nature chemical biology

ARTICLE

PUBLISHED ONLINE: 18 DECEMBER 2017 | DOI: 10.1038/NCHEMBIO.2539

Expression system for structural and functional studies of human glycosylation enzymes

Kelley W Moremen[1]*, Annapoorani Ramiah[1], Melissa Stuart[2], Jason Steel[3], Lu Meng[1], Farhad Forouhar[4], Heather A Moniz[1], Gagandeep Gahlay[2], Zhongwei Gao[1], Digantkumar Chapla[1], Shuo Wang[1], Jeong-Yeh Yang[1], Pradeep Kumar Prabhakar[1], Roy Johnson[1], Mitche dela Rosa[1], Christoph Geisler[2], Alison V Nairn[1], Jayaraman Seetharaman[4], Sheng-Cheng Wu[1], Liang Tong[4], Harry J Gilbert[1,5], Joshua LaBaer[3] & Donald L Jarvis[2]*

Vertebrate glycoproteins and glycolipids are synthesized in complex biosynthetic pathways localized predominantly within membrane compartments of the secretory pathway. The enzymes that catalyze these reactions are exquisitely specific, yet few have been extensively characterized because of challenges associated with their recombinant expression as functional products. We used a modular approach to create an expression vector library encoding all known human glycosyltransferases, glycoside hydrolases, and sulfotransferases, as well as other glycan-modifying enzymes. We then expressed the enzymes as secreted catalytic domain fusion proteins in mammalian and insect cell hosts, purified and characterized a subset of the enzymes, and determined the structure of one enzyme, the sialyltransferase ST6GalNAcII. Many enzymes were produced at high yields and at similar levels in both hosts, but individual protein expression levels varied widely. This expression vector library will be a transformative resource for recombinant enzyme production, broadly enabling structure–function studies and expanding applications of these enzymes in glycochemistry and glycobiology.

Cell-surface and secreted glycoproteins and glycolipids contribute to numerous interactions with the extracellular environment that influence cellular physiology and pathology[1,2]. These interactions are strikingly diverse and include molecular recognition events that drive cell-surface signaling, cell adhesion, and modulation of receptor function, among others[2]. Terminal glycan structures also direct cellular targeting and clearance of circulating glycoproteins, including recombinant therapeutics, and define many pathogen and toxin tropisms[2].

Numerous challenges remain in the study of glycan structures and their contributions to physiological processes. A recent study[3] concluded that transformative methods are urgently required to fill the current gaps in technology for carbohydrate-based applications in health, energy, and material sciences. The roadmap goals for technology development include the enzymes involved in glycan synthesis, modification, and catabolism. These enzymatic reagents were emphasized not only because they can advance our understanding of fundamental glycan biosynthetic and catabolic processes[3], but also for their utility in chemoenzymatic glycan synthesis, which has numerous applications in biological and biomedical studies, exemplified by the production of standards for glycomic analysis, for installation and manipulation of glycan structures in complex biological systems, and for development of glycan-related therapeutics, among others.

Glycan biosynthesis occurs predominantly within the secretory pathway[1], wherein glycosyltransferases (GTs) use sugar donor precursors to extend oligosaccharide chains and glycoside hydrolases (GHs) cleave glycan structures during oligosaccharide maturation. Several factors contribute to the diverse glycan products of these complex enzymatic reactions. Most notably, glycan elaboration is controlled by the availability, the abundance, and the specificities of the enzymes ('glycoenzymes') involved in glycan synthesis and catabolism[4]. Unlike RNA transcription and protein translation, these complex metabolic pathways are not template driven. Thus, the key to understanding glycan diversity and function lies in deciphering glycoenzyme specificities, structure–function relationships, and processes regulating their activities.

In mammals, ~700 enzymes and proteins contribute to glycan extension, modification, recognition, and catabolism[4] to generate the full collection of ≥7,000 vertebrate oligosaccharide structures[5]. These enzymes include ~200 GTs and ~80 GHs (**Supplementary Data Sets 1** and **2**) classified into 44 and 28 discrete sequence families, respectively, in the carbohydrate-active enzymes (CAZy) database[6]. Members of a given GT family generally have a common protein fold, but frequently have distinct acceptor substrate profiles and may have different sugar donor specificities[7,8]. Thus, subtle structural differences between enzymes that display similar protein folds can produce glycoenzymes with unique catalytic specificities. These differences in substrate recognition among the GTs expressed in a given cell help determine the diversity of glycan structures in animal systems[4].

Many challenges remain in advancing our understanding of vertebrate glycosylation machinery at a molecular level. Only a minor subset of the members in a given GT or GH family have been functionally characterized[6], and structural studies have been performed on an even smaller subset of these enzymes. For mammalian glycoenzymes, these observations reflect challenges associated with their recombinant expression, as evidenced by the deposition of

[1]Complex Carbohydrate Research Center, University of Georgia, Athens, Georgia, USA. [2]Department of Molecular Biology, University of Wyoming, Laramie, Wyoming, USA. [3]Biodesign Institute, Arizona State University, Tempe, Arizona, USA. [4]Department of Biological Sciences, Northeast Structural Genomics Consortium, Columbia University, New York, New York, USA. [5]Institute for Cell and Molecular Biosciences, Newcastle University, Newcastle upon Tyne, UK. *e-mail: moremen@uga.edu or DLJarvis@uwyo.edu

only 28 of 202 vertebrate GT structures in the Protein Data Bank[6,8]. A few mammalian GTs have been expressed in bacteria either as soluble truncated catalytic domains[9,10] or with *in vitro* refolding[11], but bacterial expression more commonly yields nonfunctional protein aggregates, as these enzymes require glycosylation, disulfide bonding, and chaperones that occur uniquely within the eukaryotic secretory pathway[12]. The most successful approach for recombinant glycoenzyme expression has been production in eukaryotic hosts, but this has been restricted to small-scale expression studies for relatively focused biochemical goals[8]. Broad access to large quantities of mammalian glycoenzymes would be greatly enhanced by the development of a unified, modular, high-throughput, and scalable expression platform.

We describe here an expression vector library encoding all known human glycoenzymes, comprising GTs, GHs, and sulfotransferases (STs), for production in either mammalian (HEK293) cells[13] or baculovirus-infected insect cells[14]. The strategy involved insertion of glycoenzyme coding cassettes into custom-designed plasmid or baculovirus expression vectors for recombinant production in mammalian or insect cells. We demonstrate successful expression and secretion of a large set of the glycoenzymes in mammalian cells and the baculovirus-insect cell system (65% secreted at ≥10 mg/L), generating quantities of the recombinant products sufficient for enzymatic and structural characterization and for glycan synthetic applications. We also demonstrate the purification of a subset of the GTs as a proof of concept, as well as complete enzymatic characterization and structural determination of the sialyltransferase ST6GalNAcII in complex with the donor analog, CMP. Thus, we demonstrate that the expression vector library and resulting recombinant products provide a substantial resource enabling detailed studies on these enzymes and a source of novel enzymatic tools for a wide array of biochemical and biomedical applications.

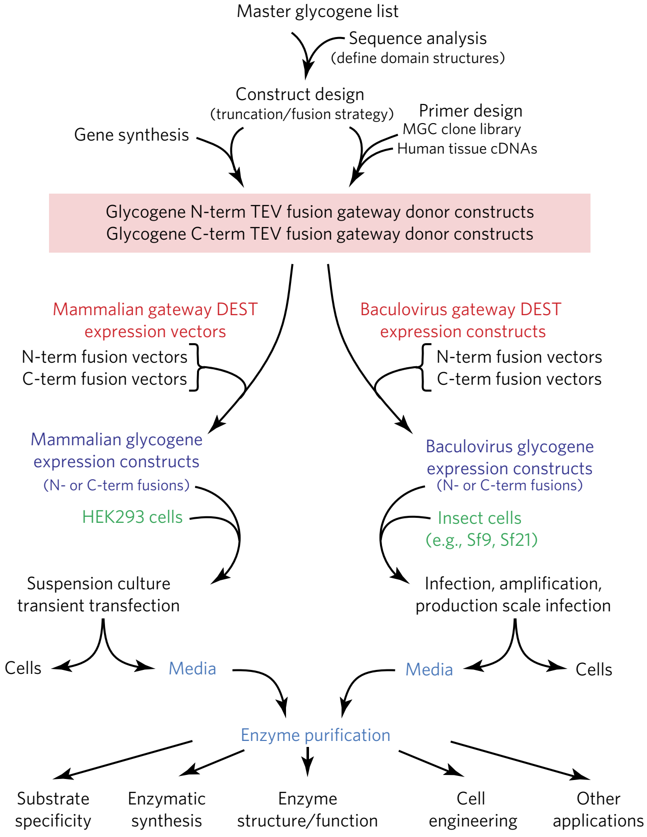

Master glycogene list
Sequence analysis
(define domain structures)
Construct design
(truncation/fusion strategy)
Gene synthesis
Primer design
MGC clone library
Human tissue cDNAs
Glycogene N-term TEV fusion gateway donor constructs
Glycogene C-term TEV fusion gateway donor constructs
Mammalian gateway DEST expression vectors
N-term fusion vectors
C-term fusion vectors
Baculovirus gateway DEST expression constructs
N-term fusion vectors
C-term fusion vectors
Mammalian glycogene expression constructs
(N- or C-term fusions)
Baculovirus glycogene expression constructs
(N- or C-term fusions)
HEK293 cells
Insect cells
(e.g., Sf9, Sf21)
Suspension culture transient transfection
Infection, amplification, production scale infection
Cells
Media
Media
Cells
Enzyme purification
Substrate specificity
Enzymatic synthesis
Enzyme structure/function
Cell engineering
Other applications

Figure 1 | Flow chart for generation of glycoenzyme expression constructs. General workflow for generating the human glycoenzyme expression constructs is depicted using a master glycogene list (**Supplementary Data Set 1**) and truncation designs based on TMD sequences, cleavable signal sequences, and available knowledge regarding the location of the enzyme active sites. Coding regions were incorporated into Gateway 'donor' vectors by PCR from cDNA sources or gene synthesis, and a TEV protease recognition site was appended on either the NH_2- or COOH-terminal end of the coding region (**Supplementary Fig. 1**). Thus, a library of all human glycoenzyme constructs ('glycogenes') in donor vectors was generated for Gateway LR recombination into custom mammalian or baculovirus DEST vectors harboring additional in-frame fusion tags (**Supplementary Figs. 2** and **3**) and employed to produce glycogene expression constructs for each recombinant host system. Expression and secretion of recombinant enzymes were examined by batch-mode transient transfection (HEK293 cells) or baculovirus infection (insect cells), and secreted products were used to produce enzyme preparations with utility for enzymology, structural studies, enzymatic glycan synthesis and numerous other applications.

RESULTS

Strategies for fusion protein expression

A comprehensive list of 339 genes comprising all human GTs, GHs, and STs, as well as numerous other glycoenzymes was targeted for protein expression (**Supplementary Table 1** and **Supplementary Data Set 1**). This list included enzymes containing NH_2-terminal or COOH-terminal transmembrane domains (TMDs), multipass enzymes with internal TMDs, enzymes with NH_2-terminal signal sequences, and cytosolic enzymes with no TMD or signal peptide. A modular design strategy was used to transfer full-length or truncated enzyme (catalytic domain) coding regions into custom, host-specific expression vectors by Gateway recombination[15] (**Fig. 1** and **Supplementary Figs. 1**–**3**). NH_2- or COOH-terminal transmembrane segments, NH_2-terminal signal sequences, or COOH-terminal endoplasmic reticulum (ER) retention sequences[16] were generally deleted and replaced by a tobacco etch virus (TEV) protease recognition and cleavage site sequence[17]. The resulting coding regions were then flanked with Gateway recombination site sequences and transferred into a plasmid vector backbone to form a library of Gateway glycoenzyme 'donor' clones (**Fig. 1** and **Supplementary Fig. 1**).

Custom expression vectors included host-specific promoter elements, in-frame fusion tags, and complementary Gateway recombination sites (Gateway destination "DEST" vectors; **Fig. 1** and **Supplementary Figs. 2** and **3**). Recombination of the Gateway donor clones harboring target gene coding regions with host-specific DEST vectors yielded the final expression vector construct library, containing host-specific transcription and translation elements that drive the recombinant production of target proteins with vector-encoded fusion tags. For example, mammalian DEST vectors included a human cytomegalovirus (CMV) promoter; fusion tag sequences; Gateway recombination and selection sequences; and artificial intron, enhancer, termination, and polyadenylation sequences. Upon Gateway recombination with the donor clones, the Gateway recombination and selection cassette was replaced with the glycoenzyme coding sequence, fusing it in-frame with the vector-encoded tag sequences (**Supplementary Figs. 2** and **3**). Three vectors appended different combinations of NH_2-terminal fusion tags (pGEn1 contained a signal sequence[18], an 8×His tag, and a StrepII tag[19]; pGEn2 contained a signal sequence, an 8×His tag, an AviTag[20], and a GFP domain[21]; pGEn3 contained a signal sequence, an 8×His tag, a GFP domain and an IgFc domain[22]; **Supplementary Fig. 2**), and two other vectors appended different combinations of COOH-terminal fusion tags (pGEc1 contained an 8×His tag and a StrepII tag; pGEc2 contained a GFP domain, an AviTag, and a 8×His tag; **Supplementary Fig. 2**). Thus, various expression vectors enabled

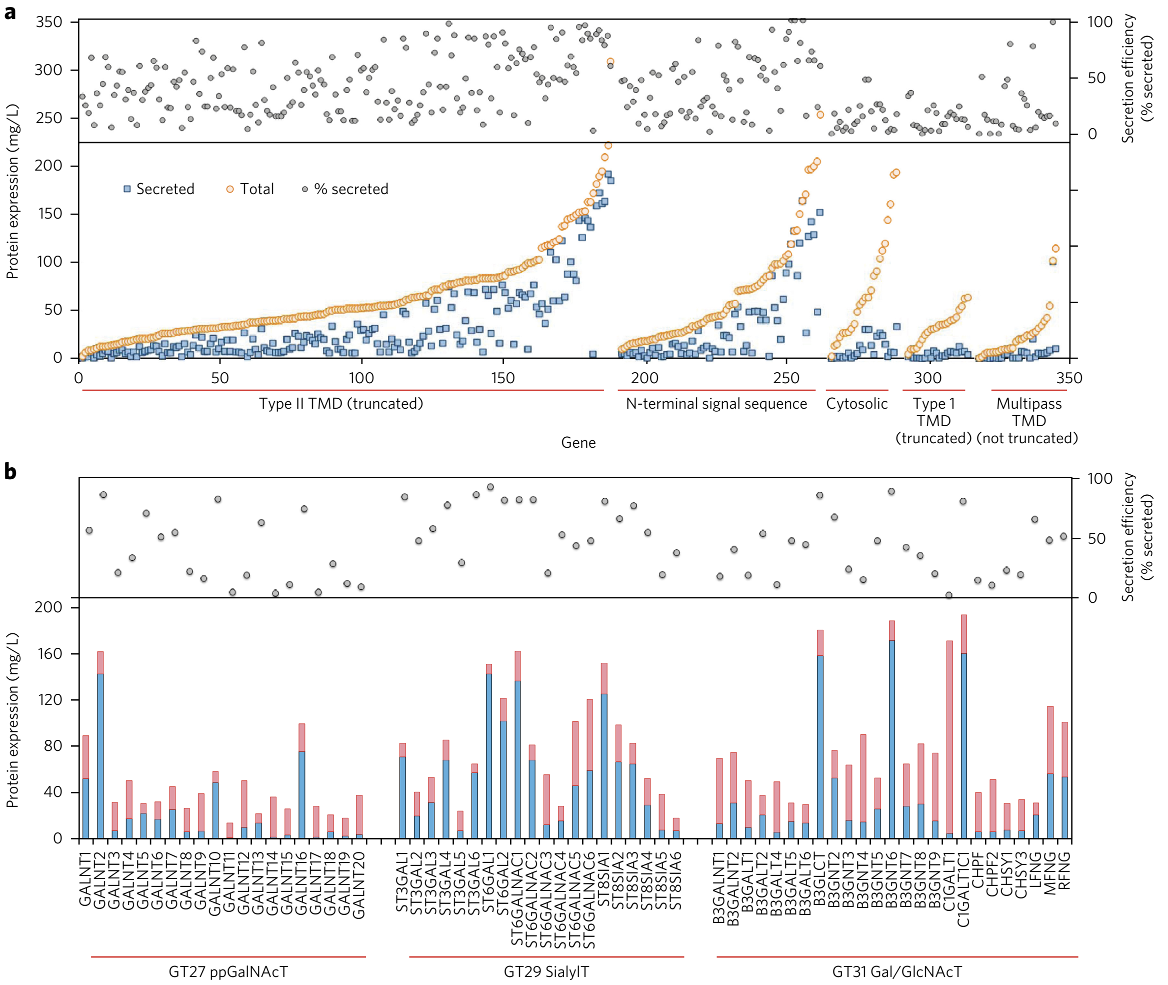


Figure 2 | Expression and secretion of the glycoenzyme constructs in mammalian cells. (**a**) The full collection of 339 human glycoenzymes as NH_2- or COOH-terminal GFP fusion proteins was expressed in HEK293 cells, and fusion protein production and secretion was profiled. Protein-coding regions were clustered by transmembrane topology and truncation strategy, including enzymes harboring NH_2-terminal TMDs, which were truncated to delete the transmembrane sequences and replaced with an NH_2-terminal fusion tag (type II TMD (truncated)); those harboring an NH_2-terminal signal sequence (N-terminal signal sequence), which were designed for protein fusions as described in Online Methods; proteins containing COOH-terminal TMDs (type I TMD (truncated)), which were truncated to delete the TMD and replaced with a COOH-terminal fusion tag; and cytosolic and multipass TMD enzymes, which were expressed as full-length coding regions containing COOH-terminal fusions. Total GFP fluorescence (cells + media) and cell-free media fluorescence values were converted into milligram quantities of fusion protein per liter and used to assess the efficiency of fusion protein secretion (% secretion). The order of the genes in each clustered transmembrane topology category was based on the rank order of overall (cell + media) expression (yellow circles). The corresponding quantity of secreted fusion protein for each coding region is indicated as blue squares. For each protein coding region, the efficiency of fusion protein secretion (% secretion) was indicated in the upper panel (gray circles). (**b**) For examination of the relative expression and secretion levels within individual GT sequence families, the values of secreted (blue bars) and cell-associated (pink bars) fusion proteins for individual polypeptide GalNAc transferases (GT27 ppGalNAcT), Sialyltransferases (GT29 SialylT), and galactosyltransferases/*N*-acetylglucosaminyltransferases (GT31 Gal/GlcNAcT) are displayed as stacked bar graphs with the respective gene name listed below each bar. Upper panel shows the efficiency of fusion protein secretion (percent secretion) for each coding region (gray circles). Fluorescence values for the secreted sialyltransferases were ± 10% of the indicated values in cases where replicate transfection experiments were performed (*n* =3).

transfer of each glycoenzyme coding region into multiple NH_2- or COOH-terminal fusion protein contexts.

