products are analyzed at every point of construction along the assembly line (Figure 2 C). The data obtained at each point informs subsequent steps, supporting model construction and iterative optimization of the synthesis.

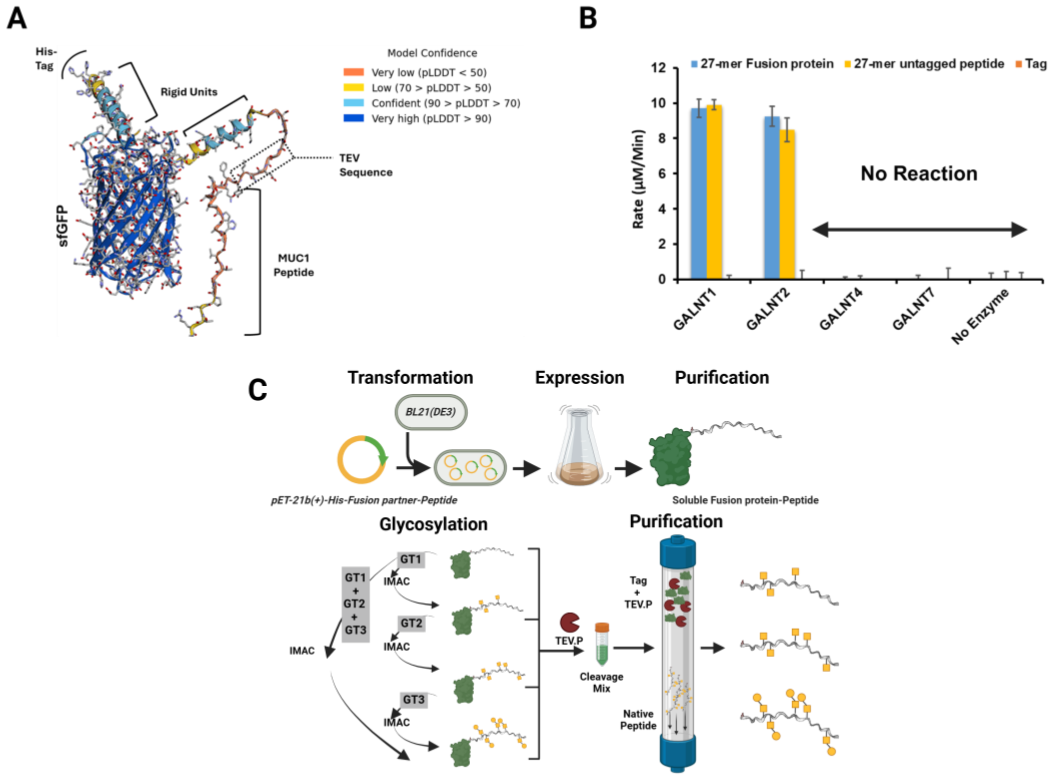


Figure 2. The one-pot biosynthesis method. A. An AlphaFold structure of the fusion protein for peptide expression. The fusion protein design connected to the MUC1 frame selected from the peptide tandem repeat. B. The initial reaction rates for the 27-mer MUC1 peptide in its tag-fused and unfused forms in addition to the tag used as acceptors for GalNAc catalysed by different GALNTs. C. The pipeline of expression and enzymatic glycosylation of the MUC1 peptide used to construct the pathway model.

The MUC1 peptide was selected to illustrate the golgi-ER simulated biosynthesis because the alteration to its glycosylation in the onset of tumor formation is not well understood despite the central role this antigen plays in epithelial cancers. A fusion protein containing the core MUC1 peptide and a carrier protein, superfolder green fluorescent protein (sfGFP), was expressed in E. coli. sfGFP was selected for its folding efficiency, minimized dimerization, and enhanced solubility.[18] To minimize the possibility of interaction with the MUC1 peptide, sfGFP was separated from the peptide using a linker (Linker 2) comprising of three rigid and three flexible units (from the N to C direction). Additionally, a rigid linker (Linker 1) was incorporated to improve steric presentation to the N-terminus His-tag, enhancing affinity-based purification. A tobacco etch virus (TEV) protease recognition sequence, was added to enable peptide cleavage at any step of the synthesis, retaining the native MUC1 sequence (Figure 2A and S1B). These features were designed to facilitate in vitro enzymatic glycosylation and enable simple one-step purification (Figure 1C). The carrier protein, along with the linkers and the His-tag, is collectively referred to as the "tag" throughout the manuscript. The design of the fusion protein, ensuring sufficient separation between the fusion protein vehicle and the glycosylation target peptide to prevent interference in the synthesis regime was achieved through the assistance of AlphaFold structure prediction tools. The rigid regions have greater conformational and structural predicted confidence compared with the flexible regions (Figure 2 A). Central to the design are the work functions of the linkers. Firstly, the rigid region on Linker 2 must maximize the peptide sfGFP distance. Secondly, the rigid linker 1 must maximize the presentation of the His-tag for later TEV protease cleavage when salvaging the glycosylated MUC1. Here the predicted low conformational and structural predicted confidence around the TEV protease cleavage region signifies flexibility and so the designed accessibility of the protease.

To ensure that the tag elements were inert and did not interfere with glycosylation, a fusion protein containing a MUC1 27-mer peptide was designed and expressed (Figure S1). The protein was cleaved using TEV protease, yielding the 27-mer peptide and the tag. The tag cleavage was confirmed by SDS-PAGE (Figure S2 B). Equal concentrations of the 27-mer in its fusion form and cleaved form in addition to the tag were tested as acceptors for GalNAc glycosylation by four GALNT enzymes: GALNT1, GALNT2, GALNT4, and GALNT7 (Figure 2B). The glycosylation rates were identical for the tagged and untagged MUC1 for all GTs. This is evidence that tag does not affect the enzymatic activity or substrate specificity of GALNTs. The fusion protein glycosylation carrier function was therefore optimized while preserving the inherent glycan recipient functions of the target peptide (MUC).

2.1.1 MUC1 model peptide design:

The biosynthesis of GalNAc O-linked glycans and forming the Tn antigen, are initiated through the polypeptide GalNAc-transferases (GALNTs) catalysis of the reactions forming the α-linkage between GalNAc and Serine or Threonine residues. Each GALNT has a catalytic and lectin-binding domain. There are three mechanisms through which GALNT glycosylation can occur, (i) glycosylation of the naked peptide using only the catalytic domain, (ii) glycosylation of pre-glycosylated peptides through the lectin domain, and (ii) a combination of the two mechanisms occurring sequentially. The lectin-binding domain functions as an anchor binding to pre-glycosylated sites limiting diffusional forces to make focused access to neighboring serine or threonine residues possible for the catalytic domain. Several structural elements including the subunits comprising the lectin domain, the properties of the linker between the catalytic and lectin domain, and the structure of the catalytic domain, determine the specificity and the mechanism of the sequential glycosylation for each GANT. [19-21]

Previously, the specificity of the isoforms GALNT1, 2, and 4 were extensively investigated either using the MUC1 peptide containing multiple tandem repeats or a single repeat with extension of the N terminal to the first threonine residue, known as TAP-24 peptide (Figure S1, A).[21-25] These designs revealed the specificities of these enzymes. However, we discovered that the TAP24 MUC1 construct employed in these studies is not a representative repeat of the natural tandem repeat able to illustrate the chemical biological stepwise glycosylation process. The optimal sequence must be inclusive of all the variables that determine GALNT specificity, and each variable must only be represented once in the peptide sequence. The commonly used TAP24 MUC1 construct is therefore unsuitable for the quantification of enzyme specificities and the biosynthesis of glycosylated MUC1.

The following criteria was therefore set for the optimal MUC1 peptide sequence: (1) includes all the five unique potential glycosylation sites of the MUC1 tandem repeat, and each site is only represented once in the sequence, (2) none of the sites is located at the peptide terminus, and the tandem repeat must be sufficiently extended in both directions of the glycosylation site to account for the motif specificity of the catalytic domain, (3) the frame of the sequence must be optimized for the position of each site in relation to the rest of the sites to accommodate the direction specificity of the lectin domain in GALNTs. To construct the optimal MUC1 peptide model, two peptides were designed, as illustrated in Figure S1, A, and Figure 3, A: the 27-mer and the 23-mer. While each peptide independently meets the first two criteria, their combined design satisfies the third criterion.

Expression of the 23mer and the 27mer was carried out in E. coli BL21 (DE3). A single-step purification using IMAC yielded around 60 mg of pure fusion protein from 250 ml cell culture. Tag cleavage using TEV protease was performed and SDS-PAGE gel confirmed complete cleavage of the naked peptide and the peptide displaying glycosylation on various sites (Figure S2).

2.1.2 In vitro healthy vs. tumor GT distribution and concentration models

The GalNAc-glycosylated sites are subject to either sialylation (addition of Sia) via ST6GALNAC1, galactosylation (addition of Gal) via C1GALT1 and its chaperone C1GALT1C1 (Cosmc), or N-acetylglucosaminylation (addition of GlcNAc) via B3GNT6 (Figure 1 A). In healthy contexts, GALNTs and C1GALT1 are localized in the cis-Golgi, while the ST6GALNAC1 is distributed across all the cisternae of the Golgi (Figure 1, B).[10, 11, 16] No specific localization of B3GNT6 has been reported. The localization of only the GALNTs were reported to be altered in response to the EGF stimulation of SRC (the proto-oncogene in cancer) via COP-1 mediated retrograde from cis Golgi to the ER.[26] This relocation results in the overexpression of Tn in the ER where a fraction of this Tn transits to the cell surface without modification, while another fraction transits with modification to T antigen. In patient samples, the same study found that the mean expression of Tn in breast cancer tissue samples was 4.5 folds higher than in normal tissues. From the samples with high Tn expression, 70 % of the samples showed ER localization of GALNTs (inferred indirectly from ER localization of Tn), whereas no significant loss of C1GALT1 was detected in these samples, pointing to the ER localization as the driving factor of the observed Tn overexpression.8, 9

Accompanying the relocation of GALNTs, the expression of O-GalNAc glycosylation enzymes is altered in cancer compared to normal cells. Furthermore, ST6GALNAC1 was reported to be upregulated in almost all cancer types[27] and C1GALT1 downregulated, mainly due to Cosmc mutation or epigenetic alteration of both C1GALT1 and Cosmc.[28, 29] The objective here is to build a one-pot in vitro synthesis model that will be representative of the impact of the redistribution of GALNTs in altering glycan structures independently of enzyme levels (Figure 1 B and Figure 2 C). To achieve this, all activities and kinetics experiments are performed at a standardized enzyme concentration of 250 nM.

