Europe PMC Funders Group
Author Manuscript
Nat Chem Biol. Author manuscript; available in PMC 2018 August 01.

Published in final edited form as:
Nat Chem Biol. 2018 February ; 14(2): 156–162. doi:10.1038/nchembio.2539.

Human glycosylation enzymes for enzymatic, structural and functional studies

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 Prabahkar**[1], **Roy Johnson**[1], **Mitche dela Rosa**[1], **Christoph Geisler**[2], **Alison V. Nairn**[1], **Sheng-Cheng Wu**[1], **Liang Tong**[4], **Harry J. Gilbert**[1], **Joshua LaBaer**[3], and **Donald L. Jarvis**[2,*]

[1]Complex Carbohydrate Research Center, University of Georgia, Athens, GA

[2]Department of Molecular Biology, University of Wyoming, Laramie, WY 82071

[3]Biodesign Institute, Arizona State University, Tempe, AZ 85287

[4]Department of Biological Sciences, Northeast Structural Genomics Consortium, Columbia University, New York, New York 10027

Abstract

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 due to 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, sulfotransferases, and 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, the sialyltransferase ST6GALNAC2. Many enzymes were

[*]To whom correspondence should be addressed: Kelley W. Moremen, Complex Carbohydrate Research Center, University of Georgia, Athens, GA 30602, Phone: 706-542-1705, Fax: 706-542-1759, moremen@uga.edu, Donald L. Jarvis, Department of Molecular Biology, University of Wyoming, Laramie, WY 82071, Phone: (307) 766-4282, DLJarvis@uwyo.edu.

[*]This research was supported by NIH grants P41GM103390 (to K.W.M.), P01GM107012 (G.J. Boons, 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 all of the glycoenzyme truncation and fusion constructs and supervised all expression efforts in mammalian cells; D.L.J. supervised the preparation of all baculovirus constructs and expression efforts in insect cells; A.V.N. generated all target mammalian glycogene lists including all gene and protein annotations; J.L. and J.S. designed primer strategies and executed high-throughput amplification of glycogene coding regions and Gateway recombination into donor and DEST vectors; A.R. and M.R. generated coding regions from human cDNAs, generated all mammalian DEST 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.M., Z.G., D.C., S.W., J.-Y.Y., L.M., P.P., and R.J. characterized expression of glycoenzymes in mammalian cells; C.G. designed baculovirus DEST expression vectors; S.-C.W. and H.G. designed and generated fusion protein constructs for expression in bacteria. L.M. expressed and purified recombinant ST6GALNAC2 for structural studies. F.F. and L.T. performed structural studies on ST6GALNAC2.

Competing financial interests: Donald L. Jarvis is the President and Christoph Geisler is now an employee of GlycoBac, LLC, a biotechnology spinout that focuses on insect host cell improvements, but that could conceivably profit from the results described herein. The other authors have no competing financial interest.

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

produced at high yields and 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 pathology1,2. These interactions are strikingly diverse and include molecular recognition events that drive cell surface signaling, cell adhesion, and modulation of receptor half-lives and dynamics, among others2. Terminal glycan structures also direct cellular targeting and clearance of circulating glycoproteins, including most recombinant human therapeutics, and define many pathogen and toxin tropisms2.

Numerous challenges remain in the study of glycan structures and their contributions to physiological processes. A recent study3 concluded that new and transformative methods are required to fill the technology gaps necessary for carbohydrate-based applications in health, energy, and materials sciences. The roadmap goals advanced for technology development included the enzymes involved in glycan synthesis, modification, and catabolism. These enzymatic reagents were emphasized not only because they would advance our understanding of fundamental glycan biosynthetic and catabolic processes3, but also for their utility in novel chemoenzymatic glycan synthesis, which has numerous applications in biological and biomedical studies, as standards for glycomic analysis, for installation and manipulation of glycan structures in complex biological systems, and use in development of glycan-related therapeutic applications, among others.

Glycan biosynthesis occurs predominantly within the secretory pathway1, where glycosyltransferases (GTs[a]) 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, abundance, and specificities of the enzymes ("glycoenzymes") involved in glycan synthesis and catabolism4. Unlike RNA transcription and protein translation, these complex metabolic pathways are not template-driven. Thus, the keys 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 catabolism4 in generating the full collection of ≥7,000 vertebrate oligosaccharide structures5. These enzymes include ~200 GTs and ~80 GHs (Supplementary Tables 1 and 2), which have been classified into 44 and 28 discrete

[a]The abbreviations used are: GT, glycosyltransferases; GH, glycoside hydrolase; ST, sulfotransferase; CAZy, Carbohydrate Active Enzymes database; PDB, Protein Data Bank; TMD, transmembrane domain; TEV, tobacco etch virus; EndoF1, endoglycosidase F1; SEEL, selective exoenzymatic labeling; DEST vector, Gateway destination vector; MGC, Mammalian Gene Collection; CMV, cytomegalovirus; AviTag, peptide recognition sequence for *in vitro* biotinylation by the BirA biotin protein ligase; GFP, green fluorescent protein; BEV, baculovirus expression vector; IMAC, immobilized metal affinity chromatography; WPRE, woodchuck hepatitis virus posttranscriptional regulatory element; BGH, bovine growth hormone.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

sequence families, respectively, in the Carbohydrate Active Enzymes (CAZy) database6. Members of a given GT family generally have a similar protein fold, but frequently have distinct acceptor substrate profiles and can have different sugar donor specificities7,8. Thus, subtle differences in protein structures between enzymes with otherwise similar protein folds produce glycoenzymes with unique catalytic specificities. These differences in substrate recognition among the GTs expressed in a given cell are critical factors determining the diversity of glycan structures in animal systems4.