We also generated custom Gateway-adapted baculovirus expression vectors (BEVs) with polyhedrin promoters that supported fusion strategies analogous to the mammalian expression plasmids (**Supplementary Fig. 3**). Thus, two BEVS supported NH_2-terminal fusion strategies similar to those of the pGEn1 and pGEn2 vectors, and an additional BEV provided a COOH-terminal fusion strategy similar to the pGEc1 vector. The same glycoenzyme donor clones were used for recombination into both the mammalian and baculovirus expression constructs to produce a parallel set of vectors for expression of NH_2-terminal or COOH-terminal fusion proteins in the two recombinant hosts.

Target protein expression, secretion, and quantitation

Glycoenzyme production and secretion efficiency was initially tested in HEK293 cells using the NH_2-terminal fusion vectors (pGEn1, pGEn2, and pGEn3) and fourteen GT29 sialyltransferase coding regions (**Supplementary Fig. 4**). The pGEn2 and pGEn3 vectors harboring the GFP fusion domain resulted in secretion of fusion proteins at higher levels than that observed with pGEn1, which did not contain this domain. Thus, pGEn2 was chosen for further screening of NH_2-terminal GFP fusion proteins, and pGEc2 was used for the COOH-terminal GFP fusions. Because generation of most of the BEV stocks was underway well before we determined that the GFP fusion cassette was the preferable strategy for the mammalian expression vectors, screening efforts using BEVs focused on the short NH_2-terminal fusion vector, analogous to pGEn1 (**Supplementary Fig. 5** and **Supplementary Data Set 1**). Protein expression was profiled by immunoblotting with antibodies that recognize the His-tag appended to the target enzymes.

The GFP fusion domain also facilitated protein quantitation during expression and purification (**Figs. 2** and **3**; **Supplementary Figs. 4** and **6**; **Supplementary Data Set 1**). Overall protein expression levels in mammalian cells were ≥10 mg/L for 94% of the recombinant products (**Fig. 2**), whereas 65% of the fusion proteins designed for secretion were secreted at levels of ≥10 mg/L. Surprisingly, there was no apparent correlation between expression levels and secretion efficiencies (**Fig. 2a**). To illustrate this variability, we examined twenty GT29 sialyltransferases that were designed using identical truncation and fusion strategies. Fluorescence measurements (**Fig. 2b**) and SDS–PAGE (**Supplementary Fig. 5**) indicated a subset of the enzymes (ST6GalI, ST3GalI, ST3GalIII, and ST3GalIV) that were exceptionally well expressed, as well as others that were well expressed but poorly secreted (ST3GalV), poorly expressed (ST3GalVI and ST6GalNAcIII), or expressed and secreted but subject to proteolysis (for example, ST6GalII and ST6GalNAcI; **Supplementary Fig. 5**). Similar variations in protein expression and secretion were also seen within other protein families (for example, GH27 and GT31 in **Fig. 2b** and **Supplementary Data Set 1**).

Comparison of insect and mammalian cell expression

A comparison of protein expression and secretion in mammalian and BEV-infected insect cells revealed similar trends (**Fig. 3a**). However, there was considerable scatter in the data, indicating that one of the two expression hosts was more effective for expression of some individual coding regions. As the initial expression screening employed different fusion protein strategies for the two host systems, we generated BEVs encoding twenty GT29 sialyltransferases with NH_2-terminal GFP fusions for direct comparison with equivalent mammalian expression constructs (**Fig. 3b** and **Supplementary Fig. 5b**). GFP fluorescence data demonstrated that protein-secretion efficiencies were quite variable in both hosts, but secreted protein levels were highly correlated for half of the fusion proteins across the two recombinant hosts ($R^2 = 0.99$; **Fig. 3b**). In contrast, eight of the sialyltransferase fusion proteins were more highly secreted in mammalian cells, and two were more highly secreted in the BEVs (**Fig. 3b**). Thus, the observations were similar to the expression data for the larger collection of the glycoenzymes (**Fig. 3a**): a subset of proteins was more effectively secreted in either human or insect cells, but most were secreted at similar levels in the two different expression hosts.

Purification of human glycoenzymes

To demonstrate purification of enzymes from both expression hosts, we used immobilized metal affinity chromatography (IMAC) to isolate the 8×His tagged sialyltransferase fusion proteins from mammalian and insect cell-free media. The results showed effective purification of even some of the more poorly expressed proteins after scale-up (**Supplementary Fig. 6**). In the mammalian host, a majority of the GT29 sialyltransferases were able to be purified at >10 mg/L of transfected cell culture, indicating that these enzymes could be produced and purified at levels required for biochemical and structural studies.

Structure determination of human ST6GalNAcII

To demonstrate the utility of the expression vector library for glycoenzyme characterization, we scaled up production of one of the sialyltransferase GFP fusions (ST6GalNAcII-pGEn2) in HEK293S ($GnTI^-$) cells[23] using a selenomethionine labeling protocol established in prior structural studies[24]. IMAC purification of the secreted fusion protein, *in vitro* cleavage with TEV protease and endoglycosidase F1 (EndoF1), and further

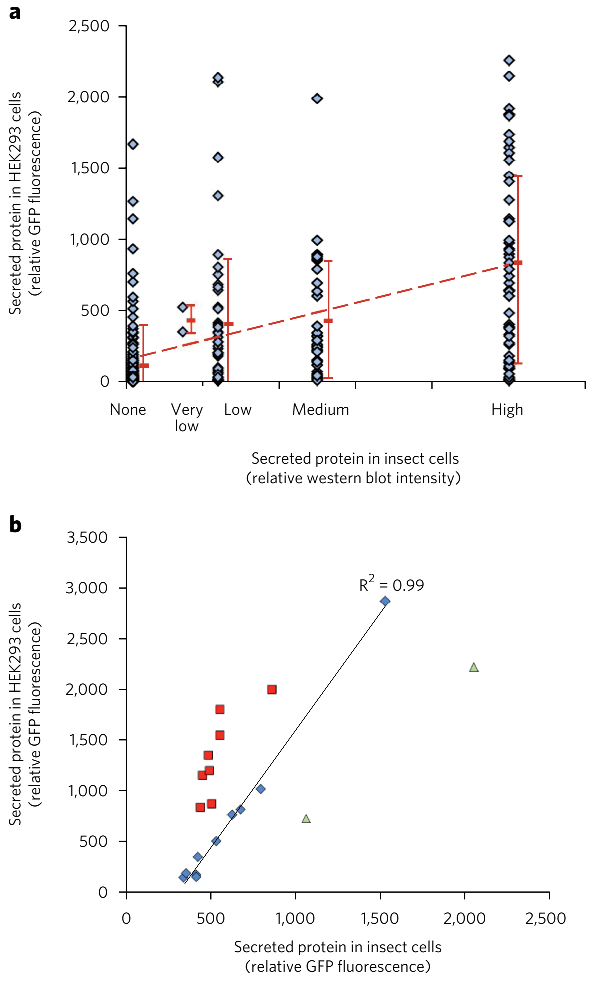


Figure 3 | Comparison of secreted enzyme expression in transfected mammalian cells and baculovirus-infected insect cells. (**a**) The collection of human glycoenzymes was expressed using mammalian expression vectors or BEVs in each respective host, and protein secretion levels were compared. For mammalian cell expression, the pGEn2 or pGEc2 GFP fusion vectors were employed, and secreted recombinant products were quantified by GFP fluorescence (*y* axis values for blue diamonds). For BEV expression, the shorter His- and StrepII-tagged constructs (Short N-term fusions; **Supplementary Fig. 3**) were employed, and the respective secreted fusion proteins were quantified using anti-His-tag immunoblots. The semiquantitative immunoblot intensities (**Supplementary Data Set 1**) were converted into arbitrary relative intensity values (*x* axis values for blue diamonds) and compared with the values for enzyme secretion in mammalian cells in a two-dimensional dot plot. The mean and s.d. for the GFP fluorescence values are indicated by the thick red bar and error bars. A trend line is shown for the mean values for GFP fluorescence (red dotted line). (**b**) As the fusion protein strategy was different in the mammalian and BEV constructs, a subset of the constructs comprising the twenty GT29 sialyltransferases was generated in the GFP N-terminal fusion baculovirus (**Supplementary Fig. 5**) for direct comparison with the equivalent pGEn2 mammalian GFP fusion constructs. Comparison of secreted GFP fluorescence values indicated a close correlation for a subset of ten sialyltransferases (blue diamonds; trend line indicated with $R^2 = 0.99$), whereas eight were more highly secreted in mammalian cells (red squares), and two were more highly secreted in BEVS (green triangles). Anti-His tag western blots of the equivalent samples for the Short N-term tag (no GFP tag) and the GFP N-term fusion baculovirus constructs are shown in **Supplementary Figure 5**. Fluorescence values for the secreted sialyltransferases were ± 10% of the indicated values in cases where replicate transfection experiments were performed ($n = 3$).

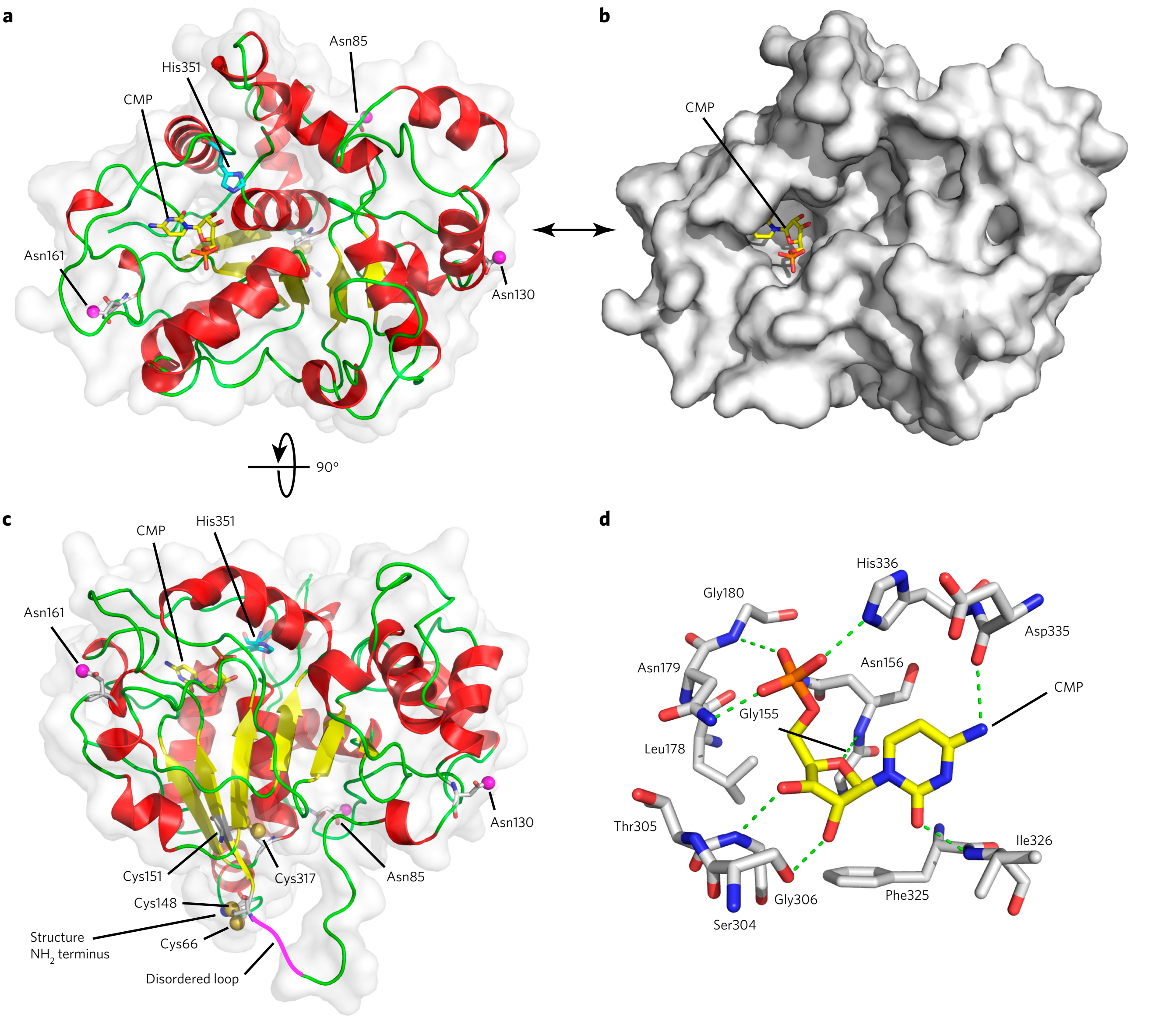


Figure 4 | Structure of human ST6GalNAcII. The human ST6GalNAcII structure was solved by X-ray diffraction, demonstrating a single Rossmann-like (GT-A variant 2) fold[24] with 6 β-strands in the domain core and 17 α-helical segments and loop regions (labeled in **Supplementary Fig. 9**). There are six protomers in the crystallographic asymmetric unit, and some of the units contain disordered loops of varying lengths. (**a**–**d**) Images represent chain C, which contains a 7-residue disordered loop between helix α5 and strand β1 (magenta line in **c** and **Supplementary Fig. 9**). (**a**) View from the top of the active site, where the bound donor analog, CMP, is shown in yellow stick representation. (**b**) Surface representation in the same orientation as in **a**. (**c**) Cartoon representation of the protein structure at a 90° rotation relative to that in **a**. The position of the His351 residue, the predicted catalytic base in the structure by comparison to structures of other GT29 sialyltransferases[24], is shown in cyan stick representation. Weak electron densities for monosaccharide residues were found to be extended from the amide side chains of Asn85, Asn130, and Asn161, as predicted for each residue found within an NxS glycosylation sequon (where x is any amino acid except Pro). The positions of the respective Asn side chains are indicated by white stick representations, and magenta spheres represent the amide nitrogens where the N-glycan is attached. Four Cys residues are found in the structure, with each participating in a disulfide bond. The NH_2 terminal residue in the structure, Cys66, is disulfide bonded to Cys148, while a disulfide bond links Cys151 with Cys317 (white stick representations with yellow sulfur atoms shown as spheres). (**d**) Extensive interactions were identified between the bound CMP donor analog and the binding site, including polar interactions (green dotted lines) and a hydrophobic stacking with Phe325.

purification yielded the ST6GalNAcII catalytic domain devoid of all tag sequences, with N-glycans trimmed to a single GlcNAc residue (**Supplementary Fig. 7**). Kinetic parameters for the TEV/EndoF1-cleaved ST6GalNAcII and the GFP fusion protein were similar (**Supplementary Table 2**), indicating that the GFP fusion did not alter enzyme activity. The structure of the catalytic domain was determined to 3.1 Å resolution by single-wavelength anomalous diffraction (SAD) phasing (**Supplementary Table 3**), and additional diffraction data were obtained to 2.35 Å for a complex of ST6GalNAcII with the donor analog, CMP (**Fig. 4** and **Supplementary Table 3**). The structure revealed a single Rossmann-like (GT-A variant 2) fold similar to those of other known GT29 sialyltransferases[9,24–26] with highly conserved 'sialylmotif' sequence elements[27,28] comprising the core of the protein fold (**Supplementary Figs. 8–10**). The bound CMP was clearly evident in the electron density at a position equivalent to other sialyltransferases (**Fig. 4** and **Supplementary Fig. 10**), confirming conservation of the sialylmotif scaffold underlying the

CMP-Neu5Ac binding site[24]. Substantial differences relative to other sialyltransferase structures were found outside the sialylmotif elements (**Supplementary Fig. 10**), reflecting the minimal primary sequence or structural similarity among the sialyltransferases in the loop regions and secondary structure elements involved in acceptor substrate recognition. While structures have been reported for three of the four vertebrate GT29 sialyltransferase subfamilies that generate Neu5Ac-α2,6Gal[24,25], Neu5Ac-α2,3Gal[9], and Neu5Ac-α2,8Neu5Ac[26] linkages, the present ST6GalNAcII structure is the first reported for a Neu5Ac-α2,6GalNAc subfamily member[24] (**Supplementary Fig. 8**). This proof of principal for the cloning, expression, biochemical analysis, and structure determination pipeline will be the basis of further studies to explore the structural basis of acceptor recognition for ST6GalNAcII.

DISCUSSION

Diversity in vertebrate cell surface and secreted glycan structures is generated in complex biosynthetic pathways[1,2,4]. The overall purpose of this study was to produce an expression vector library enabling the production of 339 enzymes involved in human glycan synthesis, modification, and catabolism for enzymatic, structural, and functional studies. We initially determined that *Escherichia coli* was not an appropriate host for these recombinant enzymes, and as a result we focused on expression in mammalian and insect cell hosts, which are amenable to high-throughput profiling of protein expression and production that can be scaled up to larger culture volumes[13,14].