The performance of the 23-mer and 27-mer peptides was compared by measuring the relative reactivity of GALNT1, GALNT2, GALNT4, and GALNT7 individually and in various combinations using the UGC assay13 (Figure 3A). The primary observation from both experiments is that GALNT4 alone does not react with either peptide, confirming it relies on a strictly lectin-dependent mechanism. GALNT7 shows no reactivity with either peptide in any combination. On the other hand, GALNT1 and GALNT2 could individually react with both peptides due to their direct catalytic mechanism. The rates of glycosylation via GALNT1 and GALNT2 were generally lower for the 23-mer than the 27-mer. This is explained by the depletion of the direct glycosylation-specific sites T13 and T20 for GALNT1 and GALNT2 respectively, in the 23-mer case, in addition to the absence of secondary lectin-assisted sites in the preferred direction. In contrast, the 27-mer permits the continuation of glycosylation via the lectin-assisted sites (Figure 3 B). When GALNT4 was tested in combination with GALNT1 or GALNT2, using the 27-mer construct, it showed no activity. However, GALNT4 exhibited significant activity using the 23mer peptide when combined with GALNT2, as indicated by the enhanced glycosylation rate when compared with the reaction of GALNT2 individually. GALNT2 is known to have a preferred specificity to T20 (in the PGST sequence) via direct catalytic domain recognition. In the 23-mer, T20 is at the N-terminal to T13 and S9 (GALNT4 specific sites), and pre-glycosylation at T20 left to T13 and S9 sites is the prerequisite for GALNT4 lectin-dependent specificity (Figure 3 B).[22] The collective evidence presented here confirms the efficiency of the fusion protein design to mimic natural MUC1 while its tag components do not affect the catalytic function of the GALNTs. Additionally, 23-mer has proved valid to capture the reaction mechanism of GALNT4 and the direct mechanism for both GALNT1 and GALNT2. However, the reported lectin-dependent mechanisms for GALNT1 and GALNT2 can be better studied using the 27mer.

2.2. In vitro synthesis of the Tn antigen

In vitro synthesis of the Tn antigen was carried out to explore the possible structures enabled by extensive glycosylation (higher enzyme concentrations for a longer period) using varying combinations of GALNTs (Figure 3 C I). LC-MS results confirmed the structure and purity of the peptide (i). The catalytic action of GALNT1 results in glycosylation of 2 or 3 sites, and 3 sites in the case of GALNT2. The glycosylation of T8 and T20 is consistent with a direct GALNT1 and GALNT2 mechanism respectively. However, the observed glycosylation of the remaining two sites cannot be attributed to the lectin-dependent mechanisms as previously reported for the GALNT1 and GALNT2 activity on MUC1.[22, 24] This is evident from short-term recorded activity on the remaining sites (Figure 3, A) since the 23mer does not provide the directionality required for this mechanism. A randomized peptide sequences study found that GALNT1 can catalyze a long-range N terminal lectin-dependent mechanism while GALNT 2 can participate in both N and C long-range lectin-dependent mechanism.[30] Another study confirmed the bi-directionality of the lectin-dependent mechanism for both enzymes with preference given to N and C long range lectin-dependent mechanisms for GALNT1 and GALNT2, respectively[21]. No further glycosylation beyond three sites was observed when GALNT1 and GALNT2 were combined, suggesting the absence of synergy between the two enzymes and that the two remaining sites do not conform to the specificity of these GTs.

When the product resulting from GALNT1 and GALNT2 activity was glycosylated with GALNT4, all five sites were glycosylated. These results indicate that the remaining two sites are GALNT4 specific and are glycosylated via a lectin-dependent mechanism. The consensus from all previously reported results is that T8 and T20 are specific sites for GALNT1 and GALNT2 respectively, and S9 and T13 are GALNT4 specific. Taken together, it can be concluded that GALNT1, GALNT2, or their combination can glycosylate S19, T20, and T8 by utilizing direct and lectin-dependent mechanisms. However, GALNT1 is less efficient than GALNT2 in completing the glycosylation of either S19 or T20 (Figure 3 C III).

It is widely accepted that the GalNAc site occupancy increases with GALNTs over expression and their ER relocation.8, 9 The effect of these two factors was simulated here in the One-pot biosynthesis model where extensive glycosylation was performed to interrogate the action of GALNTs in isolation of other GTs. It was seen that sites, such as T20, S19, and T8 were not selective and can be glycosylated by multiple GALNTs, such as GALNT1 and GALNT2 as shown here, as well as GALNT3 that was previously reported.[22, 24] We now refer to these sites that are the first ones to be glycosylated as GALNT_D-sites. The S9 and T13 sites are strictly lectin-dependent and thus are dependent on prior glycosylation of T20, S19, and T8 which we refer to now as GALNT_L-sites (Figure 4 A). The lectin-dependent mechanism is therefore responsible for high-density GalNAc-O-glycosylation.[25] This differential site occupancy was also found to be associated with cancer transformation. The site saturation was observed in MUC1 expressed in tumor cells compared with normal MUC1 in breast milk.[31] The GalNAc3-23mer (3 glycosylated sites at GALNT_D-sites only) and GalNAc5-23mer (5 glycosylated sites at GALNT_D-sites and GALNT_L-sites), synthesized here will be used as models for site occupancy, semi-glycosylated and completely-glycosylated, respectively.