Many challenges remain in advancing our understanding of vertebrate glycosylation machinery at a molecular level. In general, only ~1% of the enzymes in a given GT or GH family have been characterized6 and structural studies have been performed on only a minor subset of these enzymes. For mammalian glycoenzymes, this reflects challenges associated with their recombinant expression, as evidenced by the deposition of only 28 of 202 vertebrate GT structures in the PDB6,8. A few mammalian GTs have been expressed in bacteria as either soluble truncated catalytic domains9 or with *in vitro* refolding10, but these examples are extremely rare8. More commonly, mammalian GT expression in bacteria yields non-functional protein aggregates because these enzymes require glycosylation, disulfide bonding, and chaperones that occur uniquely within the eukaryotic secretory pathway11. Large-scale efforts to produce glycoenzymes in cell free systems have not been reported. Thus, the most successful approach has been to produce recombinant glycoenzymes in eukaryotic hosts, but this has been restricted to small-scale expression studies for relatively focused biochemical goals8. Broad access to large quantities of the mammalian glycoenzymes would be greatly enhanced by the development of a unified, high-throughput expression platform that could be applied to all glycoenzymes in a modular format.

We describe here an expression vector library encoding all known human glycoenzymes for production in either mammalian (HEK293) cells12 or baculovirus-infected insect cells13. This is the first universal platform for the production of all human GTs, GHs, sulfotransferases (STs), and various other glycan modifying enzymes and employs the transfer of glycosylation enzyme coding cassettes into custom-designed plasmid or baculovirus expression vectors for recombinant glycoenzyme production in mammalian cells, insect cells, *E. coli,* or other hosts. We demonstrate the successful expression and secretion of a large set of mammalian 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, as well as glycan synthetic applications. We also demonstrate the purification of a subset of the GTs as a proof of concept and a complete enzymatic characterization and structural determination with the sialyltransferase, ST6GALNAC2, in complex with the donor analog CMP. Thus, we demonstrate that the expression vector library and resulting recombinant products provide a significant resource enabling detailed studies on these enzymes, as well as a source of novel enzymatic tools for a wide array of biochemical and biomedical applications.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Results

Design of enzyme coding regions

A comprehensive list of >700 human glycoenzymes and proteins collected during prior glycogene transcript profiling efforts4 was used to extract a subset of 339 genes comprising all human GTs, GHs, STs, and numerous other glycan modifying enzymes (Supplementary Tables 1 and 2). Based on annotations in UniProt14, this list includes 187 protein coding regions with an NH_2-terminal transmembrane domain (TMD; 56.2%) and 22 with a COOH-terminal TMD (6.6%). In addition, 28 were multipass or internal TMD enzymes (8.4%), 72 contained NH_2-terminal signal sequences (21.6%), and 24 were predicted to be cytosolic (no TMD or signal peptide, 7.2%).

We used a modular expression vector design to enable transfer of truncated enzyme coding regions into custom, host-specific expression vectors by Gateway recombination15. Glycoenzyme coding regions were truncated at their NH_2- or COOH-termini and fused in-frame at their respective truncation sites with a sequence encoding a tobacco etch virus (TEV) protease recognition and cleavage site16. These coding region cassette sequences were then placed into a plasmid vector containing flanking Gateway recombination sites (Gateway “donor” clones, Fig. 1 and Supplementary Fig. 1). Custom expression vectors were then prepared harboring host-specific promoter elements, fusion protein sequences and complementary Gateway recombination sites (Gateway destination “DEST” vectors; Fig. 1 and Supplementary Figs. 1 and 2). Recombination of the donor cassette coding regions with host-specific DEST expression vectors yielded the final collection of expression vectors containing host-specific transcription and translation elements that drive fusion protein expression with vector-encoded tags.

For most coding regions, a single construct was designed as either an NH_2- or COOH-terminal fusion (Fig. 1 and Supplementary Fig. 1). For example, the most common native topology for the mammalian glycoenzyme collection was an NH_2-terminal TMD that initiates co-translational membrane insertion and provides membrane tethering and localization within the secretory pathway. The COOH-terminal ends of the coding regions contain the corresponding catalytic domains, which are topologically inside the lumen of the secretory pathway compartment. The TMD anchors are often dispensable for catalysis and can be deleted by truncation and replaced with other protein sequences, such as fusion tags, at the NH_2-terminal end of the coding region as long as the proteins are expressed within the context of the eukaryotic secretory pathway11. Entry into the secretory pathway is mediated by an exogenous NH_2-terminal signal sequence, which is part of the fusion tag assembly, and following entry, the catalytic domains are eventually secreted into the culture medium17. Thus, replacement of the NH_2-terminal TMD with a TEV protease cleavage site16 and transfer of these coding regions into Gateway donor plasmids harboring flanking *att*B recombination sites produces a library of glycoenzyme truncations (Fig. 1 and Supplementary Fig. 1) that can be transferred to host-specific expression vectors harboring additional upstream fusion sequences, including an NH_2-terminal signal sequence (Supplementary Figs. 2 and 3).