A variety of fusion tag strategies were tested. Inclusion of the 'superfolder' GFP domain improved protein production and secretion for many GTs tested in both human and insect cell systems (**Supplementary Figs. 4** and **5**) compared to the shorter fusion tags. The GFP fusion also facilitated detection and quantitation of recombinant products during expression and purification. Thus, all glycoenzymes were initially profiled as GFP fusions in HEK293 cells.

Most of the glycoenzyme fusion constructs were highly expressed in HEK293 cells, and a major subset (~65%) was well secreted and able to be purified at multimilligram levels from the culture media. The efficiency of secretion was surprisingly variable and did not correlate with overall expression level, transmembrane topology, or enzyme family. This variability was highlighted for the GT29 sialyltransferases, but was also observed with other GT families. These data suggest that each protein coding region has a unique set of folding constraints within the eukaryotic secretory pathway that cannot be generalized more broadly, even within an enzyme structural family.

Parallel expression studies were performed using BEV-infected insect cells, and the trend in overall expression and secretion level was similar to that in transfected mammalian cells. When enzymes with equivalent tags were compared, half were secreted at comparable levels by the two hosts, whereas the remainder were more highly secreted by either mammalian or insect cells. These data suggest that host-specific factors can also contribute to secretion efficiency for some recombinant products.

Other factors can also contribute to the efficiency of protein production in eukaryotic cells. *In vivo,* many full-length GTs are transmembrane proteins that can assemble into homodimers or heterooligomeric complexes with other proteins[29]. Sequences proximal to the TMD can also play roles in oligomer formation[30]. Thus, removal of the proteins from their transmembrane contexts and the lack of co-expression with binding partners may contribute to difficulties in folding or secretion. Nevertheless, >65% of the GTs were efficiently expressed as single coding regions in the absence of their TMDs.

Overall, most of the glycoenzymes designed for secretion can be readily purified from the culture media at high yields using IMAC (**Supplementary Fig. 6**). In addition, kinetic analysis was performed on several of these purified recombinant products before and after tag removal with TEV protease. The results, exemplified by ST6GalNAcII (**Supplementary Table 2**), indicated that the GFP fusion proteins and catalytic domains released had comparable activity[24,31]. To illustrate the utility of the expression library, we characterized several of the GTs, revealing enzyme substrate[31–37] or chaperone[38] specificities, and corrected the glycan linkage synthesized by a GT[39]. We also used some products for chemoenzymatic synthesis of novel glycan structures[40–43] and selective exoenzymatic labeling (SEEL)[44–46] of cell-surface glycans. These efforts demonstrated the utility of our expression vector library for various glycoenzyme studies and for direct applications in glycobiology.

As a proof of concept for the utility of the HEK293 platform in protein structural studies, we focused on ST6GalNAcII, an enzyme that adds α2,6-linked Neu5Ac to the GalNAc residue on Core 1 O-glycans[47]. These studies established a unified and scalable workflow for production and purification of the glycoenzymes for both enzymatic and structural studies. As the expression vector library was being completed, we continued to expand the application of the workflow for structural studies on the extracellular domain of a calcium-sensing receptor with a novel bound ligand[48] and a plant xyloglucan fucosyltransferase[49]. We also continue to apply similar approaches to study additional human glycosyltransferase structures. Thus, it is clear the constructs, expression systems, and workflows for glycoenzyme production and purification described here will reveal new insights into substrate recognition and catalysis for this diverse and important set of glycoenzymes and provide transformative reagents that will continue to expand our knowledge in glycochemistry and glycobiology.

Received 19 May 2017; accepted 8 November 2017;
published online 18 December 2017

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

References

1. Moremen, K.W., Tiemeyer, M. & Nairn, A.V. Vertebrate protein glycosylation: diversity, synthesis and function. *Nat. Rev. Mol. Cell Biol.* **13**, 448–462 (2012).
2. Varki, A. Biological roles of glycans. *Glycobiology* **27**, 3–49 (2017).
3. National Research Council of the National Academies. *Transforming Glycoscience: A Roadmap for the Future* (National Academies Press, Washington DC, 2012).
4. Nairn, A.V. & Moremen, K.W. in *Handbook of Glycomics.* (eds. R. Cummings & J.M. Pierce) 95–136 (Academic Press, Burlington, MA, 2009).
5. Cummings, R.D. The repertoire of glycan determinants in the human glycome. *Mol. Biosyst.* **5**, 1087–1104 (2009).
6. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P.M. & Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* **42**, D490–D495 (2014).
7. Lairson, L.L., Henrissat, B., Davies, G.J. & Withers, S.G. Glycosyltransferases: structures, functions, and mechanisms. *Annu. Rev. Biochem.* **77**, 521–555 (2008).
8. Taniguchi, N. *et al. Handbook of Glycosyltransferases and Related Genes*, 2nd edn (Springer Tokyo, Japan, 2014).
9. Rao, F.V. *et al.* Structural insight into mammalian sialyltransferases. *Nat. Struct. Mol. Biol.* **16**, 1186–1188 (2009).
10. Seto, N.O., Palcic, M.M., Hindsgaul, O., Bundle, D.R. & Narang, S.A. Expression of a recombinant human glycosyltransferase from a synthetic gene and its utilization for synthesis of the human blood group B trisaccharide. *Eur. J. Biochem.* **234**, 323–328 (1995).
11. Ramakrishnan, B. & Qasba, P.K. Crystal structure of lactose synthase reveals a large conformational change in its catalytic component, the beta1,4-galactosyltransferase-I. *J. Mol. Biol.* **310**, 205–218 (2001).
12. Paulson, J.C. & Colley, K.J. Glycosyltransferases. Structure, localization, and control of cell type-specific glycosylation. *J. Biol. Chem.* **264**, 17615–17618 (1989).

13. Subedi, G.P., Johnson, R.W., Moniz, H.A., Moremen, K.W. & Barb, A. High yield expression of recombinant human proteins with the transient transfection of HEK293 cells in suspension. *J. Vis. Exp.* **106**, e53568 (2015).
14. Jarvis, D.L. Recombinant protein expression in baculovirus-infected insect cells. *Methods Enzymol.* **536**, 149–163 (2014).
15. Walhout, A.J. *et al.* GATEWAY recombinational cloning: application to the cloning of large numbers of open reading frames or ORFeomes. *Methods Enzymol.* **328**, 575–592 (2000).
16. Stornaiuolo, M. *et al.* KDEL and KKXX retrieval signals appended to the same reporter protein determine different trafficking between endoplasmic reticulum, intermediate compartment, and Golgi complex. *Mol. Biol. Cell* **14**, 889–902 (2003).
17. Carrington, J.C. & Dougherty, W.G. A viral cleavage site cassette: identification of amino acid sequences required for tobacco etch virus polyprotein processing. *Proc. Natl. Acad. Sci. USA* **85**, 3391–3395 (1988).
18. Vandersall-Nairn, A.S., Merkle, R.K., O'Brien, K., Oeltmann, T.N. & Moremen, K.W. Cloning, expression, purification, and characterization of the acid alpha-mannosidase from *Trypanosoma cruzi*. *Glycobiology* **8**, 1183–1194 (1998).
19. Schmidt, T.G. & Skerra, A. The Strep-tag system for one-step purification and high-affinity detection or capturing of proteins. *Nat. Protoc.* **2**, 1528–1535 (2007).
20. Beckett, D., Kovaleva, E. & Schatz, P.J. A minimal peptide substrate in biotin holoenzyme synthetase-catalyzed biotinylation. *Protein Sci.* **8**, 921–929 (1999).
21. Pédelacq, J.D., Cabantous, S., Tran, T., Terwilliger, T.C. & Waldo, G.S. Engineering and characterization of a superfolder green fluorescent protein. *Nat. Biotechnol.* **24**, 79–88 (2006).
22. Barb, A.W. *et al.* NMR characterization of immunoglobulin G Fc glycan motion on enzymatic sialylation. *Biochemistry* **51**, 4618–4626 (2012).
23. Reeves, P.J., Callewaert, N., Contreras, R. & Khorana, H.G. Structure and function in rhodopsin: high-level expression of rhodopsin with restricted and homogeneous N-glycosylation by a tetracycline-inducible N-acetylglucosaminyltransferase I-negative HEK293S stable mammalian cell line. *Proc. Natl. Acad. Sci. USA* **99**, 13419–13424 (2002).
24. Meng, L. *et al.* Enzymatic basis for N-glycan sialylation: structure of rat α2,6-sialyltransferase (ST6GAL1) reveals conserved and unique features for glycan sialylation. *J. Biol. Chem.* **288**, 34680–34698 (2013).
25. Kuhn, B. *et al.* The structure of human α-2,6-sialyltransferase reveals the binding mode of complex glycans. *Acta Crystallogr. D Biol. Crystallogr.* **69**, 1826–1838 (2013).
26. Volkers, G. *et al.* Structure of human ST8SiaIII sialyltransferase provides insight into cell-surface polysialylation. *Nat. Struct. Mol. Biol.* **22**, 627–635 (2015).
27. Datta, A.K. & Paulson, J.C. The sialyltransferase "sialylmotif" participates in binding the donor substrate CMP-NeuAc. *J. Biol. Chem.* **270**, 1497–1500 (1995).
28. Livingston, B.D. & Paulson, J.C. Polymerase chain reaction cloning of a developmentally regulated member of the sialyltransferase gene family. *J. Biol. Chem.* **268**, 11504–11507 (1993).
29. Kellokumpu, S., Hassinen, A. & Glumoff, T. Glycosyltransferase complexes in eukaryotes: long-known, prevalent but still unrecognized. *Cell. Mol. Life Sci.* **73**, 305–325 (2016).
30. Gleeson, P.A. Targeting of proteins to the Golgi apparatus. *Histochem. Cell Biol.* **109**, 517–532 (1998).
31. Revoredo, L. *et al.* 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* **26**, 360–376 (2016).
32. Halmo, S.M. *et al.* Protein O-linked mannose β-1,4-N-acetylglucosaminyl-transferase 2 (POMGNT2) Is a gatekeeper enzyme for functional glycosylation of α-dystroglycan. *J. Biol. Chem.* **292**, 2101–2109 (2017).
33. Praissman, J.L. *et al.* The functional O-mannose glycan on α-dystroglycan contains a phospho-ribitol primed for matriglycan addition. *eLife* **5**, e14473 (2016).
34. Gao, Y. *et al.* Acceptor specificities and selective inhibition of recombinant human Gal- and GlcNAc-transferases that synthesize core structures 1, 2, 3 and 4 of O-glycans. *Biochim. Biophys. Acta* **1830**, 4274–4281 (2013).
35. Gerken, T.A. *et al.* The lectin domain of the polypeptide GalNAc transferase family of glycosyltransferases (ppGalNAc Ts) acts as a switch directing glycopeptide substrate glycosylation in an N- or C-terminal direction, further controlling mucin type O-glycosylation. *J. Biol. Chem.* **288**, 19900–19914 (2013).
36. Sheikh, M.O. *et al.* Rapid screening of sugar-nucleotide donor specificities of putative glycosyltransferases. *Glycobiology* **27**, 206–212 (2017).
37. Urbanowicz, B.R., Peña, M.J., Moniz, H.A., Moremen, K.W. & York, W.S. Two *Arabidopsis* proteins synthesize acetylated xylan *in vitro*. *Plant J.* **80**, 197–206 (2014).
38. Hanes, M.S., Moremen, K.W. & Cummings, R.D. Biochemical characterization of functional domains of the chaperone Cosmc. *PLoS One* **12**, e0180242 (2017).
39. Praissman, J.L. *et al.* B4GAT1 is the priming enzyme for the LARGE-dependent functional glycosylation of α-dystroglycan. *eLife* **3**, e03943 (2014).
40. Li, T. *et al.* Divergent chemoenzymatic synthesis of asymmetrical-core-fucosylated and core-unmodified N-glycans. *Chemistry* **22**, 18742–18746 (2016).
41. Bello, C., Wang, S., Meng, L., Moremen, K.W. & Becker, C.F. A PEGylated photocleavable auxiliary mediates the sequential enzymatic glycosylation and native chemical ligation of peptides. *Angew. Chem. Int. Edn Engl.* **54**, 7711–7715 (2015).
42. Prudden, A.R. *et al.* Synthesis of asymmetrical multiantennary human milk oligosaccharides. *Proc. Natl. Acad. Sci. USA* **114**, 6954–6959 (2017).
43. Epp, A. *et al.* Sialylation of IgG antibodies inhibits IgG-mediated allergic reactions. *J. Allergy Clin. Immunol.* DOI:10.1016/j.jaci.2017.06.021
44. Mbua, N.E. *et al.* Selective exo-enzymatic labeling of N-glycans on the surface of living cells by recombinant ST6Gal I. *Angew. Chem. Int. Edn Engl.* **52**, 13012–13015 (2013).
45. Sun, T. *et al.* One-step selective exoenzymatic labeling (SEEL) strategy for the biotinylation and identification of glycoproteins of living cells. *J. Am. Chem. Soc.* **138**, 11575–11582 (2016).
46. Yu, S.H. *et al.* Selective exo-enzymatic labeling detects increased cell surface sialoglycoprotein expression upon megakaryocytic differentiation. *J. Biol. Chem.* **291**, 3982–3989 (2016).
47. Marcos, N.T. *et al.* Role of the human ST6GalNAc-I and ST6GalNAc-II in the synthesis of the cancer-associated sialyl-Tn antigen. *Cancer Res.* **64**, 7050–7057 (2004).
48. Zhang, C. *et al.* Structural basis for regulation of human calcium-sensing receptor by magnesium ions and an unexpected tryptophan derivative co-agonist. *Sci. Adv.* **2**, e1600241 (2016).
49. Urbanowicz, B.R. *et al.* Structural, mutagenic and in silico studies of xyloglucan fucosylation in *Arabidopsis* thaliana suggest a water-mediated mechanism. *Plant J.* **91**, 931–949 (2017).

Acknowledgments

We wish to thank F. Samli, R. Collins, L. Stanton, A. Petrey, A. Yox, R. Kim and J. Aumiller for technical assistance during these studies. This research was supported by NIH grants P41GM103390 (to K.W.M.), P01GM107012 (G.J. Boons, PI), P41GM103490 (J.M. Pierce, PI) and U54GM094597 for work performed in part as a community nominated project of the Protein Structure Initiative of the National Institutes of Health (to G.T. Montelione and L.T.).

Author contributions

K.W.M., D.L.J. and J.L. formulated the project; K.W.M. designed glycoenzyme truncation and fusion constructs and supervised mammalian cell expression efforts; D.L.J. supervised preparation of baculovirus constructs and insect cell expression efforts; A.V.N. generated all target mammalian glycogene lists including gene and protein annotations; J.L. and J. Steel designed primers and executed high-throughput glycogene amplification and Gateway recombination; A.R. and M.d.R. generated human cDNAs and mammalian expression vectors, and performed Gateway recombination into mammalian expression vectors; C.G. and G.G. generated all baculovirus DEST expression vectors, M.S. and G.G. performed Gateway recombination into baculovirus expression vectors, screened and amplified viral stocks, and characterized recombinant protein expression in insect cells, H.A.M., Z.G., D.C., S.W., J.-Y.Y., L.M., P.K.P., and R.J. characterized expression of glycoenzymes in mammalian cells; C.G. designed baculovirus DEST expression vectors; S.-C.W. and H.J.G. designed and generated fusion protein constructs for expression in bacteria. L.M. expressed and purified recombinant ST6GalNAcII for structural studies. F.F., J. Seetharaman, and L.T. performed structural studies on ST6GalNAcII.

Competing financial interests

The authors declare competing financial interests: details accompany the online version of the paper.

Additional information

Any supplementary information, chemical compound information and source data are available in the online version of the paper. Reprints and permissions information is available online at http://www.nature.com/reprints/index.html. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to K.W.M. or D.L.J.

ONLINE METHODS

Choice of protein-coding regions for recombinant expression. A comprehensive list of human glycoenzymes was identified in prior efforts to profile glycogene transcription in mouse cells[4]. Starting with this comprehensive list of >700 glycogenes, we targeted 339 coding regions to generate expression constructs comprising all known human GTs involved in glycan extension, GHs involved in glycan catabolism and processing, STs involved in glycan modification, as well as a collection of additional genes involved in glycan elaboration, modification, or catabolism (**Supplementary Data Set 1**).

Truncation and fusion protein strategies for glycoenzyme expression. Each glycoenzyme coding region on the target list was examined for sequence features in the UniProt database[50] to determine the presence of TMDs, NH_2-terminal signal sequences, COOH-terminal KDEL ER retention sequences[16], or known catalytic residues. In general, proteins containing NH_2-terminal TMDs (type II transmembrane proteins) were designed to truncate all sequences spanning from the initiating Met residue to the first charged residues at the luminal boundary of the TMD. These proteins were all designed as NH_2-terminal fusions to a TEV protease recognition peptide sequence[17] during transfer to Gateway donor vectors (**Supplementary Fig. 1** and **Supplementary Data Sets 1** and **2**). Proteins containing COOH-terminal TMDs were designed to truncate the TMD between the first charged residue on the luminal side of the TMD and the COOH-terminus of the protein (**Supplementary Fig. 1**). A TEV protease recognition sequence was appended to the COOH-terminus of the coding region during transfer to the Gateway donor vector (**Supplementary Fig. 1**). Proteins containing NH_2-terminal signal sequences were generally designed as full-length coding regions with a COOH-terminal TEV site extension added in place of the termination codon during transfer to the Gateway donor vectors. Many of these latter enzymes were also designed to truncate the NH_2-terminal signal sequence with a subsequent NH_2-terminal fusion with the TEV recognition site sequence (**Supplementary Data Set 1**). Proteins containing multipass TMDs, internal TMDs, or soluble cytosolic proteins were generally designed as full-length coding regions and had a COOH-terminal TEV extension added in place of the termination codon during transfer to the Gateway donor vectors (**Supplementary Data Set 1**). In some of these latter cases NH_2-terminal TEV fusions were also generated to test alternative tagging strategies for protein expression.