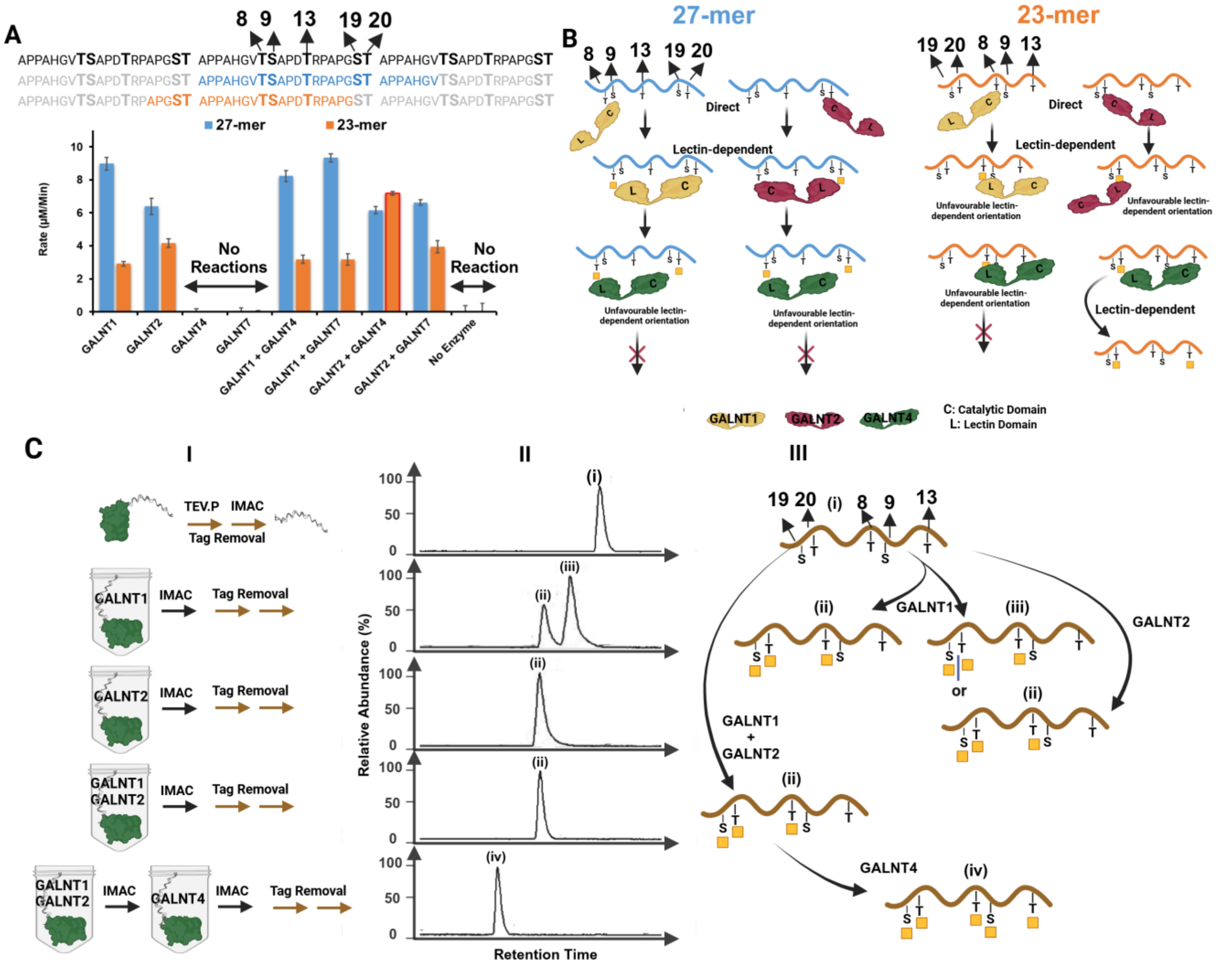
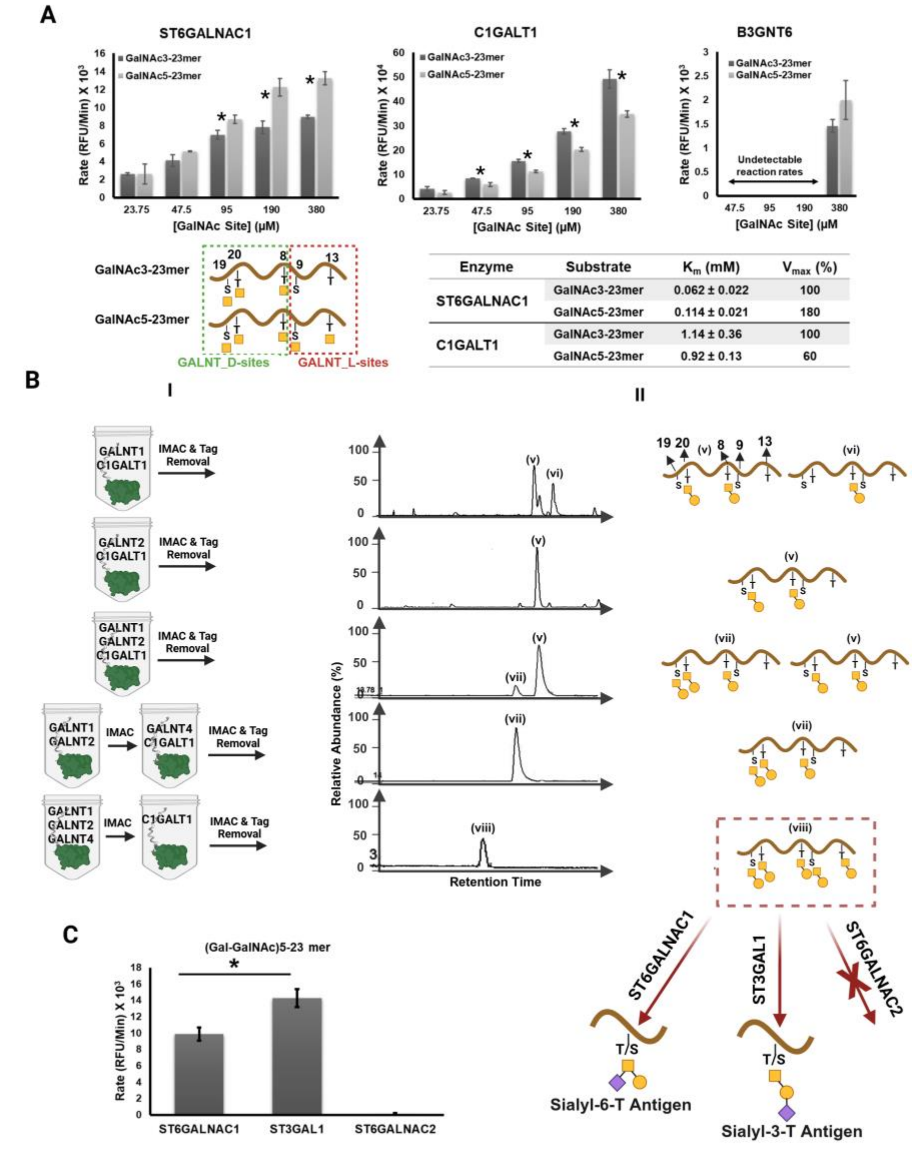


Figure 3. O-GalNAc glycosylation pathway initiation (Synthesis of Tn antigens). A. Reactivity of the 23-mer and the 27-mer peptides with the different GALNT enzymes. Initial rates were calculated from the linear range of the progress curves of the reaction and reported as (Mean ± SD). B. A schematic illustration of the direct and lectin-dependent mechanisms of the glycosylation reactions of GALNT1, GALNT2 and GALNT4 with the 23-mer and the 27-mer peptides. C. In vitro synthesis of GalNAc-glycosylated 23mer: I. Brief description of synthesis designs and steps. II. Abundance of the products of each reaction as measured by liquid chromatography. III. Structures of the products of each reaction as determined from the LC-MS analysis and the synthesis pathway leads to these structures as concluded from the preceding Data. Details of the LC-MS data and analysis are in the Supportive Information.

2.3 In vitro synthesis of the T and sTn antigen and core3

The sequential addition of one or two sugars to the Tn antigen generates diverse glycan core structures, with eight different cores identified. The predominantly structures cores 1-4 (Figure 1 A) and cores 5-8 are rare.[32] Core 1, also known as T antigen, is formed by adding galactose to Tn antigen via a β1-3 bond, a process catalyzed by T synthase (C1GALT1) with the help of its chaperone C1GALT1C1 (Cosmc). Core 3 is synthesized by adding GlcNAc to Tn antigen through a β1-3 bond, catalyzed by B3GNT6. ST6GALNAc1 sialylates Tn antigen to sialyl-Tn (sTn) that terminates the synthesis preventing the structures from undergoing further glycosylation via C1GALT1 or B3GNT3. Core 2 and Core 4 are synthesized by extending Core 1 and core 3 structures via GCNTs enzymes that can be extended to more complex structures (Figure 1 A).

The reactivity of the three enzymes were compared for GalNAc3-23mer (semi-glycosylated) and GalNAc5-23mer (completely-glycosylated) as models for site occupancy. Serial dilutions of both substrates were prepared by normalizing the concentrations to the number of GalNAc-glycosylated sites (Figure 4 A). While the overall activity of C1GALT1 was much higher than that of ST6GALNAC1, the results show that ST6GALNAC1 preferably sialylates the completely-glycosylated MUC1 whereas C1GALT1 preferably glycosylates the semi-glycosylated MUC1. The difference in the selectivity of both C1GALT1 and ST6GALNAC1 increases with an increase in MUC1 concentration. The kinetics parameters derived from the dose-response of ST6GALNAC1 for both semi- and completely-glycosylated MUC1 indicates that the enzyme has a lower affinity (Km value of 0.114 mM for GalNAc5-23mer vs 0.062 mM for GalNAc3-23mer) but higher turnover (Vmax for GalNAc5-23mer is 180 % of the Vmax of GalNAc3-23mer) of the fully glycosylated MUC1 compared with the semi glycosylated MUC1 (table in Figure 4 A). On the other hand, no significant difference in the Km values of C1GALT1 were observed between the semi-saturated and completely saturated MUC1. This indicates that C1GALT1 prefers glycosylation of the GALNT_D-sites over the GALNT_L-sites at any concentration of MUC1 and ST6GALNAC1 prefers glycosylation of the GALNT_L-sites over the GALNT_D-sites at high concentrations of MUC1. This is significant because MUC1 is frequently upregulated in tumors, consequently its upregulation in addition to its site occupancy may influence its glycosylation profile. Glycosylation of both acceptors with B3GNTs was very slow and undetectable at lower concentrations of the substrates. However, the rates at the highest tested concentration (380 µM) did not show differential specificity. Finally, the activity of ST6GALNAC2 on both acceptors was tested as well, and no significant glycosylation was detected despite the confirmation of expression of ST6GALNAC2 in active form when tested with asialofetuin.



Enzyme	Substrate	K_m (mM)	V_{max} (%)
ST6GALNAC1	GalNAc3-23mer	0.062 ± 0.022	100
	GalNAc5-23mer	0.114 ± 0.021	180
C1GALT1	GalNAc3-23mer	1.14 ± 0.36	100
	GalNAc5-23mer	0.92 ± 0.13	60

Figure 4. Assessment of the first layer of the structure extension after Tn synthesis. A. Effect of GalNAc site occupancy on the specificity of ST6GALNAC1, C1GALT1 and B3GNT6. Serial dilution of concentrations of GalNAc3-23mer and GalNAc5-23mer were calculated per GalNAc site occupancy. Data is presented as (mean ± SD, n = 3). B. Core 1 Synthesis: Effect of GALNTs vs C1GALT1 competition on site occupancy. I. synthesis designs of different combinations of sequential synthesis reflecting C1GALT1 competition with GALNTs in isolation or in different combinations. II. shows the LC peaks and structures deduced from analyzing mass spectrometry results (Supportive Information). C. Sialylation of Core 1 via ST3GAL1, ST6GALNAC1, and ST6GALNAC2. Bars in panels A and C are represented as means ± SD, n = 3), pairs marked with asterisk indicates p value < 0.01.