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Constructs encoding multipass transmembrane or cytosolic enzymes or enzymes with NH_2-terminal signal sequences were designed to include their respective full length or truncated coding regions, with a COOH-terminal TEV protease cleavage site appended in place of the termination codon during transfer to a Gateway donor plasmid (Fig. 1, Supplementary Fig. 1 and Supplementary Tables 1-3). The fusion sequences were then extended downstream of the TEV site upon transfer to the final expression vectors via Gateway recombination (Supplementary Figs. 1 and 2). Thus, all coding regions were initially generated as a library of Gateway donor clones encoding either NH_2-terminal or COOH-terminal TEV site fusions (Fig. 1 and Supplementary Fig. 1) and subsequently transferred to custom Gateway DEST vectors specific for expression in each host system (Supplementary Figs. 2 and 3).

Strategies for capturing coding regions as Gateway donor clones involved PCR with Mammalian Gene Collection18 clones or human tissue cDNAs as templates, followed by Gateway BP recombination15 into donor vectors (Fig. 1 and Supplementary Fig. 1). A total of 255 of the glycoenzyme coding regions (66%) were captured using this strategy (**Table 1**), while the remaining 132 glycoenzyme coding regions were synthesized with human codon optimization (Fig. 1, Supplementary Fig. 1 and Supplementary Tables 1-3).

Fusion protein strategies for mammalian cell expression

Custom Gateway-adapted DEST vectors for mammalian cell expression included a cytomegalovirus (CMV) promoter and alternative fusion protein strategies. Each vector had a Gateway selection cassette comprising *ccd*B and Cm^R genes flanked by *att*R sites15 that was replaced upon insertion of target gene coding regions from Gateway donor clones (Fig. 2 and Supplementary Figs. 1-3). Three different custom mammalian NH_2-terminal fusion vectors were developed using alternative tagging strategies (Supplementary Fig. 2). All three vectors contained an NH_2-terminal signal sequence followed by an 8xHis tag and other fusion sequences. The pGEn1 vector was designed to add a StrepII tag19 after the His tag and before the Gateway recombination site. Alternatively, the pGEn2 vector added an AviTag20 after the His tag followed by a “superfolder” GFP domain21. Finally, pGEn3 was designed to add the AviTag, GFP domain, and an additional Ig Fc domain22 after the His tag to facilitate homo-dimerization of the fusion protein. We also generated a parallel set of custom COOH-terminal fusion vectors (Supplementary Fig. 2). pGEc1 was designed to append a COOH-terminal 8xHis and StrepII tag, by analogy to pGEn1, whereas pGEc2 was designed to append COOH-terminal GFP, AviTag, and 8xHis tags, by analogy to pGEn2 ((Supplementary Fig. 2). These designs enabled transfer of each of the glycoenzyme coding regions into multiple NH_2- or COOH-terminal fusion protein contexts (Supplementary Fig. 2).

Target protein expression, secretion, and quantitation in mammalian cells

To test glycoenzyme production and secretion efficiencies provided by the fusion vector series, a collection of pGEn1, pGEn2 and pGEn3 vectors encoding fourteen different GT29 sialyltransferase coding regions17 were used for transient expression in HEK293 cells (Fig. 2). Most fusion proteins were secreted at significantly higher levels with pGEn2 and pGEn3, as compared to pGEn1, suggesting the GFP fusion domain enhanced protein folding and secretion. While we observed minor differences in overall protein production, secreted

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

protein levels were essentially identical with pGEn2 and pGEn3, based on GFP fluorescence (Fig. 2). Thus, pGEn2 and pGEc2 were chosen for further screening and characterization of the full collection of recombinant products as GFP fusions.

The GFP fusion domain in the pGEn2 and pGEc2 vectors also provided an effective tag for protein quantitation during expression and purification (Figs. 2 and 3, and Supplementary Table 2). The overall expression level (cells + media) was ≥10 mg/L for 94% of the recombinant products (Fig. 3). In contrast, only 53% of the proteins were secreted at a level ≥10 mg/L. However, many of the cytosolic and transmembrane proteins were not designed for secretion. When we considered only proteins designed for secretion, 65% were secreted at a level ≥10 mg/L. An interesting observation from these studies was that there was no apparent correlation between the expression level and secretion efficiency (Fig. 3, upper panel), as some well-expressed proteins were poorly secreted and many poorly expressed proteins were efficiently secreted.