Isolation of protein coding regions and capture as Gateway donor clones. Three strategies were employed to isolate full length or truncated protein coding regions with appropriate TEV protease cleavage site fusions and flanking *att*L recombination sites in Gateway donor vectors. First, full-length human cDNA clones were identified within the Mammalian Gene Collection (MGC)[51], and plasmid DNAs encoding each respective coding region were used as templates for gene-specific PCR amplification. Gene-specific primer sequences were synthesized (Eurofins MWG Operon, Louisville, KY) as shown in **Supplementary Figure 1**, with 5′ extensions to initiate the synthesis of the *att*B recombination and TEV protease recognition sites based on the design for truncation and fusion. PCR amplification products were generated using 0.1 μM gene-specific primers, 100 ng plasmid DNA, and AccuPrime Pfx DNA Polymerase (Thermo Fisher Scientific) in a total volume of 50 μl. PCR conditions were as follows: denaturation at 94 °C for 2 min followed by 15 PCR cycles comprised of 94 °C denaturation for 15 s, annealing at 57 °C for 1 min and extension at 68 °C for 2 min. After the first round of amplification, 20 μl of the gene-specific PCR reaction was then subjected to a second round of amplification using a pair of universal primers and identical PCR reaction conditions as those described above to complete the synthesis of the *att*B sites as indicated in **Supplementary Figure 1**. PCR products were then used directly for Gateway recombination with the pDONR221 vector in BP Clonase II reactions with a final volume of 10 μl, as described by the manufacturer (Thermo Fisher Scientific). The reaction products were transformed into the 5-alpha competent *E. coli* strain (New England BioLabs, Ipswich, MA), and plated on LB plates containing kanamycin. The resulting plasmid clones were screened by restriction mapping and verified by DNA sequencing of the entire coding region.

For genes that were not available in MGC or were not successfully amplified from MGC clone templates, an alternative amplification approach was employed. A library of human RNAs (FirstChoice Human RNA Survey Panel, Thermo Fisher Scientific) was reverse transcribed using 500 ng RNA, a mixture of random hexamers and oligo $(dT)_{20}$ primers and a reaction mix supplied with the SuperScript II First Strand Synthesis System, as described by the manufacturer (Thermo Fisher Scientific). Test amplifications were performed using respective gene-specific primers, the human cDNAs, and Phusion DNA polymerase reaction mix (Thermo Fisher Scientific) as described above. Annealing temperatures, extension times, primer concentrations, and cycle numbers were varied to identify optimal conditions for PCR amplification of the predicted coding region. Once optimal primary amplification conditions were identified, the resulting PCR products were purified using a QiaQuick PCR purification kit (Qiagen) and subjected to a second round of amplification as described above using the universal primers shown in **Supplementary Figure 1**. Second round amplification products were isolated using the QiaQuick PCR purification kit, subjected to Gateway BP Clonase recombination into the pDONR221 vector, and transformed into 5-alpha competent *E. coli* as described above. The resulting plasmid clones were verified by DNA sequencing of the entire coding region.

In instances where amplification of coding regions from MGC or cDNA sources was unsuccessful, or when coding regions were poorly expressed, the coding regions were generated by gene synthesis. Coding region designs containing TEV fusion sequences and flanking *att*L1 and *att*L2 recombination sites (**Supplementary Data Sets 1** and **2**) were synthesized with human codon optimization (GeneArt gene synthesis, Thermo Fisher Scientific) and subcloned into a vector backbone containing a kanamycin resistance marker (pMK-RQ). The gene synthesis products in this vector backbone were equivalent to the pDONR221 constructs and were used directly for recombination into Gateway DEST expression vectors.

Generation of mammalian expression vector backbones as Gateway DEST vectors. Five custom Gateway fusion vectors were generated for mammalian cell expression using a pGEn2 vector backbone originally employed for the expression of rat ST6GalI[24]. This latter vector was generated by gene synthesis (Thermo Fisher Scientific) and contains a CMV promoter, an artificial intron, a woodchuck hepatitis virus post-transcriptional regulatory element (WPRE), and a bovine growth hormone (BGH) termination and polyadenylation sequence to drive transcription and termination of the fusion protein coding region. Each of the mammalian Gateway expression vectors was adapted for Gateway recombination as DEST vectors by inclusion of *att*R sites flanking a selection cassette comprised of *ccd*B and chloramphenicol resistance (Cm^R) genes[15] (**Supplementary Fig. 2**). The fusion sequences in each of the five vectors were distinct, employing either NH_2-terminal (pGEn1-DEST, pGEn2-DEST, and pGEn3-DEST vectors) or COOH-terminal (pGEc1-DEST and pGEc2-DEST vectors) fusion sequences adjacent to the selection cassette. Each vector component was generated by gene synthesis (Thermo Fisher Scientific) and swapped into the original synthetic ST6GAL1-pGEn2 vector as restriction fragments to replace the ST6GAL1 and adjoining fusion sequences. The pGEn1-DEST vector contains a Kozak sequence followed by an NH_2-terminal signal sequence derived from the *Trypanosoma cruzi* lysosomal α-mannosidase[18], an 8×His tag, and a StrepII tag[19] adjacent to the *att*R1 recombination site (**Supplementary Fig. 2**). The pGEn2-DEST vector employs the same signal sequence and 8×His tag, but has an AviTag sequence[20] and a 'superfolder' GFP domain[21] adjacent to the *att*R1 recombination site (**Supplementary Fig. 2**). The pGEn3-DEST vector has the same signal sequence, 8×His tag, AviTag and GFP tag as pGEn2-DEST, but also includes an immunoglobulin Fc domain[22] sequence between the GFP and the *att*R1 site (**Supplementary Fig. 2**).

For the pGEc1-DEST and pGEc2-DEST COOH-terminal fusion vectors the CMV promoter and artificial intron led directly to the *att*R1 site followed by the selection cassette (**Supplementary Fig. 2**). For these vectors, the Kozak sequence and initiating Met residue were provided by the glycoenzyme donor clones. The fusion protein sequences for these vectors extend downstream of the *att*R2 site, where the pGEc1-DEST vector contained an 8×His tag and StrepII tag and the pGEc2-DEST vector contained a GFP domain, an AviTag sequence, and an 8×His tag (**Supplementary Fig. 2**). Each of these selection cassettes and fusion sequences were generated by gene synthesis

doi:10.1038/nchembio.2539

(Thermo Fisher Scientific) and swapped into the original ST6GAL1-pGEn2 vector as restriction fragments to replace the ST6GAL1 and adjoining fusion sequences.

Gateway recombination of donor clones into mammalian DEST vectors. Transfer of the glycoenzyme coding regions from the donor vectors into the respective mammalian DEST expression vectors was accomplished using an LR Clonase reaction. Equal quantities of donor and DEST expression vectors were used for LR Clonase reactions (Thermo Fisher Scientific) according to the manufacturer's instructions. The reaction products were transformed into 5-alpha competent *E. coli* (New England BioLabs, Ipswitch, MA) and plated on LB plates containing ampicillin, and the resulting plasmid clones were screened by restriction mapping and verified by DNA sequencing of the entire coding region.

Small-scale transient transfection of mammalian expression vectors into HEK293 cells. FreeStyle 293F cells (Thermo Fisher Scientific) were maintained in suspension at 0.5–3.0 × 10^6 cells/mL in a humidified CO_2 platform shaker incubator at 37 °C using serum-free Freestyle 293 expression medium (Thermo Fisher Scientific). FreeStyle 293F cells were grown to a density of ~2.5 × 10^6 cells/ml and transfected by direct addition of 4.5 μg/ml of the respective expression plasmid DNA and 10 μg/ml polyethylenimine (linear 25 kDa PEI, Polysciences, Inc, Warrington, PA) to the suspension cultures as previously described[13]. The cultures were diluted 1:1 with Freestyle 293 expression medium containing 4.4 mM valproic acid (2.2 mM final) 24 h after transfection, and protein production was continued for another 4–5 d at 37 °C.

Construction of baculovirus-based Gateway DEST vectors. Three different baculovirus DEST vectors were constructed for Gateway insertion of glycoenzyme coding sequences, as illustrated in **Supplementary Figure 3**. Each comprised a baculovirus genome of ~135–140 Kbp lacking the viral chitinase and cathepsin-like protease genes with key genetic features added to their polyhedrin loci. Starting with *orf603*, which has the opposite orientation, the features of the NH_2 terminal tag fusions extending downstream in the 5′ to 3′ direction include the baculoviral polyhedrin promoter (pH), a Kozak consensus sequence, an ATG initiator codon, and sequences encoding either a 'short' tag consisting of the honeybee prepromellitin signal peptide[52], an 8×His tag, and a StrepII tag, or a 'long' tag consisting of the same signal sequence and 8×His tag followed by an AviTag sequence and superfolder GFP domain followed by an *att*R1 site (short N-term fusion and GFP N-term fusion baculovirus DNAs, respectively, in **Supplementary Fig. 3**). The first *att*R1 site in both of these 'N-term tag fusion' vectors is followed by a herpes simplex virus 1 thymidine kinase gene, which is used for negative selection, and the *E. coli* β-galactosidase gene, which is used to identify parental virus clones after Gateway recombination, and these markers are followed by the second *att*R site (**Supplementary Fig. 3**). In contrast, the baculovirus DEST vector designed for COOH terminal tag fusions includes the baculovirus *orf603* and *polh* promoter followed directly by the thymidine kinase and β-galactosidase genes, which are flanked by *att*R sites (**Supplementary Fig. 3**). In addition, in this DEST vector, the downstream *att*R site is followed by sequences encoding a short tag consisting of StrepII and 8×His tags. Each of the baculovirus-based Gateway DEST vectors includes transcriptional termination signals located within the 3′UTR of the baculovirus polyhedrin gene, just upstream of *orf1629* (**Supplementary Fig. 3**).

Gateway recombination of donor clones into baculovirus DEST vectors and isolation of recombinant baculovirus expression vectors. The donor clones described above were used to insert sequences encoding glycoenzymes and TEV cleavage sites into the baculovirus DEST vectors by Gateway recombination. The encoded TEV cleavage sites were designed to appear on the COOH- or NH_2-terminal sides of the tags encoded by the NH_2 terminal or COOH terminal tag fusion vectors, respectively. Gateway recombination was performed in LR Clonase (Thermo Fisher Scientific) reactions performed with a mixture of each donor plasmid and baculovirus DEST vector, according to the manufacturer's instructions. Following the addition of Cellfectin (Thermo Fisher Scientific) and a short incubation period, each LR reaction was used to transfect Sf9 cells seeded at a density of 0.8 × 10^6 cells/well into six-well plates. The cells were incubated for 5 h at 28 °C, and the transfection mixture was then removed and replaced by insect cell growth medium containing 100 μM gancyclovir. The cells were incubated for another 5 d at 28 °C for baculovirus progeny production with gancyclovir selection against the parental virus. Cell-free media containing these progeny were then harvested and used to isolate individual baculovirus clones by plaque assays in the presence of X-gal, as previously described[14]. After 7–10 d, at least three plaque purified (PP1) clones with white plaque phenotypes were picked and amplified in Sf9 cells, and the resulting PP1P1 virus stocks were assayed for glycoenzyme expression and secretion by immunoblotting with an anti-His tag antiserum, as previously described[14]. Once identified, samples of the PP1P1 baculovirus clones passing this screen were amplified to PP1P2 by infecting fresh Sf9 cell cultures, and the PP1P2 stocks were titered by plaque assay and finally used to document glycoenzyme production and secretion levels, as described below.

Recombinant glycoenzyme production in insect cells. Glycoenzyme production and secretion in the baculovirus-insect cell system was comprehensively documented by performing small-scale infections in Sf9 cells, as previously described[14]. Briefly, the cells were seeded at a density of 1 × 10^6 cells/well into six-well plates in a serum-free growth medium (ESF921; Expression Systems), then infected with individual baculovirus vectors at a multiplicity of infection of 5 plaque forming units/cell. At 48 h post-infection, the cells were scraped into the media, the cell-free media were harvested, and the cell pellets were boiled in SDS–PAGE sample buffer. Samples of the total cell lysates and cell free supernatants, which were also mixed with SDS–PAGE sample buffer and boiled, were then analyzed by SDS–PAGE with Coomassie blue staining or immunoblotting using PVDF membranes (Immobilon-P, Millipore, Billerica, MA) and an anti-His tag antibody (Thermo Fisher Scientific), as previously described[14], to produce the results shown in **Supplementary Figure 5b**,**c**.

We also performed midscale production runs in the baculovirus-insect cell system using a subset of BEVs encoding 20 different members of the GT29 sialyltransferase family. Briefly, Sf9 cells were seeded into 50 mL shake flasks in ESF921 and allowed to reach a density of ~1 × 10^6 cells/mL in a shaking incubator set at 125 r.p.m. and 28 °C. The cells were then gently centrifuged and each culture was infected with a BEV encoding a tagged sialyltransferase catalytic domain at a multiplicity of infection of five plaque-forming units/cell, as previously described[14]. At 48 h post-infection, the cell-free supernatant was harvested and ultracentrifuged to remove extracellular baculovirus and debris, and the resulting supernatant was used to affinity purify the secreted sialyltransferase products, as previously described[14], except we used ProBond nickel chelating resin (Thermo Fisher Scientific). Finally, we analyzed samples of the cell free supernatants (starting material), unbound fraction (supernatants obtained after ProBond absorption), and elution fractions by SDS–PAGE with Coomassie blue staining or immunoblotting, as described above. In separate assays, we monitored GFP fluorescence in the starting material, flow through, and elution fractions, as described below.

Detection of recombinant glycoenzymes from HEK293 cells. For small-scale screening of fusion protein expression in transiently transfected HEK293 cells, aliquots of the cell suspensions were harvested and subjected to direct GFP fluorescence measurement on a SpectraMAX GeminiXS microplate reader (Molecular Devices, Sunnyvale, CA) using an excitation wavelength of 450 nm and emission wavelength of 515 nm. The cultures were then clarified by centrifugation at 2,500 r.p.m. for 10 min in a microfuge and the supernatants were subjected to GFP fluorescence measurement as described above. Fluorescence values in the supernatants were subtracted from total culture fluorescence values to determine cell-associated fluorescence. A recombinant form of a GFP fusion protein was expressed, purified and used to derive a factor that was used to convert GFP fluorescence values to milligrams of recombinant protein. Aliquots of the clarified cell supernatant and cell pellets were also boiled in SDS-PAGE sample buffer and analyzed by SDS–PAGE with Coomassie blue staining or immunoblotting, as described above, except a mouse anti-poly-His tag antibody (Qiagen, Inc., Germantown, MD) was used.

Scale-up of recombinant glycoenzyme production in HEK293 cells for structural studies. Recombinant enzyme production for protein structural studies was performed in a mutant HEK293 cell line defective in Asn-linked glycan maturation (HEK293S (GnTI$^-$) cells[23] (ATCC)) using modifications of workflows based on prior structural studies on rat ST6GalI[24]. The HEK293S (GnTI$^-$) cells were maintained at 0.5–3.0 × 10^6 cells/ml in a humidified CO_2 platform shaker incubator at 37 °C in cell culture medium comprised of 9 volumes of Freestyle 293 expression medium (Thermo Fisher Scientific) and 1 volume of Ex-cell 293 serum-free medium (Sigma; 9:1 medium). Transfections were accomplished by resuspending cells at ~2.8 × 10^6 cells/ml in fresh 9:1 medium, followed by direct addition of 4.5 μg/ml of the plasmid DNA (ST6GalNAcII-pGEn2) and 10 μg/ml polyethyleneimine to the suspension cell culture. The cultures were diluted 1:1 with 1:9 medium containing 4.4 mM valproic acid (2.2 mM final) 12-24 h after transfection, and protein production was continued for a further 4–5 d at 37 °C.

In cases in which recombinant proteins were labeled with selenomethionine (Sigma-Aldrich), cells were transfected as above, and the media were exchanged 12 h after transfection for custom methionine-free Freestyle 293 expression medium (Thermo Fisher Scientific) for 6 h to deplete methionine pools. The cultures were subsequently resuspended in methionine-free Freestyle 293 expression medium containing 60 mg/L selenomethionine at density of 2.0 × 10^6 cells/ml. The protein production was continued for 4–5 d at 37 °C before harvest of the conditioned medium.

Protein purification, deglycosylation, and tag removal employed workflows similar to prior structural studies on rat ST6GalI[24]. The conditioned culture medium was harvested, clarified by centrifugation, passed through a 0.45 μm filter (Millipore) and adjusted to contain 20 mM imidazole, 200 mM NaCl, and 30 mM sodium phosphate, pH 7.2, and loaded onto a 25 ml column of Ni^{2+}-NTA Superflow (Qiagen, Valencia, CA) column equilibrated with 20 mM HEPES, 300 mM NaCl, 20 mM imidazole, pH 7.2. The column was washed with column buffer and eluted successively with column buffer containing 50 mM imidazole, 100 mM imidazole, and finally 300 mM imidazole. The 300 mM imidazole elution fractions were pooled and concentrated to 1 mg/ml using a 10 kDa molecular mass cutoff ultrafiltration membrane (Millipore, Billerica, MA). Purified recombinant TEV protease and EndoF1 expressed and purified as described[24] were added at ratios of 1:15 and 1:8 relative to the GFP-ST6GalNAcII, respectively, and incubated at room temperature for 3–5 h, followed by 4 °C overnight to cleave the tag and glycans. The protein mixture was diluted 15-fold in 50 mM sodium phosphate, pH 7.2, 800 mM NaCl, 10% glycerol to lower the imidazole concentration and loaded onto a 25 ml Ni^{2+}-NTA column to remove the fusion tag and His-tagged TEV protease and EndoF1. The unbound protein preparation was then concentrated by ultrafiltration and further purified on a Superdex 75 column (GE Healthcare) preconditioned with a buffer containing 20 mM HEPES, 200 mM NaCl, 60 mM imidazole, pH 7.2. Peak ST6GalNAcII fractions were collected and concentrated by ultrafiltration to 9–10 mg/ml for crystallization.