2.3.1 Evaluating ST6GALNAC1 Tn site specificity

The formation of the sTn MUC1 is a key antigen in several cancers consequently a detailed molecular description of the location of this epitope on the MUC1 frame is essential for drug discovery as well as vaccine development. In the section detailing Michaelis-Menten kinetics experiments (Figure 4 A) it was revealed that the GALNT1_D-sites are slowly sialylated despite high MUC1 to ST6GALNAC1affinity compared with the GALNT1_L-sites that are more rapidly sialylated although, the MUC1 to ST6GALNAC1 affinity is weaker. To understand the molecular reasons for this, computational reaction dynamics simulations of ST6GALNAC1 were undertaken on the fully glycosylated peptide (GalNAc5-23mer) and its semi glycosylated form (GalNac3-23mer). The study focused on three distinct MUC1 reaction configurations: (I) the sialylation of the T20 residue on a partially glycosylated peptide, (II) the sialylation of the T20 residue on a fully glycosylated peptide, and (III) the sialylation of the T13 residue on a fully glycosylated peptide (Figure 5A). The MUC1-enzyme poses (Figure 5 B) for each reaction configuration (I, II and II) were optimised as described in 4.7 below. Following this the Free Energies of Adaptive Reaction Coordinate Forces (FEARCF)[33] computational method was used to produce reaction trajectories. Typical MUC1-ST6GALNAC1 poses along the trajectories are shown (Figure 5 B). Free energy reaction profiles were extracted in the form of minimum energy pathways for each reaction so providing the energetic details for the molecular transformation of reactants through two transition states to products (Figure 5 C). While the reaction mechanisms are common to all three sialylation processes (Figure 5 D), these simulations revealed critical insights into the differences in each of the peptide enzyme binding as well as the molecular reaction kinetics and mechanisms.

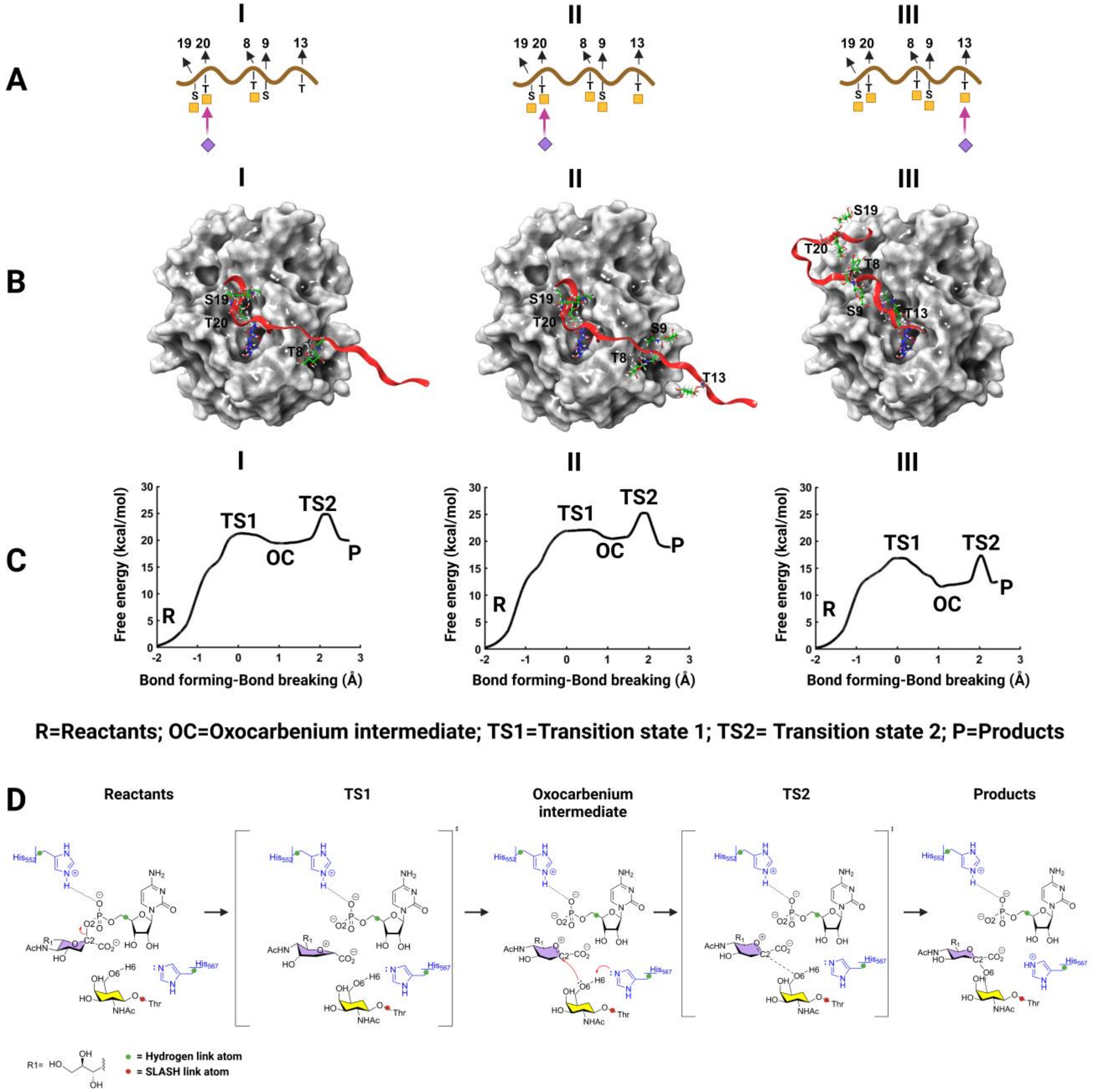

A
B
C
D
I
II
III
19 20 8 9 13
S T T S T
S19
T20
T8
S9
T13
Free energy (kcal/mol)
Bond forming-Bond breaking (Å)
R
TS1
OC
TS2
P
R=Reactants; OC=Oxocarbenium intermediate; TS1=Transition state 1; TS2= Transition state 2; P=Products
Reactants
TS1
Oxocarbenium intermediate
TS2
Products
His552
His567
R1=
= Hydrogen link atom
= SLASH link atom

Figure 5. Computational modeling of the reaction site specificity of ST6GALNAC1 A. Schematic representation of the reactions modeled. I. Sialylation of T20 on a partially glycosylated peptide. II. Sialylation of T20 on a fully glycosylated peptide. III. Sialylation of T13 on a fully glycosylated peptide. B. Snapshot of 23mer MUC1 (red ribbon) undergoing glycosylation with its GalNac residues (green ball and stick) binding to the ST6GALNAC1 surface (colored white) harboring the CMP-Neu5Ac donor (blue). C. 1D representation of the minimum energy pathway determined from the free energy surface showing the free energy associated with the bond forming and bond breaking in going from the reactants (R) to the products (P) via two transition states (TS1 & TS2) and an oxocarbenium intermediate (OC) D. SN1-like reaction mechanism of ST6GALNAC1. The position of link atoms added to prepare the system for QM/MM simulations are marked in green and red. Enzyme amino acids drawn in blue

The minimum energy pathways (MEPs) were determined as one-dimensional (1D) reaction coordinates (Figure 1C). These were expressed as a linear combination of bond forming (C2-O6) and bond breaking (C2-O2) primary reaction coordinates (Figure 5D). The calculated MEPs were consistent with previously proposed mechanisms in the literature, specifically the desiccation-driven mechanism.[34] The MEPs for each sialylation reaction reveals that the formation of an intermediate after the cleavage of CMP is a mechanistic feature common to all three setups. This intermediate is pivotal in the two-step reaction mechanism, reinforcing that glycosylation by ST6GALNAC1 involves a very coordinated series of substrate catalytic domain interactions. One of the key observations from the simulations was the significant impact of substrate glycosylation on the energy landscape and reaction kinetics. The sialylation of the T20 residue on partially glycosylated peptides (I) exhibited an energetic profile that is distinct from its fully glycosylated (II) counterpart (Figure 5 C). Similarly, the T13 residue sialylation on fully glycosylated peptides presented unique energetic and kinetic characteristics.

The sialylation of the T20 on the semi-glycosylated peptide formed the oxocarbenium intermediate (OC) with an energy of 19.42 kcal/mol after surmounting a transition state 1 (TS1) energy barrier of 21.24 kcal/mol. The sialylated product results after the Michaelis complex overcomes a second transition state 2 (TS2) barrier of 24.82 kcal/mol. In the case when Tn is sialylated at the T20 on a completely-glycosylated peptide, product formation was observed following transition state energy barriers of 22.12 kcal/mol (TS1) and 25.21 kcal/mol (TS2). The elevated reaction energy profiles of the sialyation at the T20 GALNT1_D-site for both reaction configurations is consistent with the slower reactivity observed experimentally and detailed in 2.3 above.

The sialylation of T13 on the completely glycosylated peptide proceeded via a transition state 1 (TS1) energy barrier of 16.88 kcal/mol to form a stable oxocarbenium intermediate (OC) with an energy of 11.39 kcal/mol. The formation of products was observed after overcoming a transition state 2 (TS2) energy barrier of 17.46 kcal/mol. This confirms the preference ST6GALNAC1 for GALNT1_L-sites over GALNT1_D-sites observed in the Michaelis-Menten kinetics experiments (section 2.3). The molecular reasons for this are that the sugar conformation necessary for the nucleophilic attack is critical to the sialylation reaction. If the sugar conformation is not within the near attack conformation cone angle, the histidine may covalently bind the anomeric carbon after the disassociation of the phosphate. This is the scenario in the T20 I and II T20 case where a side product forms when the catalytic histidine (HIS657) is covalently bonded to the anomeric carbon of the sialic acid lead. For the reaction to proceed without side product formation, the primary alcohol group must be sandwiched between the anomeric carbon of the sialic acid and the proton accepting nitrogen of the catalytic histidine. The most stable sandwiched structure occurs for the Michaelis complex at the T13 site. This leads to an efficient conversion of reactants to products, and with no side product formation. These molecular details explain the observed differences in rates that are experimentally measured and discussed in 2.3 (Table in Figure 4A).