To illustrate the variability of enzyme expression and secretion, we further examined twenty GT29 sialyltransferases designed using a similar fusion protein strategy. These pGEn2 constructs were expressed and cell and media samples were subjected to fluorescence measurements (Fig. 3) and SDS-PAGE (Supplementary Fig. 4). The data indicated a subset of the enzymes (ST6GAL1, ST3GAL1, ST3GAL3, and ST3GAL4) were exceptionally well expressed and secreted. A second category (e.g. ST3GAL5) were well expressed, but poorly secreted. A third category of enzymes (e.g. ST3GAL6 and ST6GALNAC3) had overall poor expression. Finally, a fourth category (e.g. ST6GAL2 and ST6GALNAC1) were well expressed and secreted, but subject to proteolysis between the catalytic domain and GFP fusion during expression (Supplementary Fig. 4). Thus, the behavior of the expression products varied significantly even within a single family of structurally similar enzymes. Variations in protein expression and secretion were also seen within other protein families, as highlighted by the enzymes from GH27, GT29, and GT31 in Figure 3b, but also observed in many other GT, GH, and ST families (Supplementary Table 2).

Fusion protein strategies for baculovirus expression

We also generated baculovirus expression vectors (BEVs) encoding most of the glycoenzymes as either NH_2- or COOH-terminal fusions using a strategy analogous to that used for the mammalian expression plasmids (Supplementary Fig. 3). Custom-designed baculoviral DEST vectors were designed containing the polyhedrin promoter, fusion protein sequences, and a Gateway recombination and selection cassette comprised of herpes simplex virus 1 thymidine kinase and *E. coli* β-galactosidase genes flanked by *att*R sites (Supplementary Fig. 3). BEVs contained either short NH_2-terminal fusions comprised of a signal sequence, 8xHis tag, and StrepII tag, analogous to pGEn1, or longer NH_2-terminal fusions comprised of a signal sequence, 8xHis tag, AviTag, and GFP domain, analogous to pGEn2. The COOH-terminal fusion BEVs (Supplementary Fig. 3) encoded StrepII and 8xHis tags, analogous to pGEc1. The same glycoenzyme donor clones used for the mammalian expression constructs were used with the custom baculovirus DEST vectors to generate the final BEVs, which encoded glycoenzymes with either NH_2-terminal or COOH-terminal fusions. Since generation of most of the BEV stocks was underway well before we

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

determined the GFP fusion cassette was the preferable strategy for the mammalian expression vectors, screening efforts with BEVs focused on the shorter His/StrepII fusions, analogous to pGEn1. Protein expression and secretion in BEV-infected insect cells was profiled by immunoblotting with anti-His tag antibodies and scored using semi-quantitative values, as indicated in Figure 4 and Supplementary Table 2.

Comparison of insect and mammalian cell expression

Comparison of protein expression and secretion in BEV-infected insect and transiently transfected mammalian cells revealed a similar overall trend (Fig. 4), in which fusion proteins that were well expressed and secreted in transiently transfected mammalian cells were also generally well expressed and secreted in BEV-infected insect cells. However, there was considerable scatter in the data, indicating one of the two expression hosts can in some cases more effectively express specific protein coding regions.

Since different fusion protein strategies had been used to produce expression vectors for human and insect hosts, we generated BEVs encoding twenty GT29 sialyltransferases with NH_2-terminal large tag fusions for direct comparison with equivalent mammalian GT29 expression constructs (Supplementary Fig. 4). Secreted fusion protein levels from BEV-infected insect cell cultures were significantly lower compared to equivalent fusion proteins from transiently transfected HEK293 cells. However, when specific protein productivities (GFP fluorescence units/cell/day) were calculated, sialyltransferase fusion protein expression levels in BEV-infected insect cells were only ~3-10-fold lower than in transiently transfected HEK293 cells (Fig. 5).

Immunoblotting and fluorescence data indicated the sialyltransferase GFP fusion secretion efficiencies were quite variable in both hosts, with some proteins efficiently secreted while others were largely retained in the cells. Secreted protein levels were highly correlated for half of the fusion proteins in the two recombinant hosts (R^2=0.99; Fig. 4). However, eight of the sialyltransferase fusion proteins were more highly secreted in mammalian cells and two were more highly secreted in the BEVs (Fig. 4b). Thus, the trend was similar to the expression data for the larger collection of the glycoenzymes (Fig. 4a), as a subset of proteins was more effectively secreted in either human or insect cells, but most were secreted similarly in the two different expression hosts.