Kinetic analysis. Sialyltransferase enzyme activity was measured using a phosphatase-coupled assay (malachite green phosphate detection kit, R&D Systems, Minneapolis, MN) essentially as previously described[24], except asialofetuin was used as the acceptor. Assays were performed in a 50 μl reaction volume containing 100 mM MES, pH 6.5, 0.1 μg of recombinant ST6GalNAcII, CD73 (0.5 ng/μl), asialofetuin (250 μM for routine assays or varied from 25 to 800 μM for kinetic analysis), and CMP-Neu5Ac (250 μM for routine assays or varied from 25 to 1,000 μM for kinetic analysis). Following 30 min incubation at 37 °C, the reaction was stopped by addition of the malachite green phosphate detection reagents and, which was followed by further incubation as described by the manufacturer[24]. Absorbance at 620 nm was determined and compared with equivalent analyses for a phosphate standard curve. Enzyme activity values (nanomoles/min) were determined at varied substrate concentrations, and kinetic data were fit using GraphPad Prism software (version 5.00, La Jolla, CA) to determine K_m and k_{cat} values.

Crystallization and structure determination. Crystallization conditions for selenomethionine-labeled human ST6GalNAcII were initially screened using the high-throughput crystallization facility at the Hauptman Woodward Medical Research Institute[53]. Based on the results, the protein was crystallized using the microbatch method at 4 °C with a protein solution (2 μl) containing the enzyme (9.3 mg/ml) in a buffer consisting of 20 mM HEPES (pH 7.4), 200 mM NaCl, 60 mM imidazole, and recombinant peptide *N*-glycosidase (1:900) expressed in *E. coli*. The protein solution was mixed with 2 μl of the precipitant solution comprising 100 mM ammonium sulfate, 100 mM sodium citrate (pH 4.2), and 24% (w/v) PEG 20K. The crystals appeared after several days, and were harvested after one week, cryoprotected by supplementing the crystallization solution with 20% (v/v) glycerol, and flash-frozen in liquid nitrogen for data collection at 100 °K. A single-wavelength anomalous diffraction (SAD) data set to resolution 3.1 Å was collected at the peak absorption wavelength (0.9792 Å) of selenium at the 24-ID-E beamline of the Advanced Photon Synchrotron (APS). The diffraction images were processed with the HKL2000 package[54]. The crystals of the apo enzyme belong to space group *P*1, and there are six protomers in the asymmetric unit (ASU) of the crystal. The selenium sites were determined by the direct method using SHELX[55], which located 23 of 30 possible selenium sites. The majority of the initial model comprising six protomers were manually built with the program XtalView[56].

The selenomethionine-labeled ST6GalNAcII in complex with CMP was subsequently crystallized using the crystallization condition for the apo enzyme, except the protein solution contained 5 mM CMP. A single-wavelength native diffraction data set to resolution 2.35 Å was collected at the 14-1 beam line of the SLAC National Accelerator Laboratory. The diffraction images were processed with the HKL2000 package[54]. The crystals of the CMP complex also belong to space group *P*1, and there are six protomers per ASU of the crystal. The structure was determined using the molecular replacement method with the program COMO[57]. All stages of the structure refinement were performed using the crystallographic programs CNS 1.3 (ref. 58) and PHENIX[59]. PHENIX was also used in the last stage of refinement. The statistics for data collection and refinement are shown in **Supplementary Table 3**.

Statistical analysis. All expression studies were performed at least twice, and fluorescence values for the expression products in mammalian cells were ± 10% of the indicated values in each case. In **Figure 3a**, the bold red bar and error bars represent the mean and s.d. for the GFP fluorescence values relative to the qualitative values for baculovirus expression. The trend line in **Figure 3a** (red dotted line) is shown for the mean GFP fluorescence values. In **Figure 3a**, the black trend line represents a linear regression for a subset of the correlated fluorescence values (blue diamonds) comparing mammalian and insect cell expression for the human sialyltransferases.

Life sciences reporting summary. Further information on experimental design and reagents is available in the **Life Sciences Reporting Summary**.

Data availability. Atomic coordinates and structure factors for human ST6GalNAcII have been deposited in the Protein Data Bank under the following accession codes: apo-enzyme (6APJ) and CMP complex (6APL). Construct designs, annotations and sequences are summarized on our website (http://glycoenzymes.ccrc.uga.edu/). Plasmids are available from DNASU (https://dnasu.org) and baculovirus stocks are available from the Jarvis laboratory (dljarvis@uwyo.edu). All data generated or analyzed during this study are included in this published article (and its supplementary information files) and from the corresponding author on reasonable request.

50. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).
51. Gerhard, D.S. *et al.* The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). *Genome Res.* **14**, 2121–2127 (2004).
52. Tessier, D.C., Thomas, D.Y., Khouri, H.E., Laliberté, F. & Vernet, T. Enhanced secretion from insect cells of a foreign protein fused to the honeybee melittin signal peptide. *Gene* **98**, 177–183 (1991).

doi:10.1038/nchembio.2539

53. Luft, J.R. *et al.* A deliberate approach to screening for initial crystallization conditions of biological macromolecules. *J. Struct. Biol.* **142**, 170–179 (2003).
54. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
55. Sheldrick, G.M. A short history of SHELX. *Acta Crystallogr. A* **64**, 112–122 (2008).
56. McRee, D.E. XtalView/Xfit--A versatile program for manipulating atomic coordinates and electron density. *J. Struct. Biol.* **125**, 156–165 (1999).
57. Jogl, G., Tao, X., Xu, Y. & Tong, L. COMO: a program for combined molecular replacement. *Acta Crystallogr. D Biol. Crystallogr.* **57**, 1127–1134 (2001).
58. Schröder, G.F., Levitt, M. & Brunger, A.T. Super-resolution biomolecular crystallography with low-resolution data. *Nature* **464**, 1218–1222 (2010).
59. Adams, P.D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010).

doi:10.1038/nchembio.2539

Supplementary Results for:

Expression system for structural and functional studies of human glycosylation enzymes

Kelley W. Moremen, Annapoorani Ramiah, Melissa Stuart, Jason Steel, Lu Meng, Farhad Forouhar, Heather Moniz, Gagandeep Gahlay, Zhongwei Gao, Digantkumar Chapla, Shuo Wang, Jeong-Yeh Yang, Pradeep Prabhakar, Roy Johnson, Mitche dela Rosa, Christoph Geisler Alison V. Nairn, Jayaraman Seetharaman, Sheng-Cheng Wu, Liang Tong, Harry J. Gilbert, Joshua LaBaer, and Donald L. Jarvis

Supplementary Table 1. Summary of human glycoenzyme expression constructs.

Supplementary Table 2. Kinetic analysis of human ST6GalNAcII

Supplementary Table 3. Summary of crystallographic data collection and refinement statistics

Supplementary Figure 1. Strategy for cloning truncated and full length glycoenzymes in Gateway donor vectors.

Supplementary Figure 2. Design strategy for NH_2-terminal and COOH-terminal fusion vectors for expression in mammalian cells.

Supplementary Figure 3. Design strategy for NH_2-terminal and COOH-terminal fusion vectors for expression in baculovirus-infected insect cells.

Supplementary Figure 4. Expression and secretion of representative human GT29 sialyltransferases in different NH_2-terminal fusion vectors.

Supplementary Figure 5. Sialyltransferase fusion protein expression in HEK293 cells and insect cells.

Supplementary Figure 6. Purification of secreted recombinant fusion proteins.

Supplementary Figure 7. ST6GalNAcII expression, purification, and tag/glycan cleavage

Supplementary Figure 8. Sequence alignment and secondary structures of mammalian sialyltransferases.

Supplementary Figure 9. Structure and putative membrane tethering of human ST6GalNAcII.

Supplementary Figure 10. Comparisons of mammalian sialyltransferase structures.

Supplementary Data Set 1. Summary of human glycosylation enzyme fusion protein expression strategies and expression results for production in HEK293 cells and BEVS (see separate **Supplementary Data Set 1.xlsx** Microsoft Excel file)

Supplementary Data Set 2. DNA sequences of human glycosylation enzyme expression constructs (see separate **Supplementary Data Set 2.xlsx** Microsoft Excel file)

Supplementary Table 1. Summary of human glycoenzyme expression constructs. Four categories of enzymes were captured as Gateway donor clones for transfer into baculovirus and mammalian expression vectors: glycosyltransferases (*GT*), glycoside hydrolases (*GH*), sulfotransferases (*ST*), and a miscellaneous collection of glycan modifying enzymes (*Other*) as indicated by the gene list (**Supplementary Data Set 1**). Each enzyme coding region was designed for truncation and fusion protein tagging strategies based on its transmembrane topology (e.g. NH_2-terminal TMD (*Type 2 TMD*), COOH-terminal TMD (*Type 1 TMD*), multipass TMD or internal TMD (*Multipass or internal TMD*), the presence of an NH_2-terminal signal sequence (*N-terminal signal sequence*), or the absence of signal sequence or TMD segments (*Cytosolic*; see Methods for design strategies). As a result, protein coding regions in each enzyme category were designed either as NH_2-terminal or COOH- terminal fusions as indicated in the table. A subset of genes was designed as both NH_2-terminal fusions and separately as COOH-terminal fusions in an effort to find the most effective strategy for expression and secretion. Coding regions for each enzyme were isolated either by PCR from MGC clone templates or from human cDNA sources (*cDNA*) or by gene synthesis (*Gene Synthesis*) as described in Methods. The tables summarizing the fusion protein and protein coding region capture strategies and details for each gene are found in **Supplementary Data Sets 1 and 2**.

GT	**CAZy families = 44**			***Multipass***	***N-terminal***		
		Type 2 TMD	***Type 1 TMD***	***or internal TMD***	***signal sequence***	***Cytosolic***	***Total***
	Total genes	136	21	21	12	11	201
	Fusion strategy						
	N-term fusions	136	12	1	11	4	164
	C-term fusions	0	21	21	5	9	56
	Coding region capture strategy						
	cDNA	77	23	14	9	10	133
	Gene synthesis	59	10	8	7	3	87

GH	**CAZy families = 28**			***Multipass***	***N-terminal***		
		Type 2 TMD	***Type 1 TMD***	***or internal TMD***	***signal sequence***	***Cytosolic***	***Total***
	Total genes	12	1	3	46	12	74
	Fusion strategy						
	N-term fusions	12	1	3	40	7	63
	C-term fusions	0	0	0	35	10	45
	Coding region capture strategy						
	cDNA	8	1	3	68	10	90
	Gene synthesis	4	0	0	7	7	18

ST				***Multipass***	***N-terminal***		
		Type 2 TMD	***Type 1 TMD***	***or internal TMD***	***signal sequence***	***Cytosolic***	***Total***
	Total genes	35	0	0	0	0	35
	Fusion strategy						
	N-term fusions	35	0	0	0	0	35
	C-term fusions	0	0	0	0	0	0
	Coding region capture strategy						
	cDNA	20	0	0	0	0	20
	Gene synthesis	15	0	0	0	0	15

Other				***Multipass***	***N-terminal***		
		Type 2 TMD	***Type 1 TMD***	***or internal TMD***	***signal sequence***	***Cytosolic***	***Total***
	Total genes	4	0	4	14	1	23
	Fusion strategy						
	N-term fusions	4	0	1	7	0	12
	C-term fusions	0	0	3	7	1	11
	Coding region capture strategy						
	cDNA	0	0	3	8	0	11
	Gene synthesis	4	0	1	6	1	12

Supplementary Table 2: Kinetic parameters for human ST6GalNAcII

	GFP	Donor CMP-Neu5Ac			Acceptor Asialofetuin		
Enzyme form	Fluorescence[a]	K_m	k_{cat}	k_{cat}/K_m	K_m	k_{cat}	k_{cat}/K_m
		(μM)	*(min^{-1})*	*(mM^{-1}min^{-1})*	*(μM)*	*(min^{-1})*	*(mM^{-1}min^{-1})*
GFP-ST6GalNAcII[b]	935	102 ± 18	5.4 ± 0.3	53	130 ± 21	5.1 ± 0.7	39
ST6GalNAcII[c]	520	104 ± 17	5.9 ± 0.2	57	108 ± 14	4.9 ± 0.6	45

[a] Expression and secretion of the GFP-ST6GalNAcII fusion protein in transiently transfected HEK293 cells was determined by measuring the fluorescence of the recombinant protein secreted into the media. Fusion protein production in the HEK293S (GnTI-) cells resulted in reduced expression efficiency.

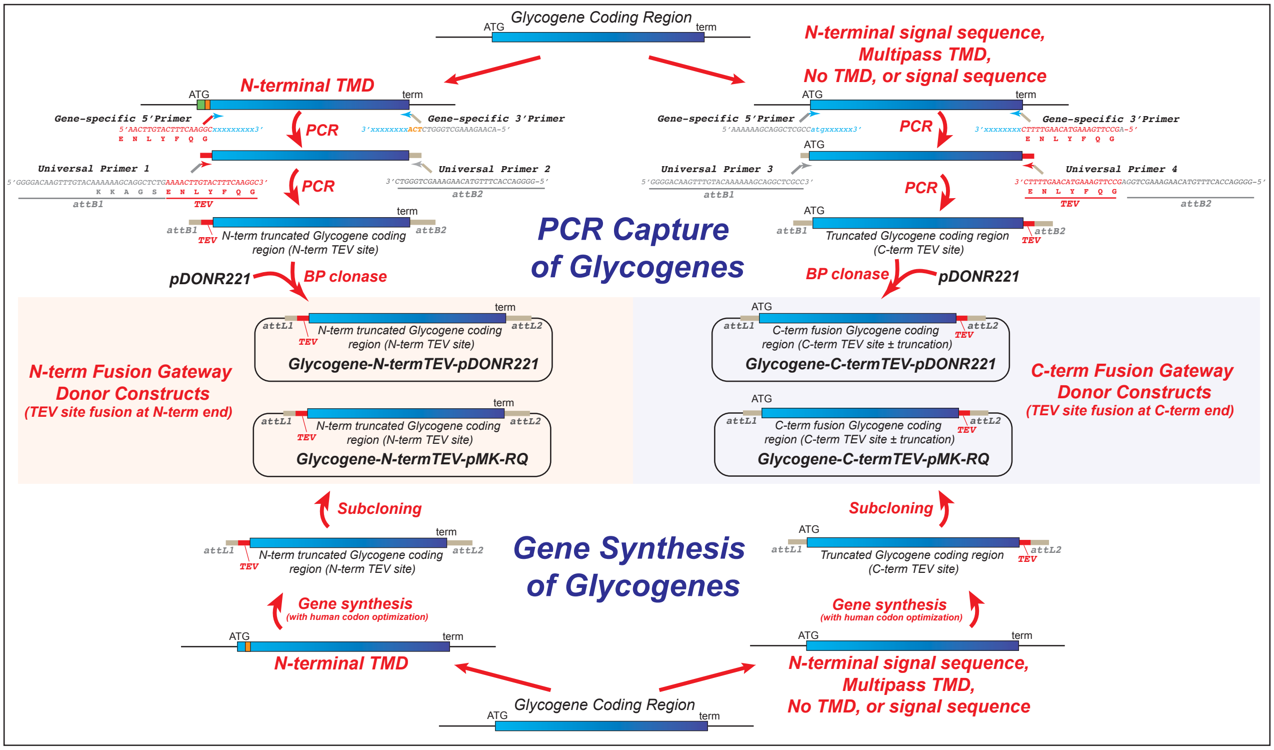
[b] The GFP-ST6GalNAcII fusion protein was expressed in Freestyle 293-F cells and purified by Ni^{2+}-NTA chromatography. The fusion protein tags and complex N-glycans were retained in this enzyme preparation.

[c] The GFP-ST6GalNAcII fusion protein expressed in HEK293S (GnTI-) cells was purified by Ni^{2+}-NTA chromatography and cleaved with TEV protease and endoglycosidase F as described in Methods. The ST6GalNAcII catalytic domain was purified from the fusion protein tags and cleavage enzymes over an additional Ni^{+2}-NTA column and gel filtration prior to enzyme assays.