2.3.2 In vitro synthesis of Core 1 (T antigen)

The synthesis of Core 1 via C1GALT1 was performed using different one-pot designs to simulate the varying distribution of GALNTs across the ER-Golgi system between healthy and tumor settings. Two scenarios were modeled: (1) when C1GALT1 competes with GALNTs, mimicking their co-localization in the cis-Golgi, and (2) when GALNTs act on the peptide first in isolation, representing their re-localization to the ER. When C1GALT1 was mixed with GALNT1 (Figure 4B), the chromatogram displayed two peaks, indicating the incorporation of Gal-GalNAc at one or two sites. When C1GALT1 was mixed with GALNT2, a uniform product with two-site occupancy was observed. These findings demonstrate that mixing either GALNT1 or GALNT2 with C1GALT1 reduces site occupancy by one compared to when GALNTs act alone. This suggests that after GalNAc is added to T8 or T20 via the direct mechanism of GALNT1 or GALNT2, respectively, C1GALT1 competes with these enzymes' lectin domains for the modified sites. This competition results in the synthesis of Core 1 structures and inhibits subsequent GalNAc glycosylation at other sites via the lectin-dependent mechanism of GALNTs.

In a one-pot reaction containing GALNT1, GALNT2, and C1GALT1, a major peak corresponding to a peptide with two Gal-GalNAc modifications was produced. This represents a reduction of one glycosylation site compared to the reaction with GALNT1 and GALNT2 alone (Figure 3C). These results suggest that GalNAc glycosylation of S19 by GALNT1 and/or GALNT2 occurs exclusively via a lectin-dependent mechanism. A similar inhibition was observed for GALNT4: When C1GALT1 and GALNT4 reacted with the product of glycosylation by GALNT1 and GALNT2, only three sites with Core 1 structures were identified (Figure 4B), indicating that the lectin-dependent mechanism of GALNT4 is also inhibited.

Finally, when simulating the ER localization of GALNTs (i.e., when the product of GALNT1,2 and 4 in combination was incubated with C1GALT1), C1GALT1 generated five sites occupied by Core 1 structures. The impact of GALNT co-localization with C1GALT1 on site occupancy can be extended to their co-localization with ST6GALNAC1. This is supported by a study showing that overexpression of ST6GALNAC1 reduced site occupancy by 25% in Chinese Hamster Ovary (CHO) cells.[16]

2.4 Sialylation of Core 1

Core 1 can be sialylated via ST3GAL1 to form the sialyl-3-T antigen or via ST6GALNAC family to form sialyl-6-T antigen (Figure 4 C). However, despite the reported activities of the three enzymes on the T antigen, the specificities of these enzymes were not tested comprehensively using standardized substrates. Additionally, details of ST6GALNAC1 vs ST6GALNAC2 specificities for Tn and T are conflicting across studies and the models (in vitro vs in vivo).[35, 36] Both ST6GALNAC1 and ST6GALNAC2 show no reactivity with the stand alone GalNAc or Gal-GalNAc acceptor, which confirms the peptide core requirement for both enzymatic activities.[35, 36] That was confirmed when both showed significant reactivity with asialofetuin (data not shown). We showed above that ST6GALNAC1 (but not ST6GALNAC2) reacts with Tn antigen. The same observation was true for the T antigen (Figure 4, C). When compared to ST6GALNAC1, ST3GAL1 showed significantly higher reactivity with the T antigen, thus the results suggest that sialyl-3-T is more predominant than sialyl-6-T.

3. Conclusion: in vitro constructed O-GalNAc glycosylation MUC1 model

The in vitro reconstruction of the MUC1 GalNAc-O-glycosylation pathway delivered new insights and previously unknown details of a) the competitive specificity of each enzyme to the substrate's MUC1 locale, b) the possible glycosylated combinations of MUC1 (glycoforms) produced by each enzyme as well as the predominance of the glycoforms, c) the effect of the sequence of the GTs in the glycosylation procession and d) the effect of GT compartmentation (co-localization) on the glycosylation products profile. To illustrate the utility of the approach, the reported effect of GALNTs relocating to the ER, on site occupancy was confirmed. However, insight into the nature of the GALNT relocation effect showed that two mechanisms are at the root of this effect. Firstly, the isolation of the GALNT in the ER extends the exposure to the substrate and so the time to react which leads to complete glycosylation and the saturation of the MUC1 glycosylation target sites through lectin-dependent mechanisms. Secondly, further glycosylation of the GALNT_D-sites by C1GALT1 and ST6GALNAC1 leads to the inhibition of the lectin-dependent mechanism of GALNTs, so preventing complete saturating MUC1 GALNT_L-sites with a primary GalNAc.

The in vitro enzyme specificity and competition at each step of the synthesis revealed that B3GNT6 specificity to MUC1 Tn sites are negligible compared with C1GALT1 and ST6GALNAC1. This finding corresponds with previous summations that core 3 and core 4 structures are less predominant in MUC1. Previously a comparison between normal epithelial breast cell lines and breast cancer cell lines recorded the absence of core 3 and core 4 structures from both types of cells.[37] Of greater note, the hypothesis that ST6GALNAC1 and ST6GALNAC2 may have an equal propensity to Tn and T antigens was debunked with the observation that the relative specificity of ST6GALNAC1 toward Tn and T antigens is significantly greater than ST6GALNAC2. The conclusion is that ST6GALNAC1 and not ST6GALNAC2 is responsible for the α-2 sialylation of these two antigens. When the reactivity of ST3GAL1 was compared to ST6GALNAC1 towards T antigen, ST3GAL1 was found more reactive, suggesting that sialyl-3-T (the product of ST3GAL1) is the more likely route to ST than the route via sialyl-6-T.

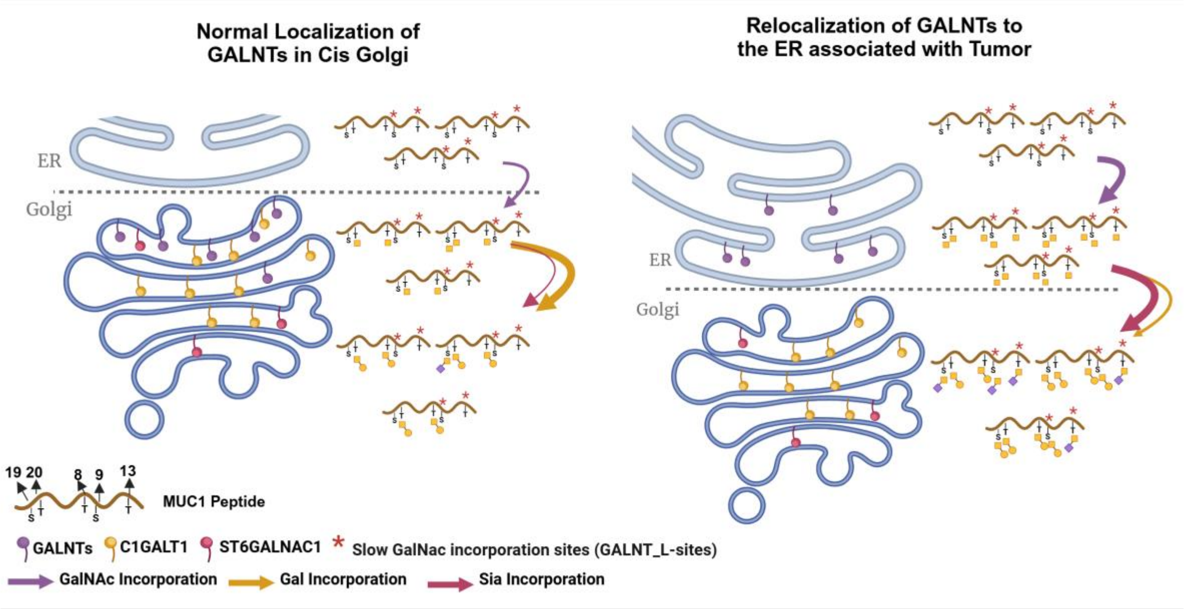

Normal Localization of GALNTs in Cis Golgi
Relocalization of GALNTs to the ER associated with Tumor
ER
Golgi
ER
Golgi
19 20 8 9 13
MUC1 Peptide
GALNTs
C1GALT1
ST6GALNAC1
* Slow GalNac incorporation sites (GALNT_L-sites)
GalNAc Incorporation
Gal Incorporation
Sia Incorporation

Figure 6. A proposed model of the effect of the localization of GALNTs on site occupancy and sialylation of the Tn antigen. The left panel illustrates normal cellular MUC1 glycosylation, while the right panel illustrates tumor cellular MUC1 glycosylation.

Interestingly, the specific reactivity of C1GALT1 and ST6GALNAC1 to the fully glycosylated or the semi glycosylated MUC1 by the Tn antigen revealed that C1GALT1 preferred adding a galactose to the GALNT_D-sites (associated with T8, S19, and T20) whereas ST6GALNAC1 prefers the completely glycosylated GALNT_L-sites (associated with S9 and T13). In fact, when revisiting the chemoenzymatic synthesis approach followed by Yoshimura et al. 20196 it is apparent that ST6GALNAC1 displayed significant specificity to the threonine T13 compared to other threonine residues T8 and T20. Moreover, in the same study, S19 was not sialylated by ST6GALNAC1 while low reactivity with S9 was detected.

The collective summary of these observations leads to a model explaining the differences in observed normal epithelial cell glycosylation compared with tumor epithelial cell glycosylation (Figure 6). In the normal case, the co-localization of GALNTs with C1GALT1 and ST6GALNAC1 in the cis-Golgi results in competition between GALNTs and ST6GALNAC1, with GALNTs preventing full occupancy of the five potential glycosylation sites on MUC1. In contrast in tumor cells, the localization of GALNTs in the ER, isolated from cis-Golgi GTs, allows slower GalNAc incorporation at specific sites to progress further. The site specificity study undertaken here leads to the proposal that this alteration in occupancy is a determinant to the specificity of ST6GALNAC1 and C1GALT1. The specificity of C1GALT1 is greater toward the to the GALNT_D-sites, which leads to enrichment of glycans with an extended Core1, whereas the tumor-associated, completely GalNAc glycosylated state is preferentially recognized by ST6GALNAC1 preferring sialylation at GALNT_L-sites that results in glycan truncation and the synthesis of the tumor-associated sTn antigen (at high density).