Purification of human glycoenzymes from mammalian and insect cell cultures

The 8xHis tags associated with the secreted enzyme fusion proteins allow direct affinity purification of the recombinant products from the conditioned media by immobilized metal affinity chromatography (IMAC). To demonstrate enzyme purification from both expression host systems, we used IMAC to isolate the GH29 sialyltransferase GFP fusion proteins from mammalian or insect cell-free media. The results showed we could obtain >50% yields for the more highly expressed enzymes, and could readily purify even some of the more poorly expressed proteins by scaling up the culture volumes (Fig. 5). In the mammalian expression host, a majority of the GT29 sialyltransferases could be purified at >10 mg/L of transfected cell culture, indicating a majority of the enzymes could be produced and purified at levels required for biochemical and structural studies.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Structure determination ST6GALNAC2

To demonstrate the utility of the expression vector library for glycoenzyme characterization, we scaled up production of one of the sialyltransferase GFP fusions (ST6GALNAC2-pGEn2) in HEK293S (GnTI-) cells23 using a selenomethionine labeling protocol established in prior structural studies on rat ST6GAL117. IMAC purification of the secreted fusion protein was followed by concurrent *in vitro* cleavage with TEV protease and endoglycosidase F1 (EndoF1). Further IMAC purification and gel filtration yielded the ST6GALNAC2 catalytic domain devoid of all tag sequences with *N*-glycans trimmed to a single GlcNAc residue (Supplementary Fig. 5). Sialyltransferase kinetic parameters for the TEV/EndoF1 cleaved product and the GFP fusion were similar (Supplementary Table 4), indicating the GFP fusion did not alter enzyme activity. The catalytic domain was concentrated and crystallized, leading to diffraction to 3.1 Å, and the structure was solved by SAD phasing, revealing six molecules in the asymmetric unit (Supplementary Table 5). Additional diffraction data were obtained to 2.35 Å for a labeled co-complex of ST6GALNAC2 with the donor analog, CMP, and the structure was solved by molecular replacement (Fig. 6 and Supplementary Table 5). This revealed a single Rossmann-like (GT-A variant 2) fold similar to other known GT29 sialyltransferases9,17,24,25, especially within the conserved “sialylmotif” sequence elements (Supplementary Fig. 6). The bound CMP was clearly evident in the electron density at a position equivalent to CMP bound in other sialyltransferases (Fig. 6), indicating the sialylmotif scaffold underlying the CMP-NeuAc binding site is conserved17. Significant structural differences were found outside the sialylmotif elements, as compared to other sialyltransferases, reflecting the minimal primary sequence similarity among the sialyltransferases in the loop regions and secondary structure elements involved in acceptor substrate recognition (Supplementary Fig. 6). While structures have been reported for three of the four vertebrate GT29 sialyltransferase subfamilies that generate NeuAc-α2,6Gal17,24, NeuAc-α2,3Gal9, and NeuAc-α2,8NeuAc25 linkages, the present ST6GALNAC2 structure is the first reported for a NeuAc-α2,6GalNAc subfamily member17. Efforts to identify enzyme-CMP-acceptor structural co-complexes and perform kinetic analyses of site directed mutants are presently underway and will be reported elsewhere.

Discussion

The diversity of vertebrate cell surface and secreted glycans are produced by complex biosynthetic pathways1,2,4. The overall purpose of the present study was to produce an expression vector library enabling production of the 339 enzymes involved in human glycan synthesis, modification, and catabolism for enzymatic, structural and functional studies. We initially examined the potential for recombinant glycoenzyme expression in *E. coli,* but found that all the recombinant products were insoluble aggregates. As a result, we focused on eukaryotic expression in HEK293 and insect cells. These two expression hosts were chosen because they are both amenable to high-throughput profiling of protein expression and production can be scaled up in larger volume cultures12,13.

A variety of fusion tag strategies were tested. Inclusion of the “superfolder” GFP domain led to improved protein production and secretion for many of the GTs tested in both human and

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

insect cell systems (Fig. 2 and Supplementary Figs. 4B and 4C), as compared to shorter fusion tags. The GFP fusion also enabled direct detection and quantitation of recombinant products during expression and purification. Thus, all of the 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 could be purified at multi-milligram levels from the culture media. The efficiency of secretion was surprisingly variable and did not correlate with overall expression level, transmembrane topology or enzyme sequence family. This variability was highlighted for the GT29 sialyltransferases (Fig. 4 and Supplementary Fig. 4), but was also observed with other GT families (Supplementary Table 2). These data suggest 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 also performed in BEVs and a common trend in overall expression and secretion levels was observed in both insect and mammalian recombinant hosts. When enzymes with equivalent tags were compared, half of the enzymes were secreted at similar levels by the two hosts, while the remainder were more highly secreted by either mammalian or insect cells. These data suggest host-specific factors can also contribute to secretion efficiency for some recombinant products.

Additional factors can also contribute to the efficiency of protein production in eukaryotic cells. *In vivo,* many of the full length GTs are transmembrane proteins that can assemble into homo-dimers or hetero-oligomeric complexes with other proteins26 and sequences proximal to the TMD can also play roles in oligomer formation27. 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. Initial efforts to co-express potential oligomeric partners or to enforce homo-dimerization using the Ig Fc domain in the pGEn3 vector did not lead to enhanced protein secretion, but is still being pursued.