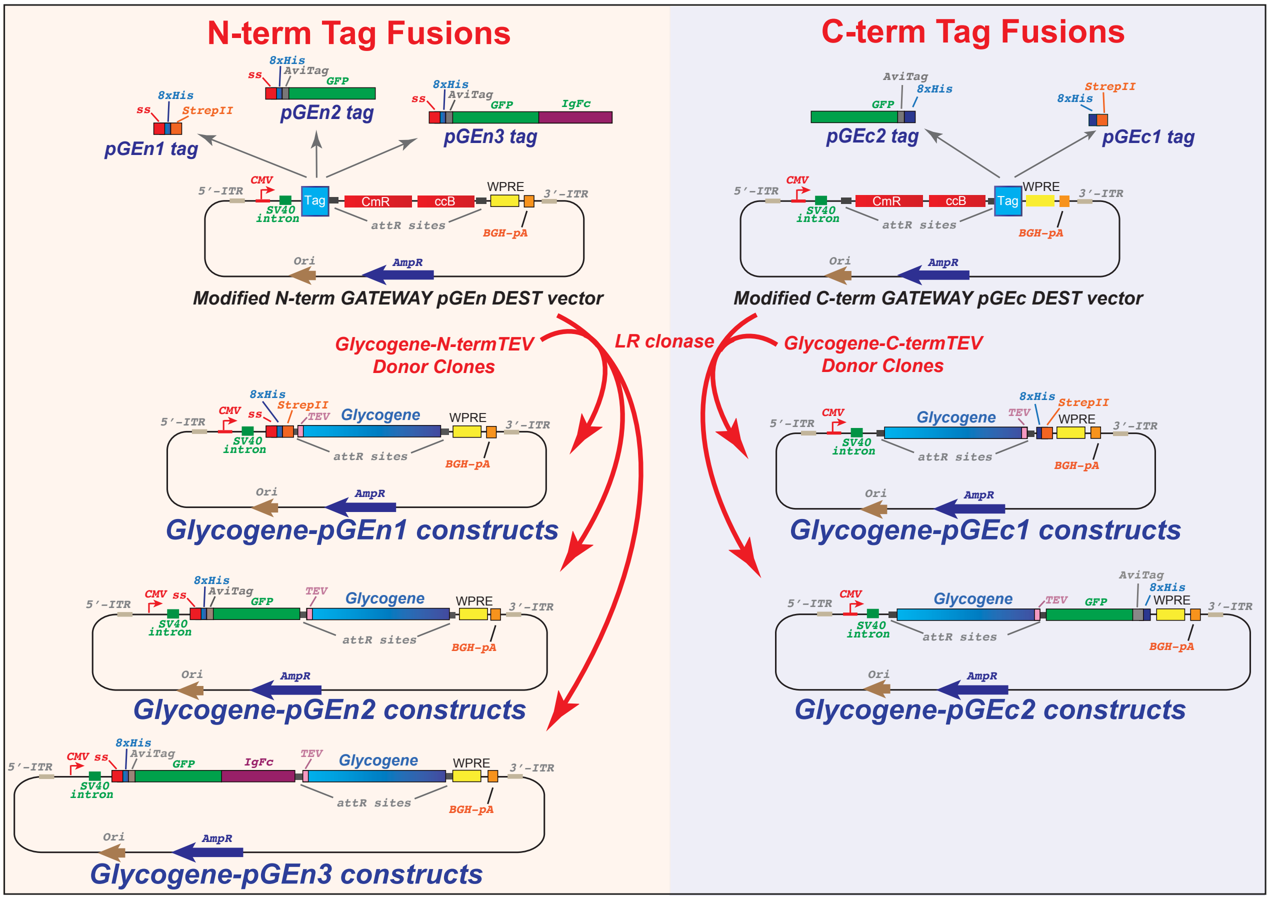
Supplementary Table 3. Summary of crystallographic data collection and refinement statistics

	APO	CMP complex
Data collection		
Space group	*P*1	*P*1
Cell dimensions		
a, *b*, *c* (Å)	71.3, 71.8, 134.2	71.2, 71.1, 138.6
α, β, γ (°)	98.8, 101.9, 103.5	103.7, 97.3,103.0
Resolution (Å)	44.3-3.1(3.25-3.10)*	41.6-2.35(2.39-2.35)
R_{sym} or R_{merge}	7.2(56.7)	4.2(42.1)
I / σI	8.8(1.9)	20.5(1.2)
Completeness (%)	93.2(83.3)	97.5(96.0)
Redundancy	1.4(1.4)	1.6(1.6)
Refinement		
Resolution (Å)	44.3-3.1(3.14-3.10)	41.6-2.35(2.39-2.35)
No. reflections	41,880(4,185)	102,962 (3,064)
R_{work} / R_{free}	20.1/25.2	19.6/23.6
No. atoms		
Protein	12,002	14,169
Ligand/ion	56	266
Water	0	215
B-factors		
Protein	83	48.6
Ligand/ion	95.2	74.8
Water	0	63.2
R.m.s. deviations		
Bond lengths (Å)	0.002	0.004
Bond angles (°)	0.507	0.729

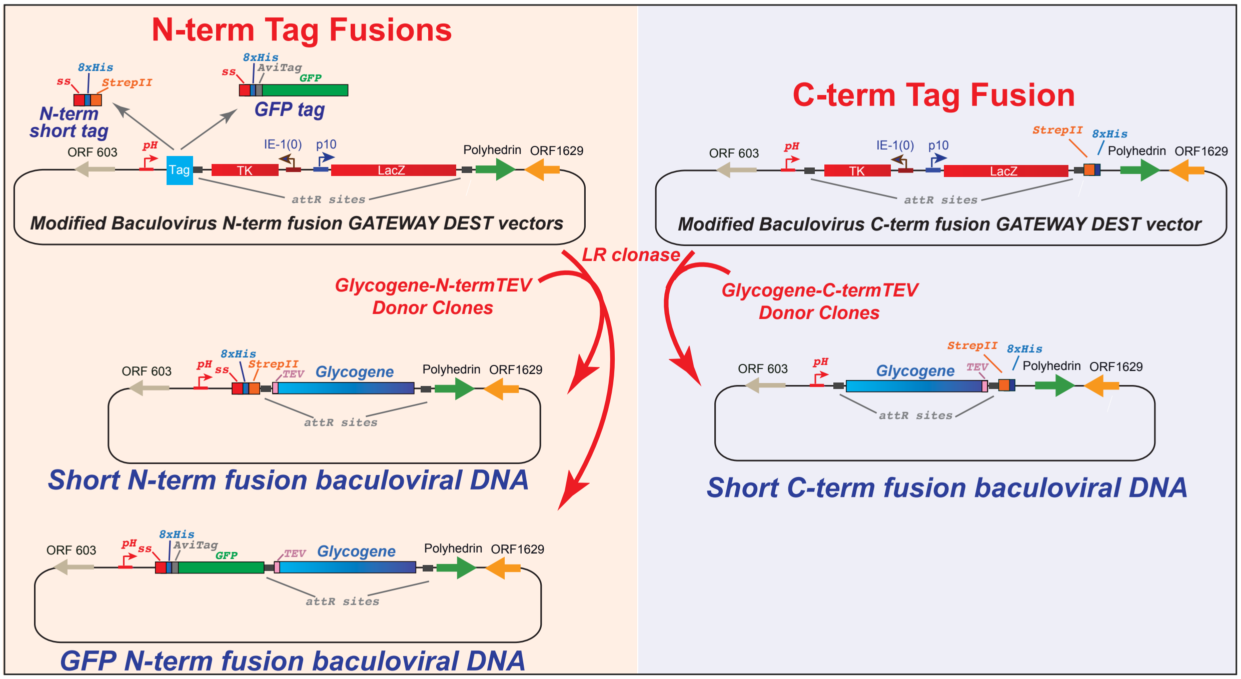
*Values in parentheses are for highest-resolution shell.


Glycogene Coding Region
ATG
term
N-terminal TMD
N-terminal signal sequence,
Multipass TMD,
No TMD, or signal sequence
Gene-specific 5'Primer
5'AACTTGTACTTTCAAGGCxxxxxxxxx3'
E N L Y F Q G
Gene-specific 3'Primer
3'xxxxxxxxACTCTGGGTCGAAAGAACA-5'
PCR
Universal Primer 1
5'GGGGACAAGTTTGTACAAAAAAGCAGGCTCTGAAAACTTGTACTTTCAAGGC3'
K K A G S E N L Y F Q G
attB1
TEV
Universal Primer 2
3'CTGGGTCGAAAGAACATGTTTCACCAGGGG-5'
attB2
N-term truncated Glycogene coding region (N-term TEV site)
pDONR221
BP clonase
Gene-specific 5'Primer
5'AAAAAAGCAGGCTCGCCatgxxxxxx3'
Gene-specific 3'Primer
3'xxxxxxxxCTTTTGAACATGAAAGTTCCGA-5'
E N L Y F Q G
Universal Primer 3
5'GGGGACAAGTTTGTACAAAAAAGCAGGCTCGCC3'
attB1
Universal Primer 4
3'CTTTTGAACATGAAAGTTCCGAGGTCGAAAGAACATGTTTCACCAGGGG-5'
E N L Y F Q G
TEV
attB2
Truncated Glycogene coding region (C-term TEV site)
PCR Capture of Glycogenes
attL1
attL2
N-term truncated Glycogene coding region (N-term TEV site)
Glycogene-N-termTEV-pDONR221
Glycogene-N-termTEV-pMK-RQ
N-term Fusion Gateway Donor Constructs
(TEV site fusion at N-term end)
C-term fusion Glycogene coding region (C-term TEV site ± truncation)
Glycogene-C-termTEV-pDONR221
Glycogene-C-termTEV-pMK-RQ
C-term Fusion Gateway Donor Constructs
(TEV site fusion at C-term end)
Subcloning
Gene Synthesis of Glycogenes
Gene synthesis
(with human codon optimization)
N-term truncated Glycogene coding region (N-term TEV site)
Truncated Glycogene coding region (C-term TEV site)
N-terminal TMD
N-terminal signal sequence,
Multipass TMD,
No TMD, or signal sequence
Glycogene Coding Region

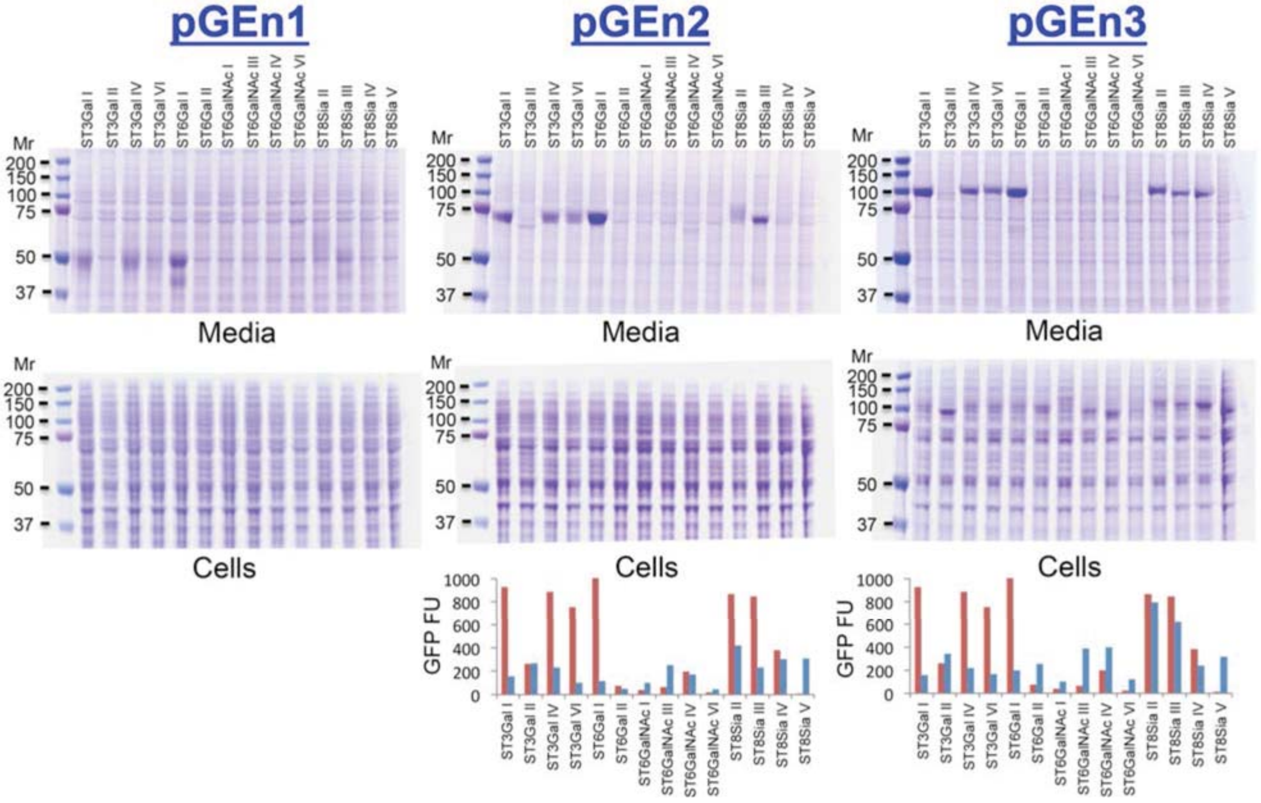
Supplementary Figure 1. Strategy for cloning truncated and full-length glycoenzymes in Gateway donor vectors. The strategy for PCR amplification of the truncated or full-length glycoenzyme coding regions is shown at the top of the figure, including primer designs for initial gene-specific amplification based on the truncation strategy described in Methods and **Supplementary Tables 1 and Supplementary Data Sets 1 and 2**. Enzymes with NH_2-terminal truncations and fusions were amplified using gene specific primers to delete NH_2-terminal sequences and initiate the addition of NH_2-teminal primer-encoded extensions at the NH_2-terminus comprising a TEV protease recognition site and COOH-terminal extensions comprising a Gateway *att*B2 recombination site following the enzyme's termination codon. A second round amplification was then performed using universal primers that completed the TEV recognition site and *att*B1 and *att*B2 recombination sites. The resulting PCR amplification products were then transferred to the Gateway pDONR221 vector by BP recombination (Glycogene-N-TermTEV-pDONR221, see Methods). A similar strategy was used to generate COOH-terminal TEV fusion constructs. In this case, the 5' gene-specific PCR primers were designed to anneal at the ATG initiation codon and initiate the extension of an *att*B1 Gateway recombination site upstream of the Kozak sequence. The 3' gene-specific primers were designed to delete the termination codon (and, in some cases, additional upstream sequences to produce COOH-terminal truncations) and initiated the primer-encoded TEV protease recognition site. A second round amplification was then performed using universal primers to complete the TEV recognition site and *att*B1 and *att*B2 recombination sites. The resulting PCR amplification products were then transferred to the Gateway pDONR221 vector by BP recombination (Glycogene-C-TermTEV-pDONR221). The templates for the coding region PCR reactions were either human MGC clones encoding the respective genes or human tissue cDNA preparations (see Methods). In cases where gene-specific PCR amplification could not be achieved, the desired coding sequences with appropriate truncations, adjoining TEV protease recognition sites, and flanking *att*L recombination sites were generated by gene synthesis in vectors equivalent to pDONR221 (Glycogene-N-termTEV-pMK-RQ and Glycogene-C-termTEV-pMK-RQ). The library of NH_2-terminal and COOH-terminal TEV fusion donor vectors were subsequently used for Gateway recombination reactions, resulting in transfer of the desired coding sequences into the desired eukaryotic DEST expression vectors (**Supplementary Figs. 2** and **3**)


N-term Tag Fusions
C-term Tag Fusions
ss
8xHis
StrepII
pGEn1 tag
AviTag
GFP
pGEn2 tag
IgFc
pGEn3 tag
pGEc2 tag
pGEc1 tag
5′-ITR
CMV
SV40 intron
Tag
CmR
ccB
WPRE
3′-ITR
attR sites
BGH-pA
Ori
AmpR
Modified N-term GATEWAY pGEn DEST vector
Modified C-term GATEWAY pGEc DEST vector
Glycogene-N-termTEV Donor Clones
LR clonase
Glycogene-C-termTEV Donor Clones
TEV
Glycogene
Glycogene-pGEn1 constructs
Glycogene-pGEc1 constructs
Glycogene-pGEn2 constructs
Glycogene-pGEc2 constructs
Glycogene-pGEn3 constructs

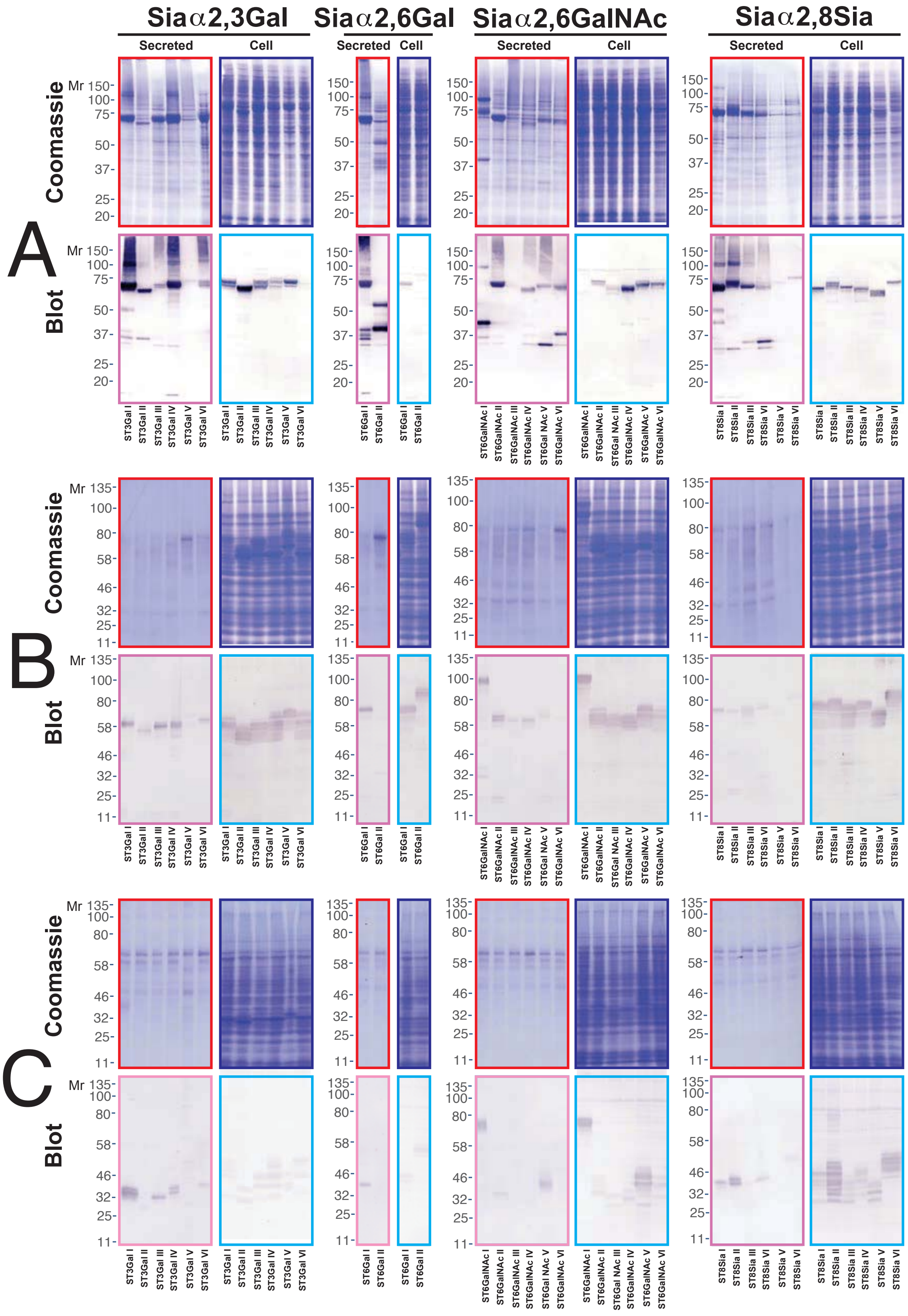
Supplementary Figure 2. Design strategy for NH_2-terminal and COOH-terminal fusion vectors for expression in mammalian cells. Multiple custom mammalian fusion vectors were designed with alternative NH_2-terminal and COOH-terminal fusion tags. All the mammalian expression vectors contained a cytomegalovirus (CMV) promoter, woodchuck hepatitis virus post-transcriptional regulatory element (WPRE), and a bovine growth hormone (BGH) termination and polyadenylation signal, and were adapted for Gateway recombination by inclusion of *att*R sites and a selection cassette to form DEST vectors as described in Methods. Distinctions between the vectors resulted from differences in fusion tag strategies. The pGEn1-DEST vector initiates translation with an NH_2-terminal signal sequence, an 8xHis tag and StrepII tag as described in Methods. The pGEn2-DEST vector employs a signal sequence, 8xHis tag, AviTag, and GFP domain as described in Methods. The pGEn3-DEST vector included the same fusion tag components as pGEn2-DEST, but also included an Ig-Fc domain sequence as described in Methods. Gateway LR recombination of the glycoenzyme coding regions from the donor vectors into these pGEn DEST vectors results in the formation of an 8 amino acid recombination scar followed by the TEV cleavage site and the corresponding enzyme coding region. The COOH-terminal fusion vectors were analogous to the pGEn1 and pGEn2 vectors. In this latter case, the CMV promoter led directly to the *att*R1 site while the Kozak sequence and initiating Met residue were provided by the glycoenzyme donor clone. The COOH-terminal fusion tags were appended downstream of the *att*R2 site in each vector. The pGEc1-DEST vector contained an 8xHis tag and StrepII tag and the pGEc2-DEST vector contained a GFP domain, AviTag sequence, and 8xHis tag. Gateway LR recombination of the glycoenzyme coding regions from the donor vectors into these latter pGEc DEST vectors results in the expression of the full-length protein coding region (or potential COOH-terminal truncation) followed by an 8 amino acid recombination scar, the TEV cleavage site and the respective fusion tags.