4. Materials and Methods

4.1 Expression of MUC1 peptide fusion protein and tag removal

Gene sequences encoding MUC1 peptide fusion proteins were synthesize and inserted in the pET-21b(+) E. coli expression vector by BIOMATIK (Ontario, Canada). The sequence of the synthesized fusion protein is provided in the supplementary information (Figure S1).

Transformation of E. Coli. BL21 (Sigma-Aldrich, Cat. no. CMC0014) was carried out using the heat shock protocol and grown on Luria Broth (LB) agar plates supplemented with ampicillin to a final concentration of 50 µg/ml at 37 °C. Expression was carried out in Terrific Broth (TB) medium overnight at 16 °C with shaking at 150 rpm. Glucose was added to a concentration of 2 % wt/vol at all stages of expression. Expression was induced by isopropyl β-D-thiogalactoside (IPTG, Sigma-Aldrich, Cat. No. 16758) to a final concentration of 1 mM at OD600 of 0.4-0.6.

The cell pellets from E. coli expression were lysed by incubating at 4 °C for 4 hours in an IMAC binding buffer consisting of 20 mM Tris-HCl, 5 mM Imidazole-HCl, 500 mM NaCl, 0.05% (w/v) sodium azide, and 10% (v/v) glycerol at pH 7.9. This buffer was supplemented with a protease inhibitor tablet (complete, Sigma Aldrich, Cat. no. 11873580001) and 20 mg of lysozyme per 10 ml of buffer. The cell lysate was then clarified through centrifugation at 48,000 RCF for 30 minutes followed by filtration using a 0.45 µM sterile filter. IMAC was conducted on a protein liquid chromatography (FPLC) system ÄKTA Start utilizing 1 ml HiTrap Chelating High-Performance columns (Cytiva, Cat. no. 17-0408-01). Proteins were eluted with a gradient ranging from 5 to 500 mM imidazole-HCl over 15 minutes at a flow rate of 1 ml/min. The eluted fractions were collected, and SDS-PAGE was used to verify protein purity and size. The purified enzymes were quantified with the Bradford protein assay kit (ThermoFisher, A55866) and stored at -80 °C in a freezing buffer comprising 20 mM Tris-HCl, 150 mM NaCl, and 10% (v/v) glycerol at pH 7.6.

TEV protease was expressed from the expression plasmid pRK793 (Addgene plasmid #8827; http://n2t.net/addgene:8827; RRID: Addgene_8827), in E. coli. BL21(DE3)-RIL cells as described previously. 13 Briefly, Tag removal was performed overnight at 4 °C by incubating the TEV protease with the fusion proteins at an optimized ratio (Supportive Information, Figure S2) in a buffer of 50 mM Tris-HCl, 0.5 mM EDTA, 1 mM DTT, pH 8.0

4.2 Expression of Glycosyltransferases

HEK293F cells (R79007, Thermo Fisher) were a gift from E. Sturrock (University of Cape Town, South Africa). All expression plasmids for GTs used in this work were purchased from the plasmid repository DNASU: HsCD00522282 for GALNT1, HsCD00413124 for GALNT2, HsCD00413161 for GALNT4, HsCD00413129 for GALNT7, HsCD00413109 for ST6GALNAC2, HsCD00413042 for C1GALT1C1, HsCD00413169 for ST3GAL1. As for the pGEn2 vectors of C1GALT1, ST6GALNAC1 and B3GNT6, they were obtained directly from Kelley Moremen (University of Georgia, The USA). GTs expressed from pGEn2 vectors are N-terminally tagged with signalling peptide-8X His-Avi tag-super folder GFP-Tev protease recognition sites. To remove the tag and obtain the enzymes in their native sequence, purified (tagged) proteins were incubated overnight at 4 °C with TEV protease at a ratio of 1:5 (TEV protease: fusion protein) in TEV protease buffer (50 mM Tris-HCl, 0.5 mM EDTA, 1 mM DTT, pH 8.0). The mixture was then processed by an IMAC to collect the untagged enzymes from the flow through fractions. Enzymes were stored at −80 °C in freezing buffer (20 mM Tris-HCl, 150 mM NaCl and 10% (v/v) Glycerol, pH 7.6).

4.3 Enzyme assay

The enzymatic reactions were monitored using the previously developed UGC assay,[13] the components of the assay are: l-Lactic dehydrogenase (LDH, L2500), PK (P1506), donors such as CMP-sialic acid (CMP-Neu5Ac, C1006), phosphoenolpyruvic acid monopotassium salt (PEP-K, 860077), adenosine 5′-triphosphate disodium salt trihydrate (ATP, 10519979001), β-Nicotinamide adenine dinucleotide, reduced disodium salt hydrate (NADH, 10128023001), bovine serum albumin (BSA, A3059), N-(2-hydroxyethyl) piperazine-N′-(2-ethanesulfonic acid), 4-(2-hydroxyethyl) piperazine-1-ethanesulfonic acid (HEPES, H3375), all purchased from Sigma Aldrich. The 27-mer peptide used as a standard naked peptide in Figure 2, B was synthesized at GL Biochem, Shanghai. Nucleoside diphosphate kinase (NDK) and cytidylate kinase (CMK) were expressed in house as described in the original protocol. Peptides used as fusion proteins were quantified using Bradford assay. The concentration of the GT enzymes was standardized in all the kinetics and activity assays at 250 nM. Donor concertation was standardized at 2mM. The induction of reaction and preparation of reaction mixtures were performed following the standard protocol.[13] Briefly, Master mixtures were prepared in CMK-NDK-PK-LDH format for sialyltransferase enzymes and in NDK-PK-LDH for the other enzymes. In addition to the coupling enzymes, master mixtures included: reaction buffer, BSA, NADH, phosphoenolpyruvate, ATP, donor, and the constant reaction components (enzyme or acceptor). The variable reaction component was distributed to the wells in triplicate and the induction was carried out by adding the master mixture to the well simultaneously. Corresponding controls including reactions without the variable component were also included. Fluorescence signals were recorded in real time and initial rates were calculated from the slope of progress curves within the initial linear range and normalized to the corresponding blank controls. Rates were converted from fluorescence unit to molarity/minute using the standard curved obtained for the nucleotide corresponding to the donor used in the reaction as explained in the original protocol.

4.4 Preparative glycosylation

The glycosylation of the peptides in their fusion form was performed in a one-pot reaction mixture containing the peptide fusion protein at a concentration of 100 µM, donors at 2 mM, GT enzymes at 560 nM. The buffer used was 50 mM HEPES-NaOH buffer, 50 mM KCl, 10 mM MnCl2, 10 mM MgCl2, and 0.02 vol/vol Triton X-100 pH 7.4. A total of 1 ml reaction mixtures were prepared and incubated for 8 hours at 37 °C. After completion of the reaction, the glycosylated peptides in their fusion form were purified in a one-step IMAC. The eluted fusion proteins were then subjected to buffer exchange to either freezing buffer (20 mM Tris-HCl, 150 mM NaCl and, 10% (v/v) Glycerol, pH 7.6) when a subsequent glycosylation reaction is performed or to TEV protease cleavage buffer (50 mM Tris-HCl, 0.5 mM EDTA, 1 mM DTT, pH 8.0) when LC-MS was performed.

4.5 LC-MS analysis

LC-MS analysis was performed using a Q-Exactive quadrupole-Orbitrap mass spectrometer (Thermo Fisher Scientific, USA) coupled with a Dionex Ultimate 3000 nano-UPLC system, and data was acquired with Xcalibur v4.1.31.9, Chromeleon v6.8 (SR13), Orbitrap MS v2.9

(build 2926), and Thermo Foundations 3.1 (SP4). Peptides were prepared in a solution containing 0.1% (v/v) formic acid (FA) and 2% (v/v) acetonitrile (ACN). Final concentrations of the peptides were estimated to be around 10 nM, with a volume equivalent to 50 fmol of peptide injected per sample. Samples were trapped on a PepMap100 C18 column and separated using a ReproSil-Pur 120 C-18-AQ column with a multi-step gradient of Solvent A (0.1% FA in LC water) and Solvent B (0.1% FA in ACN). The mass spectrometer was operated in positive ion mode at a capillary temperature of 320°C and an electrospray voltage of 1.95 kV. Full scan and data-dependent MS/MS settings were used to determine MS1 and MS2 m/z distributions, details of sample preparation and procedure are in the Supportive Information.

4.6 Statistics and Regression

All the data points were repeated three times. Shapiro–Wilk normality tests were performed to confirm the normal distribution of the data. Normality tests and Michaelis-Menten regression were performed using GraphPad Prism 8 software as described previously.13 The data are presented as (mean value) ± (standard deviation).

4.7 Computational methods

The ST6GALNAC1 structure was obtained from the SWISS-MODEL[38] database and used to prepare the Michaelis complex. The protein sequence of the CMP-binding position was determined by structural alignment with the crystal structure of ST6GALNAC2 (PDB ID: 6APL), which was subsequently modified to CMP-Neu5Ac. The binding position of the 23mer MUC1 peptide was determined using the HADDOCK[39] webserver to perform protein-protein docking. HADDOCK generated 10 top clusters, each with 4 representative poses. The lowest-energy cluster placed T13 of the MUC1 23mer peptide closest to the acceptor GalNAc residue in the catalytic pocket. The glycosidic bond between T13 and the GalNAc was modeled, and the resulting glycopeptide structure was energy-minimized. Additional GalNAc residues were modeled, producing glycopeptide III (Figure 5A). Subsequent mutations, peptide extensions, and glycosylations yielded glycopeptides I and II (Figure 5A).