Overall, many of the glycoenzymes designed for secretion could be readily purified from the culture media at high yields using IMAC (Fig. 5). 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 ST6GALNAC2 (Supplementary Table 4), indicated that the enzymes had comparable activity as GFP fusion proteins or released catalytic domains17,28. As the construct library was being completed, we further characterized several of these expression products and the results effectively identified GT substrate specificities28–30, corrected the glycan linkage elaborated by a GT31, resulted in chemoenzymatic synthesis of novel glycan structures32, and supported selective exoenzymatic labeling (SEEL)33 of cell surface glycans, demonstrating the utility of our expression vector library as a tool for various glycoenzyme studies and applications in glycobiology.

As a proof of concept for the utility of the HEK293 cell expression platform for protein structural studies, we focused on ST6GALNAC2, an enzyme that adds α2,6 linked NeuAc

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

residues to the GalNAc residue on Core 1 *O*-glycans34. The enzyme was produced and purified as a GFP fusion protein, tags were removed with TEV protease, *N*-glycans were enzymatically trimmed, and the product was further purified, crystallized, and used for structure determination. This effort was enhanced by expressing the enzyme in HEK293S (GnTI-) cells23, which allowed glycan cleavage with EndoF117, and metabolic incorporation of selenomethionine, which facilitated phasing of diffraction data by single wavelength anomalous diffraction17. Thus, a unified workflow was achieved for production and purification of glycoenzymes for both enzymatic and structural studies. We have performed structural studies on additional glycosylation enzymes as the expression vector library was being completed. These efforts have elucidated six additional mammalian GT structures (many in complex with donor and acceptor analogs, manuscripts in preparation), as well as the structure of the extracellular domain of a calcium-sensing receptor with a novel bound ligand35, and two plant xyloglucan modifying enzymes (manuscripts in preparation). Thus, it is clear that the constructs, expression systems, and workflows described here for glycoenzyme production and purification will provide novel insights into substrate recognition and catalysis for this diverse and important set of glycoenzymes and will provide transformative reagents that will continue to expand our knowledge in glycochemistry and glycobiology.

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 cells4. 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 Table 2).

Truncation and fusion protein strategies for glycoenzyme expression

Each glycoenzyme coding region on the target list was examined for sequence features in the UniProt database14 to determine the presence of TMDs, NH_2-terminal signal sequences, COOH-terminal KDEL ER retention sequences36, 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 sequence16 during transfer to Gateway donor vectors (Supplementary Fig. 1). 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-

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

terminal signal sequence and subsequent NH_2-terminal fusion to the TEV site sequence. 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. 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)18 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 (ThermoFisher Scientific) in a total volume of 50 μl. PCR conditions were denaturation at 94°C for 2 min followed by 15 PCR cycles comprised of 94°C denaturation for 15 seconds, annealing at 57°C for 1 minute 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 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 (ThermoFisher 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 not successfully amplified from MGC clone templates, an alternative amplification approach was employed. A library of human RNAs (FirstChoice Human RNA Survey Panel, ThermoFisher 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 (ThermoFisher Scientific). Test amplifications were performed using respective gene-specific primers, the human cDNAs, and Phusion DNA polymerase reaction mix (ThermoFisher 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 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

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

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 Tables 2 and 3) were synthesized with human codon optimization (GeneArt gene synthesis, ThermoFisher 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 ST6GAL117. This latter vector was generated by gene synthesis (ThermoFisher Scientific) and contains a CMV promoter, artificial intron, woodchuck hepatitis virus posttranscriptional regulatory element (WPRE), and 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 Cm^R genes15 (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 (ThermoFisher 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-mannosidase37, an 8xHis tag, and a StrepII tag19 adjacent to the *att*R1 recombination site (Supplementary Fig. 2). The pGEn2-DEST vector employs the same signal sequence and 8xHis tag, but has an AviTag sequence20 and a “superfolder” GFP domain21 adjacent to the *att*R1 recombination site (Supplementary Fig. 2). The pGEn3-DEST vector has the same signal sequence, 8xHis tag, AviTag and GFP tag as pGEn2-DEST, but also includes an Ig Fc domain22 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 8xHis tag and StrepII tag and the pGEc2-DEST vector contained a GFP domain, AviTag sequence, and 8xHis tag (Supplementary Fig. 2). Each of these selection cassettes and fusion

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

sequences were generated by gene synthesis (ThermoFisher 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 (ThermoFisher Scientific) according to the manufacturer's instructions. The reaction products were transformed into 5-alpha competent *E. coli* (New England Biolabs, Ipswitch, MA), 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 293-F cells (ThermoFisher 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 (ThermoFisher 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 described12. 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 days 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 peptide38, an 8xHis tag, and a StrepII tag, or a "long" tag consisting of the same signal sequence and 8xHis 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

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

short tag consisting of StrepII and 8xHis 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, also described above, 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 (ThermoFisher 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 (ThermoFisher 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 6-well plates. The cells were incubated for 5 h at 28°C, then the transfection mixture was removed and replaced by insect cell growth medium containing 100 μM gancyclovir. The cells were incubated for another 5 days 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 described13. After 7-10 days, at least three plaque purified (PP1) clones with white plaque phenotypes were picked, 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 described13. 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 described13. Briefly, the cells were seeded at a density of 1×10^6 cells/well into 6-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 (ThermoFisher Scientific), as previously described13, to produce the results shown in Supplementary Figures 4B-C.