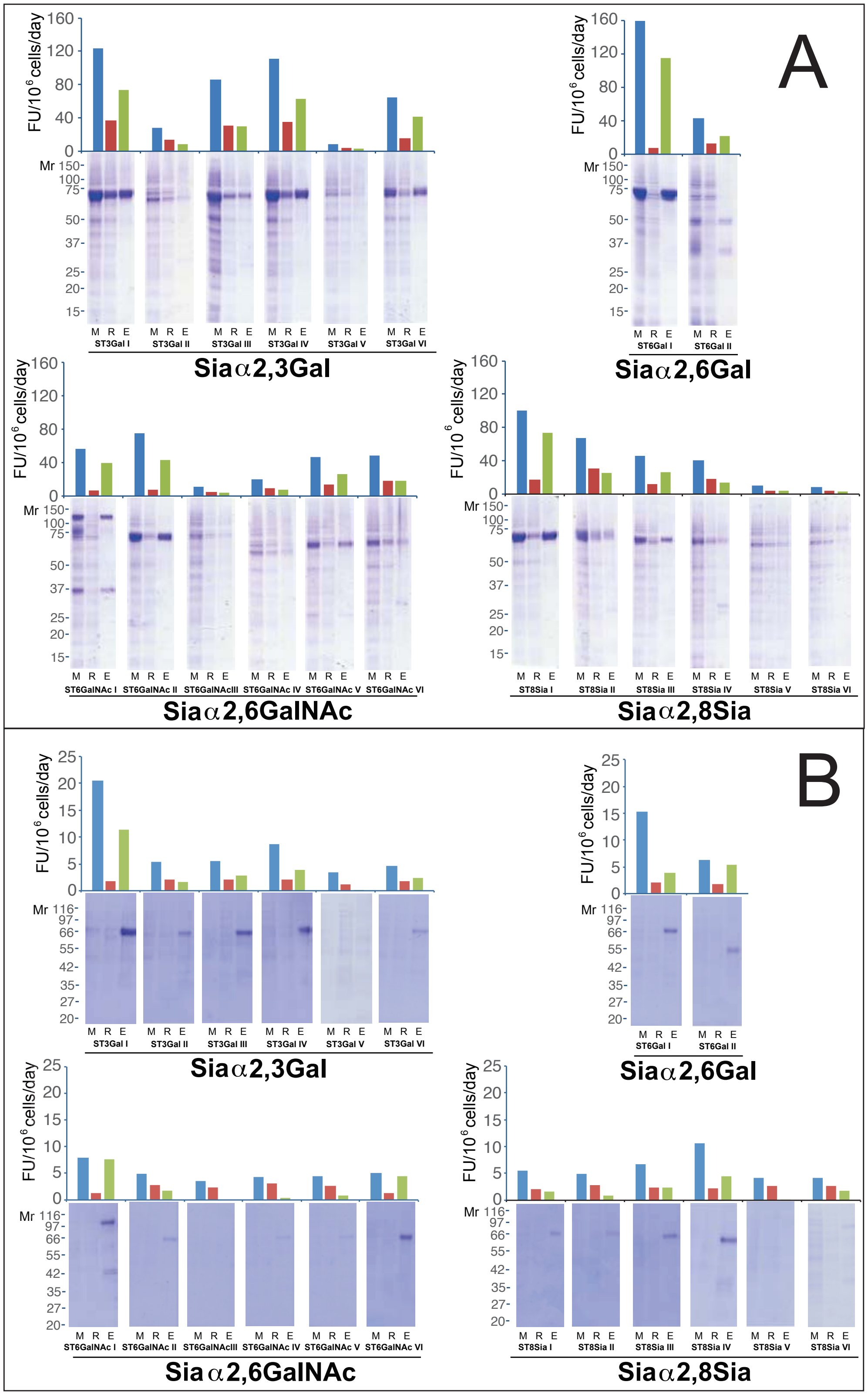
Supplementary Figure 3. Design strategy for NH_2-terminal and COOH-terminal fusion vectors for expression in baculovirus-infected insect cells. Multiple custom baculovirus fusion vectors were designed with alternative NH_2-terminal and COOH-terminal fusion tags. All of the baculovirus expression vectors employed the polyhedrin promoter (*pH*) and polyhedrin termination and polyadenylation sequences (*Polyhedrin*). Each vector was adapted for Gateway recombination by inclusion of *att*R sites flanking a selection cassette as described in Methods. Distinctions between the vectors resulted from differences in fusion tag strategies. For vectors containing short NH_2-terminal fusions, the coding region was initiated with a honeybee mellitin signal sequence, an 8xHis tag, and a StrepII tag. For longer NH_2-terminal GFP fusions the vector employed the honeybee mellitin signal sequence, 8xHis tag, AviTag sequence, and GFP domain as described in Methods. Gateway LR recombination from the donor vectors into these baculoviral DEST vectors results in the formation of an 8 amino acid recombination scar followed by the TEV cleavage site and the corresponding enzyme coding region. A COOH-terminal baculovirus fusion vector was also generated analogous to the mammalian pGEc1 vector containing a StrepII tag and 8xHis tag downstream from the selection cassette. Gateway LR recombination from the donor vectors into this latter expression vector results in the expression of full-length protein coding regions (or potential COOH-terminal truncations) followed by an 8 amino acid recombination scar, the TEV cleavage site and the respective fusion tags.



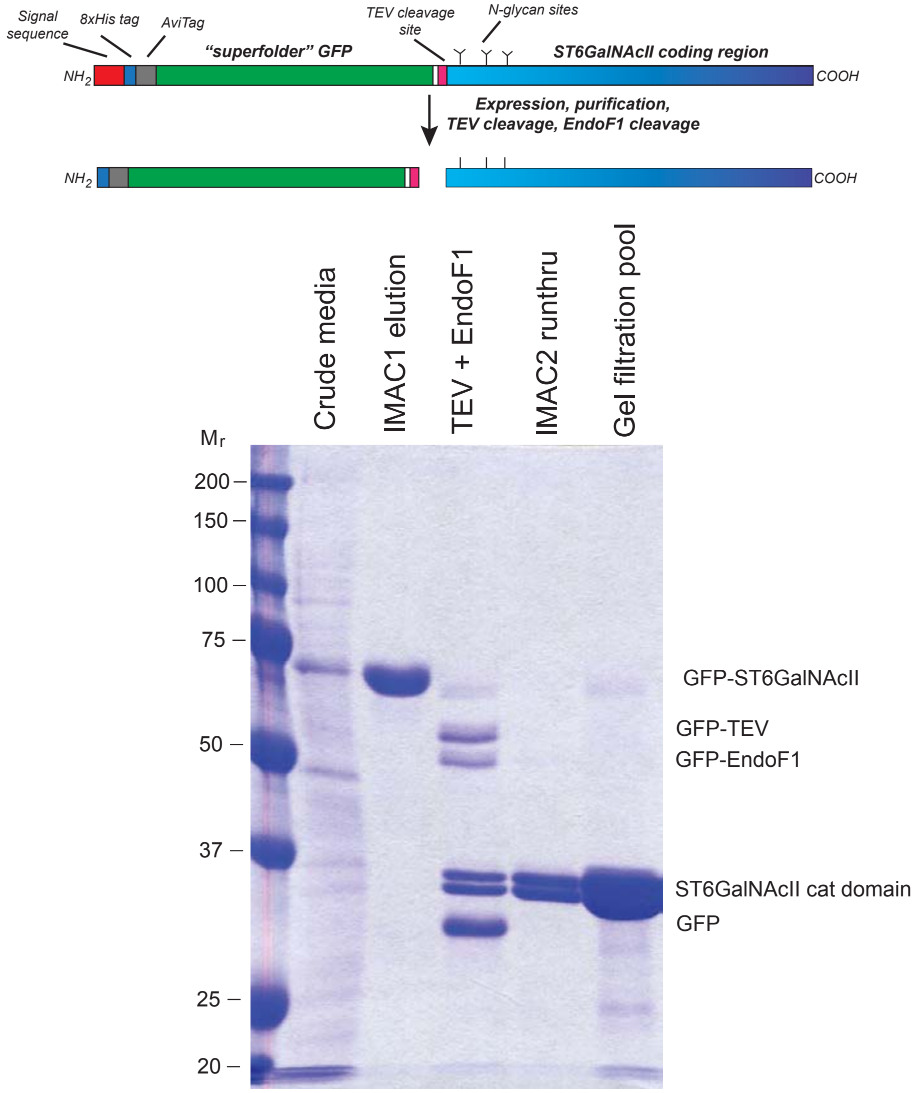
Supplementary Figure 4. Expression and secretion of representative human GT29 sialyltransferases in different NH_2-terminal fusion vectors. A collection of fourteen GT29 sialyltransferase coding regions were transferred into the pGEn1, pGEn2, and pGEn3 NH_2-terminal fusion vectors and tested for expression and secretion by transient transfection of HEK293F cells. Clarified conditioned media and cell extracts were collected and resolved by SDS-PAGE with subsequent Coomassie protein staining. For the pGEn2 and pGEn3 expression constructs harboring a GFP fusion domain, samples of the conditioned media and cell pellets were also tested for fluorescence to quantify the recombinant products (bar charts for pGEn2 and pGEn3 constructs). Red bars represent GFP in the conditioned media and blue bars represent cell associated fluorescence. Anticipated sizes for the sialyltransferase fusion proteins on the SDS-PAGE gels were 37-50 kDa (pGEn1), 64-75 kDa (pGEn2), and 91-102 kDa (pGEn3) depending on the coding region and contribution of glycosylation to molecular mass.



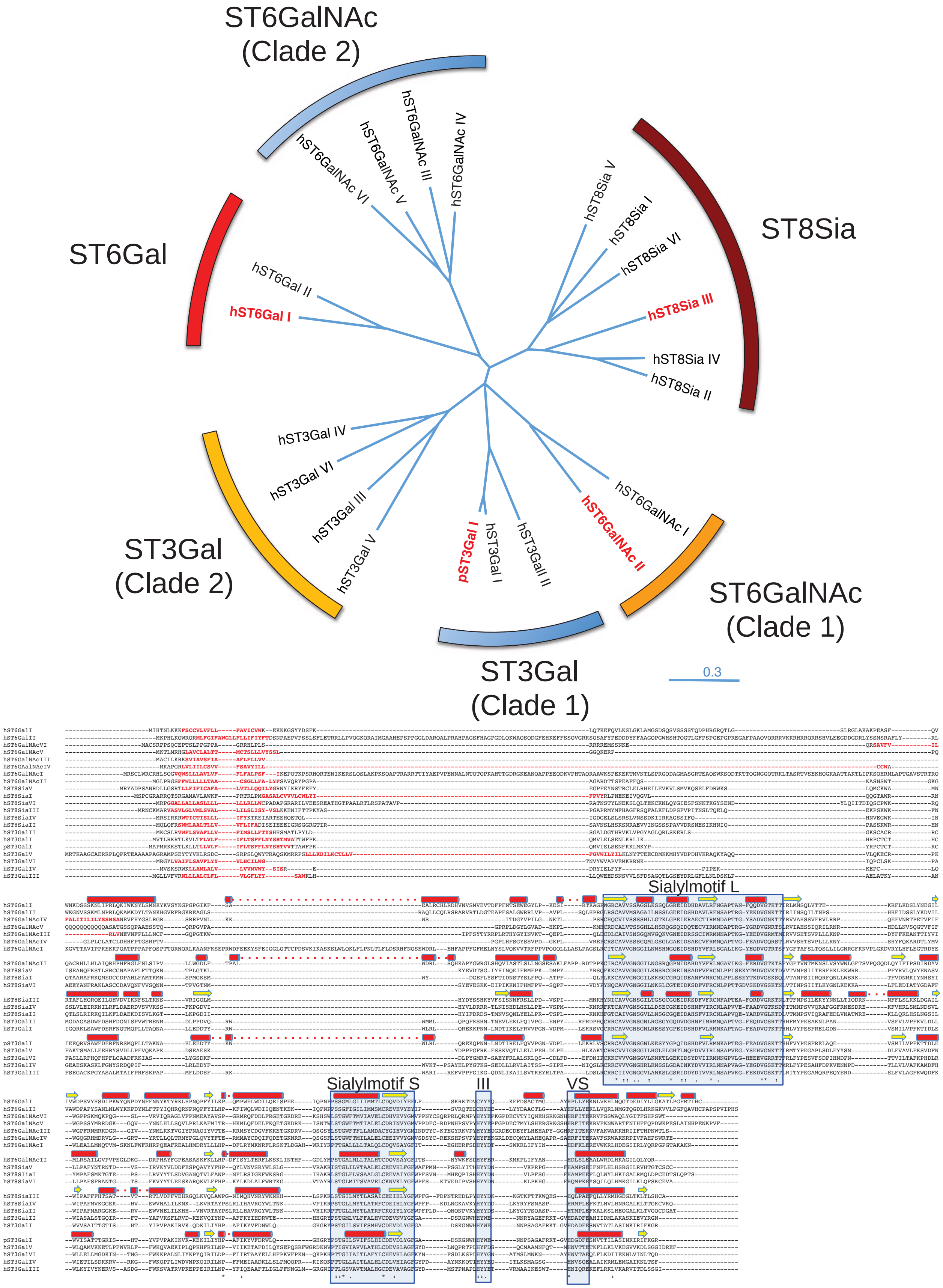
Supplementary Figure 5. Sialyltransferase fusion protein expression in HEK293 cells and insect cells. A collection of expression constructs encoding twenty GT29 sialyltransferases as NH_2-terminal fusions were produced in HEK293 cells (***Panel A***, pGEn2 vector constructs as 8xHis/AviTag/GFP fusions) and in insect cells as NH_2-terminal 8xHis/AviTag/GFP (***Panel B***) and 8xHis/StrepII tag fusions (***Panel C***). Transfected mammalian cell cultures and baculovirus infected insect cell cultures were centrifuged to generate conditioned media (***Secreted***, red or pink boxes) or cell pellets (***Cell***, dark blue or light blue boxes) and samples were resolved by SDS-PAGE. Gels were then either stained with Coomassie brilliant blue (***Coomassie***, red or blue boxes) or blotted and probed with an anti-His tag antibody (***Blot***, pink or light blue boxes). Recombinant products were detected in cell or media samples by immunoblots and in some cases by Coomassie staining. Most of the glycosylated sialyltransferase GFP fusion proteins were visible in blots and protein staining as ~55-75 kDa polypeptides with the exception of the ST6GalNAcI coding region, which is expressed as a ~110 kDa glycosylated fusion protein. The BEV 8xHis/StrepII expression products were ~32-50 kDa polypeptides except for the ST6GalNAcI coding region (~70 kDa). Immunodetection of cell-associated proteins in BEVS often revealed a ladder-like banding appearance that resulted from heterogeneity in glycosylation and potentially proteolysis. A sharper banding pattern was commonly found for the secreted products in the media of BEV cultures. In transfected HEK293 cell cultures, several sialyltransferase coding region products (ST6GalII, ST6GalNAcI, ST6GalNAcV, ST6GalNAcVI, ST8SiaI, ST8SiaII, ST8SiaIII, and ST8SiaIV) were apparently subject to partial proteolysis following secretion, since ~35-45 kDa His/AviTag/GFP bands were evident in the media in addition to the intact fusion proteins. Only the intact fusion proteins were detected in the cell pellets.