The Michaelis complex for each system was modelled with the CHARMM36 force field[40]. All systems were solvated in TIP3P water molecules in a cubic water box with a 12Å buffer from the edge of the protein. The system was neutralized and NaCl ions were added to a concentration of 0.15M. Initial energy minimization was performed for 1,000 steps using the ABNR method followed by a 1000 step minimization using the SD method. Thermal equilibration was performed for 100 ps with an NPT ensemble and 100 ns classical molecular dynamics simulation was performed in an NVT ensemble. Particle-mesh Ewald (PME) summation was used to calculate electrostatic interactions. The Verlet cutoff scheme was applied to van der Waals (vdW) interactions with a cutoff distance of 12 Å and a switch distance of 10 Å. SHAKE algorithm is used in all simulations to constrain hydrogen bonds. Simulations were carried out with a time step of 2 fs. All simulations were performed at a temperature of 300 K and under a pressure of 1 bar.

The reaction dynamics simulations were modeled using hybrid quantum mechanical/molecular mechanical methods (QM/MM).[41] The QM region comprised HIS567 (catalytic base) sidechain atoms, HIS552 (stabilizing the CMP phosphate group), the Neu5Ac moiety, and the CMP phosphate group. The nonpolar C-C bonds were treated with hydrogen link atoms, and the polar C-O bonds were treated with the Simple Link Atom Saccharide Hybrid (SLASH)[42] method.

Free energy simulations were conducted along two reaction coordinates to monitor bond formation and breaking using the Free Energies of Adaptive Reaction Coordinate Forces (FEARCF)[33] method. Histograms for each reaction coordinate were generated using 110 bins spanning a sampling range of 0.5–6 Å. Twelve FEARCF iterations were performed, each comprising 120 simulations of 30 ps duration that was preceded by a 2 ps equilibration run culminating in production runs each totalling of 43.2 ns.

Simulations were performed in a 37 Å water sphere using stochastic boundaries. A buffer shell (30–37 Å from the system origin) was applied. The reaction region was modeled with the mio1-1 DFTB3 parameter set, which includes corrections for hydrogen bonding and dispersion interactions. Forces between partial charges beyond 12 Å were zeroed using an atom-wise force-shifting function. Dynamics in the system were modeled using Langevin dynamics with a 2 fs time step.

Funding Sources

The National Research Foundation (NRF) CPPR 466624 and the South African Medical Research Council Self-Initiated Grant (SAMRC SIG) 416090.

ACKNOWLEDGMENT

This work is based in part upon research supported by the National Research Foundation (NRF) CPPR 466624 grant (K.J.N.) and the South African Medical Research Council Self-Initiated Grant (SAMRC SIG) 416090 grant. A.N. Thanks to the Scientific Computing Research Unit (SCRU) for graduate fellowship funding. We thank the Centre for High Performance Computing (CHPC) for computational resources (CHEM0840).

REFERENCES

(1) Pothukuchi, P.; Agliarulo, I.; Russo, D.; Rizzo, R.; Russo, F.; Parashuraman, S. Translation of Genome to Glycome: Role of the Golgi Apparatus. *FEBS Lett* **2019**, *593* (17), 2390-2411. DOI: 10.1002/1873-3468.13541.

(2) Ashkani, J.; Naidoo, K. J. Glycosyltransferase Gene Expression Profiles Classify Cancer Types and Propose Prognostic Subtypes. *Scientific Reports* **2016**, *6*, 26451. DOI: 10.1038/srep26451.

(3) Pinho, S. S.; Reis, C. A. Glycosylation in Cancer: Mechanisms and Clinical Implications. *Nat Rev Cancer* **2015**, *15* (9), 540-555, Review. DOI: 10.1038/nrc3982.

(4) Murrell, M. P.; Yarema, K. J.; Levchenko, A. The Systems Biology of Glycosylation. *ChemBioChem* **2004**, *5* (10), 1334-1347. DOI: 10.1002/cbic.200400143.

(5) Goth, C. K.; Mehta, A. Y.; McQuillan, A. M.; Baker, K. J.; Hanes, M. S.; Park, S. S.; Stavenhagen, K.; Hjortø, G. M.; Heimburg-Molinaro, J.; Chaikof, E. L.; et al. Chemokine Binding to Psgl-1 Is Controlled by O-Glycosylation and Tyrosine Sulfation. *Cell Chemical Biology* **2023**, *30* (8), 893-905.e897. DOI: 10.1016/j.chembiol.2023.06.013 (acccessed 2024/05/02).

(6) Yoshimura, Y.; Denda-Nagai, K.; Takahashi, Y.; Nagashima, I.; Shimizu, H.; Kishimoto, T.; Noji, M.; Shichino, S.; Chiba, Y.; Irimura, T. Products of Chemoenzymatic Synthesis Representing Muc1 Tandem Repeat Unit with T-, St- or Stn-Antigen Revealed Distinct Specificities of Anti-Muc1 Antibodies. *Scientific Reports* **2019**, *9* (1), 16641. DOI: 10.1038/s41598-019-53052-1.

(7) Narimatsu, Y.; Joshi, H. J.; Nason, R.; Van Coillie, J.; Karlsson, R.; Sun, L.; Ye, Z.; Chen, Y. H.; Schjoldager, K. T.; Steentoft, C.; et al. An Atlas of Human Glycosylation Pathways Enables Display of the Human Glycome by Gene Engineered Cells. *Mol Cell* **2019**, *75* (2), 394-407 e395. DOI: 10.1016/j.molcel.2019.05.017.

(8) Gill, D. J.; Chia, J.; Senewiratne, J.; Bard, F. Regulation of O-Glycosylation through Golgi-to-Er Relocation of Initiation Enzymes. *J Cell Biol* **2010**, *189* (5), 843-858. DOI: 10.1083/jcb.201003055.

(9) Gill, D. J.; Tham, K. M.; Chia, J.; Wang, S. C.; Steentoft, C.; Clausen, H.; Bard-Chapeau, E. A.; Bard, F. A. Initiation of Galnac-Type O-Glycosylation in the Endoplasmic Reticulum Promotes Cancer Cell Invasiveness. *Proc Natl Acad Sci U S A* **2013**, *110* (34), E3152-3161. DOI: 10.1073/pnas.1305269110.

(10) Bui, S.; Mejia, I.; Diaz, B.; Wang, Y. Adaptation of the Golgi Apparatus in Cancer Cell Invasion and Metastasis. *Front Cell Dev Biol* **2021**, *9*, 806482. DOI: 10.3389/fcell.2021.806482.

(11) Rabouille, C.; Hui, N.; Hunte, F.; Kieckbusch, R.; Berger, E. G.; Warren, G.; Nilsson, T. Mapping the Distribution of Golgi Enzymes Involved in the Construction of Complex Oligosaccharides. *J Cell Sci* **1995**, *108 (Pt 4)*, 1617-1627. DOI: 10.1242/jcs.108.4.1617.

(12) Xi, X.; Wang, J.; Qin, Y.; Huang, W.; You, Y.; Zhan, J. Glycosylated Modification of Muc1 Maybe a New Target to Promote Drug Sensitivity and Efficacy for Breast Cancer Chemotherapy. *Cell Death Dis* **2022**, *13* (8), 708. DOI: 10.1038/s41419-022-05110-2.

(13) Nashed, A.; Naidoo, K. J. Universal Glycosyltransferase Continuous Assay for Uniform Kinetics and Inhibition Database Development and Mechanistic Studies Illustrated on St3gal1, C1galt1, and Fut1. *ACS Omega* **2024**, *9* (15), 17518-17532. DOI: 10.1021/acsomega.4c00485.

(14) Taylor-Papadimitriou, J.; Burchell, J.; Miles, D. W.; Dalziel, M. Muc1 and Cancer. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* **1999**, *1455* (2), 301-313. DOI: https://doi.org/10.1016/S0925-4439(99)00055-1.

(15) Nath, S.; Mukherjee, P. Muc1: A Multifaceted Oncoprotein with a Key Role in Cancer Progression. *Trends in Molecular Medicine* **2014**, *20* (6), 332-342. DOI: 10.1016/j.molmed.2014.02.007 (acccessed 2024/06/11).

(16) Sewell, R.; Bäckström, M.; Dalziel, M.; Gschmeissner, S.; Karlsson, H.; Noll, T.; Gätgens, J.; Clausen, H.; Hansson, G. C.; Burchell, J.; et al. The St6galnac-I Sialyltransferase Localizes Throughout the Golgi and Is Responsible for the Synthesis of the Tumor-Associated Sialyl-Tn

-Glycan in Human Breast Cancer. *J Biol Chem* **2006**, *281* (6), 3586-3594. DOI: 10.1074/jbc.M511826200.