We also performed mid-scale 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

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

~1×10^6 cells/mL in a shaking incubator set at 125 rpm 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 described13. At 48 h post-infection, the cell-free supernatant was harvested, ultracentrifuged to remove extracellular baculovirus and debris, and the resulting supernatant was used to affinity purify the secreted sialyltransferase products, as previously described13, except we used ProBond nickel chelating resin (ThermoFisher 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 2500 rpm 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 purified recombinant form of a GFP fusion protein expressed in pGEn2 (ROBO1-pGEn239) was used to derive a factor we 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-) cells23 (ATCC) using modifications of workflows based on prior structural studies on rat ST6GAL117. 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 Freestyle 293 expression medium (ThermoFisher Scientific) and 1 volume 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 (ST6GALNAC2-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) 12-24h after transfection, and protein production was continued for a further 4-5 days at 37°C.

In cases where 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 (ThermoFisher Scientific) for 6 h

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

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 days at 37 °C before harvest of the conditioned medium.

Protein purification, deglycosylation, and tag removal employed workflows similar to prior structural studies on rat ST6GAL117. The conditioned culture medium was harvested, clarified by centrifugation, passed through a 0.45 μm filter (Millipore) and adjusted to contain 20 mM imidazole, 20 mM NaCl, and 3 mM sodium phosphate, pH 7.2, and loaded onto a 25 ml 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 described17 were added at ratios of 1:15 and 1:8 relative to the GFP-ST6GALNAC2, respectively, and incubated at room temperature for 3-5 hours, 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 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 ST6GALNAC2 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 described17, except asialofetuin was used as the acceptor. Assays were performed in 50 μl containing 100 mM MES, pH 6.5, 0.1 μg of recombinant ST6GALNAC2, CD73 (0.5 ng/l), 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 1000 μM for kinetic analysis). Following a 30 min incubation at 37 °C, the reaction was stopped by addition of the malachite green phosphate detection reagents and further incubation as described by the manufacturer17. 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 ST6GALNAC2 were initially screened using the high throughput crystallization facility at the Hauptman Woodward Medical Research Institute40. Based on the results, the protein was crystallized

Europe PMC Funders Author Manuscripts

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*41. 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, 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 data set to resolution 2.9 Å was collected at the peak absorption wavelength (0.9792 Å) of selenium at the 24-ID-E beam line of the Advanced Photon Synchrotron (APS). The diffraction images were processed with the HKL2000 package42. The crystals of the apo enzyme belong to space group P1 and there are six protomers in the asymmetric unit (ASU) of the crystal. The selenium sites were determined by the direct method using SHELX43, which located 23 of 30 possible selenium sites. A crude model was subsequently built using SOLVE/RESOLVE44. The majority of the initial model comprising six protomers were manually built with the program XtalView45.

The selenomethionine-labeled ST6GALNAC2 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 Accelator Laboratory (SLAC). The diffraction images were processed with the HKL2000 package46. The crystals of the CMP-complex also belong to space group P1, albeit having different cell parameters as compared with those of the apo enzyme crystal, and there are six protomers per ASU of the crystal. The structure was determined using the molecular replacement method with the program COMO47. All stages of the structure refinement were performed using the crystallographic programs CNS 1.348 and PHENIX49. PHENIX was also used in the last stage of refinement. The statistics for data collection and refinement are shown in Supplementary Table 5.

Availability of constructs

All mammalian glycoenzyme donor constructs and final mammalian expression constructs are archived in the DNASU plasmid repository (dnasu.org). Baculovirus stocks are available directly from the Jarvis laboratory (dljarvis@uwyo.edu). A website has been generated (glycoenzymes.ccrc.uga.edu) summarizing the strategy for construct designs, and provides annotations and sequences of all glycoenzyme constructs in the library, as well as links to construct availability at DNASU.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We wish to thank Farah Samli, Robert Collins, Leslie Stanton, Aaron Petrey, Alexander Yox, Rosemary Kim and Jared Aumiller for technical assistance during these studies.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