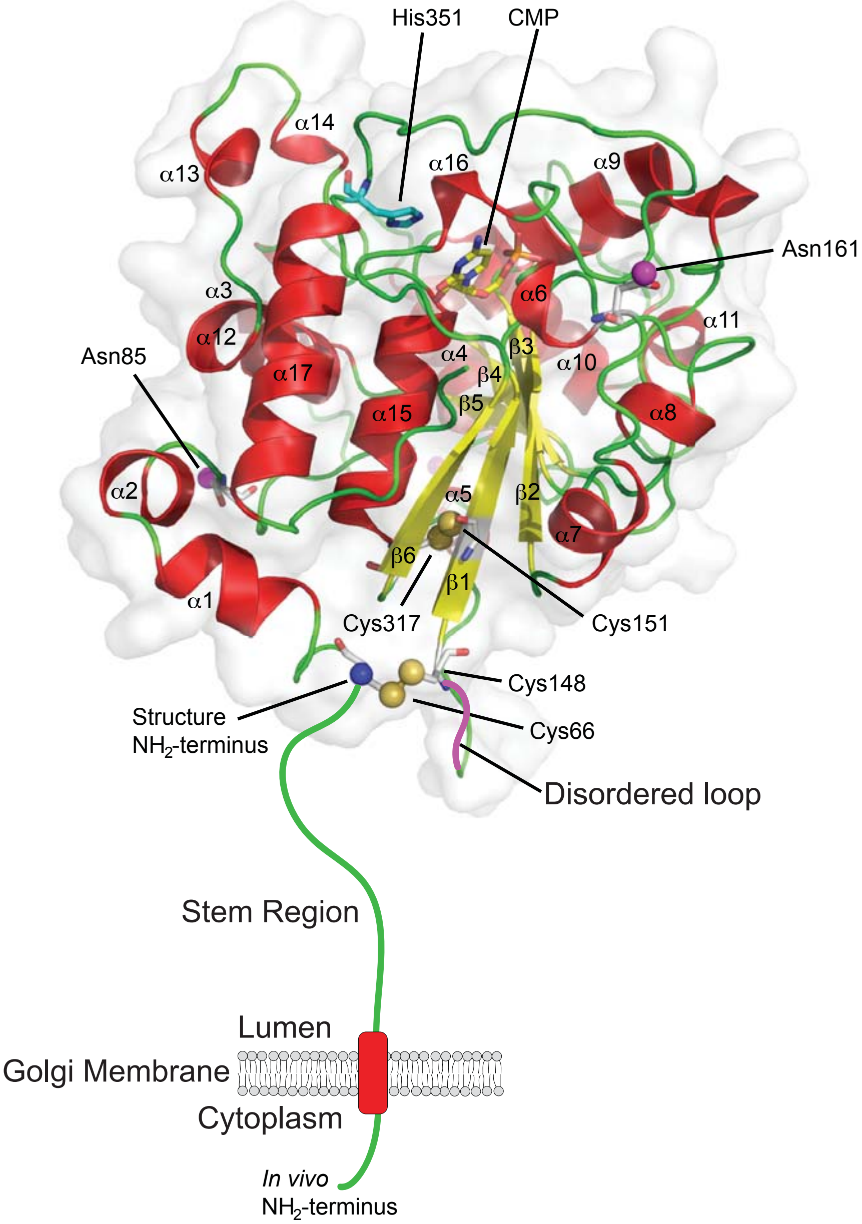
Supplementary Figure 6. Purification of secreted recombinant fusion proteins. The collection of recombinant fusion protein constructs encoding the twenty GT29 sialyltransferases were expressed in HEK293 cells (*Panel A*) and BEVs (*Panel B*) and the conditioned media were used for protein purification by IMAC. Samples corresponding to the crude expression media (M), the column run-through fractions (R), and imidazole elution fractions (E) from each enzyme purification were analyzed, and the data are shown with genes clustered by enzyme specificity subfamily (see **Supplementary Fig. 8**). GFP fluorescence values were also obtained for each fraction and are plotted at the top of each panel for the media (blue bars), column run-through (red bars), and elution (green bars) fractions. Most of the glycosylated sialyltransferase GFP fusion proteins were visible by protein staining as ~55-75 kDa polypeptides with the exception of the ST6GalNAcI coding region, which was expressed as a ~110 kDa glycosylated fusion protein.



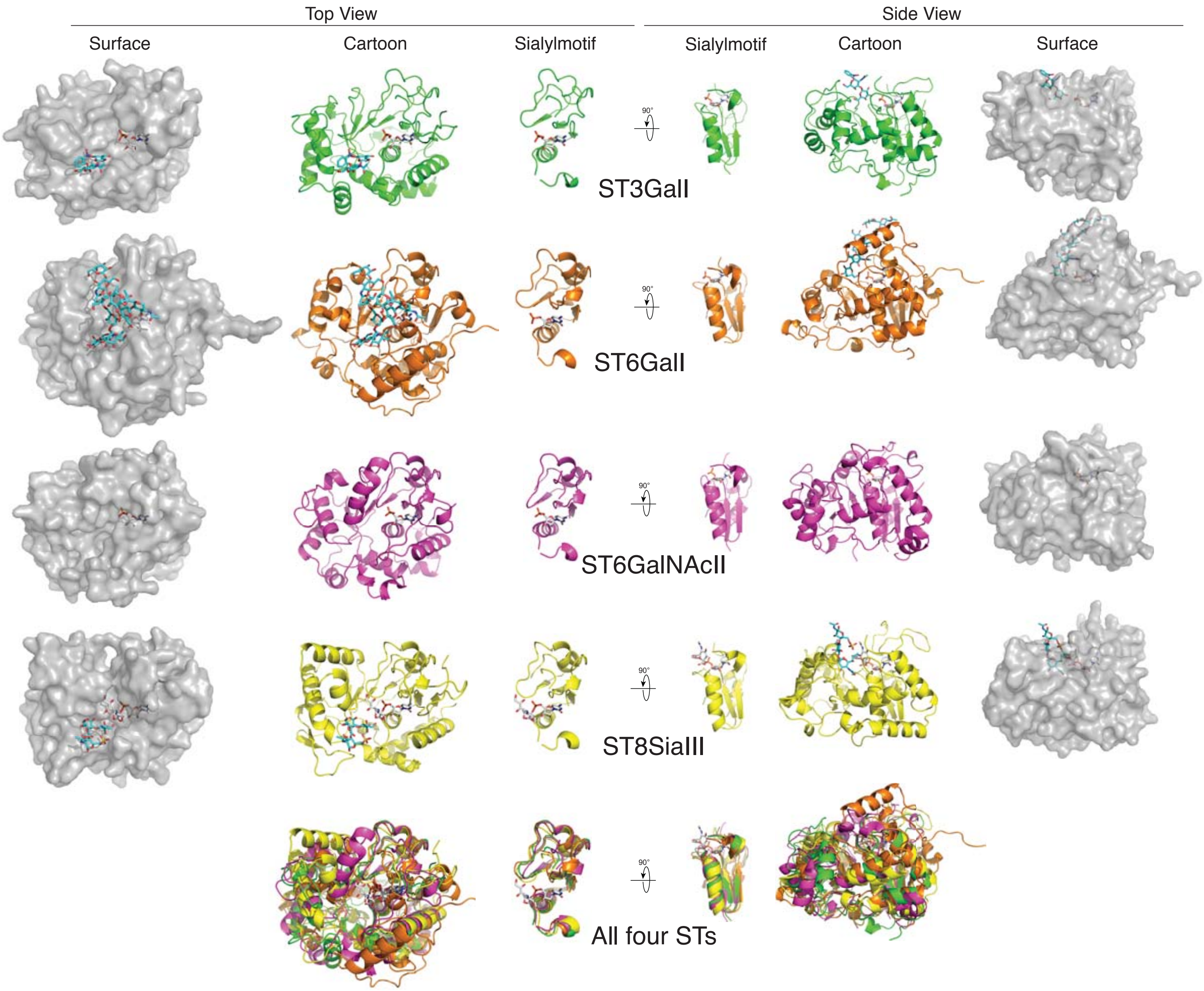
Supplementary Figure 7. ST6GalNAcII expression, purification, and tag/glycan cleavage. A diagrammatic representation of the recombinant ST6GalNAcII fusion protein coding region is shown at the top of the figure. This fusion protein has an NH_2-terminal signal sequence followed by an 8xHis tag, AviTag, superfolder GFP, TEV protease cleavage site, and the catalytic domain of ST6GalNAcII containing three *N*-glycan consensus sequons (see Methods for details). Expression of the recombinant product in HEK293S (GnTI$^-$) cells resulted in secretion of the fusion protein into the culture medium (*Crude media*, lower panel), and subsequent Ni^{2+}-NTA purification yielded a highly-enriched enzyme preparation (*IMAC1 elution*, lower panel). Cleavage of the enzyme with TEV protease and EndoF1 resulted in removal of the tag sequences and glycans, leaving only a single GlcNAc residue attached to the peptide backbone (*TEV + EndoF1*, lower panel), with the exception of one partially EndoF1-resistant *N*-glycan indicated by the doublet representing two nearly co-migrating forms of the ST6GalNAcII catalytic domain (*ST6GalNAcII cat domain*). Ni^{2+}-NTA chromatography separated the unbound ST6GalNAcII catalytic domain (*IMAC2 runthru*) from the bound tag sequences, TEV protease, and EndoF1, as the latter were all His-tagged. The enzyme was further purified over Superdex-75, and we observed an overall yield of ~35% in the pooled peak fractions of purified ST6GalNAcII (*Gel filtration pool*) isolated from the crude culture medium, based on recovery of the ST6GalNAcII catalytic domain.



Supplementary Figure 8. Sequence alignment and secondary structures of mammalian sialyltransferases. The full length primary sequences all twenty human GT29 sialyltransferases and porcine ST3GalI were aligned using the MUSCLE alignment algorithm provided by Geneious software (version 6.1.5, Biomatters, Auckland, New Zealand). The dendrogram derived from the sequence alignment (*upper panel*) shows the distinct clades of the ST6GAL, ST8SIA, ST3GALNAC (two sub-clades) and ST3GAL (two sub-clades) enzyme subfamilies. Proteins that have published structures (porcine ST3GalI, PDB 2WNB[1]; human ST6GalI, PDB 4JS2[2]; human ST8SiaIII, PDB 5BO9[3], and the present ST6GalNAcII structure) are indicated in bold red text in the dendrogram. The aligned primary sequences (*lower panel*) were used as the framework to display the respective secondary structures derived from the corresponding protein structures (α-helices indicated by red cylinders and β-strands indicated by yellow arrows) above each respective primary sequence. Conserved sequence elements (Sialylmotifs L, S, III, and VS) that comprise the Rossmann fold scaffold underlying the CMP-Neu5Ac binding site[23] are indicated by the blue boxes. Residues that are identical (*) and similar (: or .) are indicated under the sequence alignment. Predicted transmembrane domain sequences are indicated in red text. Sequences and structures outside the sialylmotif elements are highly variable among the sialyltransferases and comprise loop regions and secondary structure elements that are unique to each member of the GT29 sialyltransferase subfamily.



Supplementary Figure 9. Structure and putative membrane tethering of human ST6GalNAcII. The structure of human ST6GalNAcII represents a single Rossmann-like (GT-A variant 2) fold with 6 β-strands in the domain core and 17 α-helical segments and loop regions (α-helices and β-sheets are labeled from the NH_2-terminus). While the crystallographic asymmetric unit contains six protomers with disordered loops of varying lengths, the image represents Chain C, containing a 7 residue disordered loop between helix α5 and strand β1 (magenta line). The orientation of this figure is a 90° rotation relative to **Figure 4C**. The bound donor analog, CMP, is shown in yellow stick representation. As in **Figure 4**, the predicted catalytic base, His351, is shown in cyan stick representation. Predicted glycosylated amide side chains of Asn85, Asn130, and Asn161 are indicated by a white stick representation and a magenta sphere for the amide nitrogen where the *N*-glycan is attached. Four Cys residues participating in a disulfide bond are shown in white stick representations with yellow sulfur atoms shown as spheres, as in **Figure 4**. The full length ST6GalNAcII is a transmembrane, Golgi-localized enzyme *in vivo* and contains a non-conserved 37 amino acid linker "stem region" between the transmembrane span and the catalytic domain represented as a green line.



Supplementary Figure 10. Comparisons of mammalian sialyltransferase structures. The four known mammalian GT29 sialyltransferase structures (porcine ST3GalI, PDB 2WNB[1]; human ST6GalI, PDB 4JS2[2]; human ST6GalNAcII, the present study; and human ST8SiaIII, PDB 5BO9[3] and) were aligned using Coot[4] and the structures of the respective proteins were displayed as either a surface (*Surface*) or cartoon (*Cartoon*) representations of the GT-A (variant 2) Rossmann-like folds. Two orientations are shown for each protein, including a face-on view of the respective enzyme active site (Top View) or a Side View rotated 90°. The bound CMP donor analogs (ST3GalI, ST6GalI, and ST6GalNAcII) or CMP-Neu5Ac donor (ST8SiaIII) are shown with white stick representation in the respective structures. Bound glycan acceptor complexes were identified for ST3GalI, ST6GalI, and ST8SiaIII and the acceptor complexes are shown in cyan stick representation. A subset of each protein structure comprises the sialylmotif sequence elements (**Supplementary Fig. 8**) and these structural elements were extracted from the overall structures and are also shown in top and side views as a cartoon secondary structure representation. The sialylmotif elements comprise the scaffold that underlies the CMP-Neu5Ac binding site and the respective donor analog (CMP) or donor (CMP-Neu5Ac) is shown in white stick representation in association with each of the sialylmotif structures. The panels at the bottom of the figure show an overlay of the aligned structures and sialylmotif sequences of all four proteins (All four STs) illustrating the high structural similarity for the respective sialylmotifs, while the remainder of the protein structures for each of the sialyltransferases were quite distinct.

References

1. Rao, F. V. et al. Structural insight into mammalian sialyltransferases. *Nat Struct Mol Biol* **16**, 1186-1188 (2009).
2. Kuhn, B. et al. The structure of human alpha-2,6-sialyltransferase reveals the binding mode of complex glycans. *Acta Crystallogr D Biol Crystallogr* **69**, 1826-1838 (2013).
3. Volkers, G. et al. Structure of human ST8SiaIII sialyltransferase provides insight into cell-surface polysialylation. *Nat Struct Mol Biol* **22**, 627-635 (2015).
4. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* **60**, 2126-2132 (2004).

Supplementary Data Set 1: Summary of human glycosylation enzyme fusion protein expression strategies and expression results for production in HEK293 cells and BEVS (see separate Supplementary Data Set 1.xlsx Excel spreadsheet). Protein sequences for human GTs and GHs were collated from CAZy, NCBI, and the primary literature to create the table of glycosylation enzymes for human glycosyltransferases, glycoside hydrolases, and other glycan modifying enzymes. Glycosyltransferases (GTs) and glycoside hydrolases (GHs) are listed by CAZy family designation. Database index information is provided for each gene product including gene symbol, alternative names (synonym), and additional annotation information (UniProt accession, GenBank protein and DNA reference sequence annotation, GeneID, and accession number for Mammalian Gene Collection clones). The Description field contains the protein title from the UniProt entry. The domain structure column lists the transmembrane topology of the protein (N-term, C-term TMD or multipass TMD), if the coding region contains an NH_2-terminal signal sequence, COOH-terminal KDEL sequence, or if the protein is cytosolic in localization). The template columns indicate whether the protein coding region was generated by PCR from an MGC clone template or a human cDNA source, or by gene synthesis. Expression in mammalian cells employed the pGEn2 or pGEc2 GFP fusion vectors and the GFP fluorescence of the recombinant fusion protein in the media (FU secreted) or cells (FU cells) was measured and converted to values of mg/ml as described in Methods. The efficiency of secretion was listed as a percentage of total protein expression (% secretion). A qualitative description of cell associated or secreted fusion protein in HEK293 (based on GFP fluorescence) and BEVs (based on immunoblot intensity) is indicated in columns labeled as "Level cell" and "Level sec".

Supplementary Data Set 2. DNA sequences of human glycosylation enzyme expression constructs (see separate Supplementary Data Set 2.xlsx Excel spreadsheet). The DNA coding sequences employed for the expression of human glycosyltransferases, glycoside hydrolases, and other glycan modifying enzymes are listed for each gene as NH_2-terminal and COOH-terminal fusion constructs including flanking attL recombination sequences. These are the sequences found in the Gateway donor clones for the respective enzyme expression constructs. Gateway recombination into the respective mammalian or baculovirus destination vectors yields the final expression construct fusions for production of the enzymes in the HEK293 and insect host cell systems. Glycosyltransferases (GTs) and glycoside hydrolases (GHs) are listed by CAZy family designation. Database annotation information includes gene symbol and UniProt accession code. Additional annotation information for each gene can be found in **Supplementary Data Set 1**. The domain structure column lists the transmembrane topology of the protein (N-term TMD, C-term TMD, multipass TMD, or internal TMD, or if the coding region contains an NH_2-terminal signal sequence, a COOH-terminal KDEL sequence, or if the protein is soluble and cytosolic in localization). The template source columns indicate whether the respective protein coding region was generated by PCR from a human MGC clone template or a human cDNA source, or if the coding region was generated by codon optimized gene synthesis (see Methods). The sequences of the attL1 and attL2 recombination sites were used in the Gateway LR recombination to transfer the coding regions into the respective fusion protein expression vectors. The columns labeled as "TEV site and coding region" contain the full length or truncated enzyme coding region and flanking TEV protease recognition site sequence (as an NH_2- or COOH-terminal fusion) that was used in the generation of the respective expression construct.