(17) Pantazopoulou, A.; Glick, B. S. A Kinetic View of Membrane Traffic Pathways Can Transcend the Classical View of Golgi Compartments. *Frontiers in Cell and Developmental Biology* **2019**, *7*. DOI: ARTN 153

10.3389/fcell.2019.00153.
(18) Pedelacq, J. D.; Cabantous, S.; Tran, T.; Terwilliger, T. C.; Waldo, G. S. Engineering and Characterization of a Superfolder Green Fluorescent Protein. *Nat Biotechnol* **2006**, *24* (1), 79-88. DOI: 10.1038/nbt1172.
(19) Pratt, M. R.; Hang, H. C.; Ten Hagen, K. G.; Rarick, J.; Gerken, T. A.; Tabak, L. A.; Bertozzi, C. R. Deconvoluting the Functions of Polypeptide N-Alpha-Acetylgalactosaminyltransferase Family Members by Glycopeptide Substrate Profiling. *Chem Biol* **2004**, *11* (7), 1009-1016. DOI: 10.1016/j.chembiol.2004.05.009.
(20) de Las Rivas, M.; Lira-Navarrete, E.; Gerken, T. A.; Hurtado-Guerrero, R. Polypeptide Galnac-Ts: From Redundancy to Specificity. *Curr Opin Struct Biol* **2019**, *56*, 87-96. DOI: 10.1016/j.sbi.2018.12.007.
(21) Collette, A. M.; Hassan, S. A.; Schmidt, S. I.; Lara, A. J.; Yang, W.; Samara, N. L. An Unusual Dual Sugar-Binding Lectin Domain Controls the Substrate Specificity of a Mucin-Type O-Glycosyltransferase. *Sci Adv* **2024**, *10* (9), eadj8829. DOI: 10.1126/sciadv.adj8829.
(22) Coelho, H.; Rivas, M. L.; Grosso, A. S.; Diniz, A.; Soares, C. O.; Francisco, R. A.; Dias, J. S.; Companon, I.; Sun, L.; Narimatsu, Y.; et al. Atomic and Specificity Details of Mucin 1 O-Glycosylation Process by Multiple Polypeptide Galnac-Transferase Isoforms Unveiled by Nmr and Molecular Modeling. *JACS Au* **2022**, *2* (3), 631-645. DOI: 10.1021/jacsau.1c00529.
(23) Hassan, H.; Reis, C. A.; Bennett, E. P.; Mirgorodskaya, E.; Roepstorff, P.; Hollingsworth, M. A.; Burchell, J.; Taylor-Papadimitriou, J.; Clausen, H. The Lectin Domain of Udp-N-Acetyl-D-Galactosamine: Polypeptide N-Acetylgalactosaminyltransferase-T4 Directs Its Glycopeptide Specificities. *J Biol Chem* **2000**, *275* (49), 38197-38205. DOI: 10.1074/jbc.M005783200.
(24) Wandall, H. H.; Hassan, H.; Mirgorodskaya, E.; Kristensen, A. K.; Roepstorff, P.; Bennett, E. P.; Nielsen, P. A.; Hollingsworth, M. A.; Burchell, J.; Taylor-Papadimitriou, J.; et al. Substrate Specificities of Three Members of the Human Udp-N-Acetyl-Alpha-D-Galactosamine:Polypeptide N-Acetylgalactosaminyltransferase Family, Galnac-T1, -T2, and -T3. *J Biol Chem* **1997**, *272* (38), 23503-23514. DOI: 10.1074/jbc.272.38.23503.
(25) Wandall, H. H.; Irazoqui, F.; Tarp, M. A.; Bennett, E. P.; Mandel, U.; Takeuchi, H.; Kato, K.; Irimura, T.; Suryanarayanan, G.; Hollingsworth, M. A.; et al. The Lectin Domains of Polypeptide Galnac-Transferases Exhibit Carbohydrate-Binding Specificity for Galnac: Lectin Binding to Galnac-Glycopeptide Substrates Is Required for High Density Galnac-O-Glycosylation. *Glycobiology* **2007**, *17* (4), 374-387. DOI: 10.1093/glycob/cwl082.
(26) Luo, J.; Zou, H.; Guo, Y.; Tong, T.; Ye, L.; Zhu, C.; Deng, L.; Wang, B.; Pan, Y.; Li, P. Src Kinase-Mediated Signaling Pathways and Targeted Therapies in Breast Cancer. *Breast Cancer Res* **2022**, *24* (1), 99. DOI: 10.1186/s13058-022-01596-y.
(27) Hugonnet, M.; Singh, P.; Haas, Q.; von Gunten, S. The Distinct Roles of Sialyltransferases in Cancer Biology and Onco-Immunology. *Front Immunol* **2021**, *12*, 799861. DOI: 10.3389/fimmu.2021.799861.
(28) Ju, T.; Lanneau, G. S.; Gautam, T.; Wang, Y.; Xia, B.; Stowell, S. R.; Willard, M. T.; Wang, W.; Xia, J. Y.; Zuna, R. E.; et al. Human Tumor Antigens Tn and Sialyl Tn Arise from Mutations in Cosmc. *Cancer Res* **2008**, *68* (6), 1636-1646. DOI: 10.1158/0008-5472.CAN-07-2345.
(29) Zeng, J.; Mi, R.; Wang, Y.; Li, Y.; Lin, L.; Yao, B.; Song, L.; van Die, I.; Chapman, A. B.; Cummings, R. D.; et al. Promoters of Human Cosmc and T-Synthase Genes Are Similar in Structure, yet Different in Epigenetic Regulation. *J Biol Chem* **2015**, *290* (31), 19018-19033. DOI: 10.1074/jbc.M115.654244.
(30) Revoredo, L.; Wang, S.; Bennett, E. P.; Clausen, H.; Moremen, K. W.; Jarvis, D. L.; Ten Hagen, K. G.; Tabak, L. A.; Gerken, T. A. Mucin-Type O-Glycosylation Is Controlled by Short- and Long-Range Glycopeptide Substrate Recognition That Varies among Members of the Polypeptide Galnac Transferase Family. *Glycobiology* **2016**, *26* (4), 360-376. DOI: 10.1093/glycob/cwv108.

(31) Hanisch, F. G.; Muller, S. Muc1: The Polymorphic Appearance of a Human Mucin. *Glycobiology* **2000**, *10* (5), 439-449. DOI: 10.1093/glycob/10.5.439.

(32) Wang, S.; Chen, C.; Gadi, M. R.; Saikam, V.; Liu, D.; Zhu, H.; Bollag, R.; Liu, K.; Chen, X.; Wang, F.; et al. Chemoenzymatic Modular Assembly of O-Galnac Glycans for Functional Glycomics. *Nat Commun* **2021**, *12* (1), 3573. DOI: 10.1038/s41467-021-23428-x.

(33) Naidoo, K. J. Fearcf a Multidimensional Free Energy Method for Investigating Conformational Landscapes and Chemical Reaction Mechanisms. *Science China Chemistry* **2011**, *54* (12), 1962-1973. DOI: 10.1007/s11426-011-4423-7.

(34) Hansen, T.; Lebedel, L.; Remmerswaal, W. A.; van der Vorm, S.; Wander, D. P. A.; Somers, M.; Overkleeft, H. S.; Filippov, D. V.; Désiré, J.; Mingot, A.; et al. Defining the Sn1 Side of Glycosylation Reactions: Stereoselectivity of Glycopyranosyl Cations. *ACS Central Science* **2019**, *5* (5), 781-788. DOI: 10.1021/acscentsci.9b00042.

(35) Kono, M.; Tsuda, T.; Ogata, S.; Takashima, S.; Liu, H.; Hamamoto, T.; Itzkowitz, S. H.; Nishimura, S.; Tsuji, S. Redefined Substrate Specificity of St6galnac Ii: A Second Candidate Sialyl-Tn Synthase. *Biochem Biophys Res Commun* **2000**, *272* (1), 94-97. DOI: 10.1006/bbrc.2000.2745.

(36) Marcos, N. T.; Pinho, S.; Grandela, C.; Cruz, A.; Samyn-Petit, B.; Harduin-Lepers, A.; Almeida, R.; Silva, F.; Morais, V.; Costa, J.; et al. Role of the Human St6galnac-I and St6galnac-Ii in the Synthesis of the Cancer-Associated Sialyl-Tn Antigen. *Cancer Res* **2004**, *64* (19), 7050-7057. DOI: 10.1158/0008-5472.CAN-04-1921.

(37) Lloyd, K. O.; Burchell, J.; Kudryashov, V.; Yin, B. W.; Taylor-Papadimitriou, J. Comparison of O-Linked Carbohydrate Chains in Muc-1 Mucin from Normal Breast Epithelial Cell Lines and Breast Carcinoma Cell Lines. Demonstration of Simpler and Fewer Glycan Chains in Tumor Cells. *J Biol Chem* **1996**, *271* (52), 33325-33334. DOI: 10.1074/jbc.271.52.33325.

(38) Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R.; Heer, F. T.; de Beer, T. A P.; Rempfer, C.; Bordoli, L.; et al. Swiss-Model: Homology Modelling of Protein Structures and Complexes. *Nucleic Acids Research* **2018**, *46* (W1), W296-W303. DOI: 10.1093/nar/gky427 (acccessed 12/17/2024).

(39) Honorato, R. V.; Koukos, P. I.; Jiménez-García, B.; Tsaregorodtsev, A.; Verlato, M.; Giachetti, A.; Rosato, A.; Bonvin, A. M. J. J. Structural Biology in the Clouds: The Wenmr-Eosc Ecosystem. *Frontiers in Molecular Biosciences* **2021**, *8*, Technology and Code. DOI: 10.3389/fmolb.2021.729513.

(40) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; MacKerell, A. D. Charmm36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nature Methods* **2017**, *14* (1), 71-73. DOI: 10.1038/nmeth.4067.

(41) Field, M. J.; Bash, P. A.; Karplus, M. A Combined Quantum Mechanical and Molecular Mechanical Potential for Molecular Dynamics Simulations. *J. Comput. Chem.* **1990**, *11* (6), 700-733.

(42) Crous, W.; Field, M. J.; Naidoo, K. J. Simple Link Atom Saccharide Hybrid (Slash) Treatment for Glycosidic Bonds at the Qm/Mm Boundary. *Journal of Chemical Theory and Computation* **2014**, *10* (4), 1727-1738. DOI: 10.1021/ct400903n (acccessed 2014/05/19).