References

1. Moremen KW, Tiemeyer M, Nairn AV. Vertebrate protein glycosylation: diversity, synthesis and function. Nat Rev Mol Cell Biol. 2012; 13:448–462. [PubMed: 22722607]
2. Varki A. Biological roles of glycans. Glycobiology. 2017; 27:3–49. [PubMed: 27558841]
3. National, Research Council of the National Academies. Transforming Glycoscience: A Roadmap for the Future. National Academies Press; Washington (DC): 2012.
4. Nairn, AV., Moremen, KW. Handbook of Glycomics. Cummings, R., Pierce, JM., editors. Academic Press; Burlington, MA: 2009. p. 95-136.
5. Cummings RD. The repertoire of glycan determinants in the human glycome. Mol Biosyst. 2009; 5:1087–1104. [PubMed: 19756298]
6. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic acids research. 2014; 42:D490–495. [PubMed: 24270786]
7. Lairson LL, Henrissat B, Davies GJ, Withers SG. Glycosyltransferases: structures, functions, and mechanisms. Annu Rev Biochem. 2008; 77:521–555. [PubMed: 18518825]
8. Handbook of Glycosyltransferases and Related Genes. Edn. 2nd. Springer; Tokyo, Japan: 2014.
9. Rao FV, et al. Structural insight into mammalian sialyltransferases. Nat Struct Mol Biol. 2009; 16:1186–1188. [PubMed: 19820709]
10. Ramakrishnan B, Qasba PK. Crystal structure of lactose synthase reveals a large conformational change in its catalytic component, the beta1,4-galactosyltransferase-I. J Mol Biol. 2001; 310:205–218. [PubMed: 11419947]
11. Paulson JC, Colley KJ. Glycosyltransferases. Structure, localization, and control of cell type-specific glycosylation. J Biol Chem. 1989; 264:17615–17618. [PubMed: 2681181]
12. Subedi GP, Johnson RW, Moniz HA, Moremen KW, Barb A. High Yield Expression of Recombinant Human Proteins with the Transient Transfection of HEK293 Cells in Suspension. J Vis Exp. 2015:e53568. [PubMed: 26779721]
13. Jarvis DL. Recombinant protein expression in baculovirus-infected insect cells. Methods Enzymol. 2014; 536:149–163. [PubMed: 24423274]
14. The UniProt, Consortium. UniProt: the universal protein knowledgebase. Nucleic acids research. 2017; 45:D158–D169. [PubMed: 27899622]
15. Walhout AJ, et al. GATEWAY recombinational cloning: application to the cloning of large numbers of open reading frames or ORFeomes. Methods Enzymol. 2000; 328:575–592. [PubMed: 11075367]
16. Carrington JC, Dougherty WG. A viral cleavage site cassette: identification of amino acid sequences required for tobacco etch virus polyprotein processing. Proc Natl Acad Sci U S A. 1988; 85:3391–3395. [PubMed: 3285343]
17. Meng L, et al. Enzymatic basis for N-glycan sialylation: structure of rat alpha2,6-sialyltransferase (ST6GAL1) reveals conserved and unique features for glycan sialylation. J Biol Chem. 2013; 288:34680–34698. [PubMed: 24155237]
18. Gerhard DS, et al. The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). Genome Res. 2004; 14:2121–2127. [PubMed: 15489334]
19. Schmidt TG, Skerra A. The Strep-tag system for one-step purification and high-affinity detection or capturing of proteins. Nat Protoc. 2007; 2:1528–1535. [PubMed: 17571060]
20. Beckett D, Kovaleva E, Schatz PJ. A minimal peptide substrate in biotin holoenzyme synthetase-catalyzed biotinylation. Protein Sci. 1999; 8:921–929. [PubMed: 10211839]
21. Pedelacq JD, Cabantous S, Tran T, Terwilliger TC, Waldo GS. Engineering and characterization of a superfolder green fluorescent protein. Nat Biotechnol. 2006; 24:79–88. [PubMed: 16369541]
22. Barb AW, et al. NMR characterization of immunoglobulin G Fc glycan motion on enzymatic sialylation. Biochemistry. 2012; 51:4618–4626. [PubMed: 22574931]
23. Reeves PJ, Callewaert N, Contreras R, Khorana HG. Structure and function in rhodopsin: high-level expression of rhodopsin with restricted and homogeneous N-glycosylation by a tetracycline-

Europe PMC Funders Author Manuscripts

inducible N-acetylglucosaminyltransferase I-negative HEK293S stable mammalian cell line. Proc Natl Acad Sci U S A. 2002; 99:13419–13424. [PubMed: 12370423]

24. Kuhn B, et al. The structure of human alpha-2,6-sialyltransferase reveals the binding mode of complex glycans. Acta Crystallogr D Biol Crystallogr. 2013; 69:1826–1838. [PubMed: 23999306]
25. Volkers G, et al. Structure of human ST8SiaIII sialyltransferase provides insight into cell-surface polysialylation. Nat Struct Mol Biol. 2015; 22:627–635. [PubMed: 26192331]
26. Kellokumpu S, Hassinen A, Glumoff T. Glycosyltransferase complexes in eukaryotes: long-known, prevalent but still unrecognized. Cell Mol Life Sci. 2016; 73:305–325. [PubMed: 26474840]
27. Gleeson PA. Targeting of proteins to the Golgi apparatus. Histochem Cell Biol. 1998; 109:517–532. [PubMed: 9681632]
28. 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. 2016; 26:360–376. [PubMed: 26610890]
29. Halmo SM, et al. Protein O-Linked Mannose beta-1,4-N-Acetylglucosaminyl-transferase 2 (POMGNT2) Is a Gatekeeper Enzyme for Functional Glycosylation of alpha-Dystroglycan. J Biol Chem. 2017; 292:2101–2109. [PubMed: 27932460]
30. Praissman JL, et al. The functional O-mannose glycan on alpha-dystroglycan contains a phospho-ribitol primed for matriglycan addition. Elife. 2016; 5
31. Praissman JL, et al. B4GAT1 is the priming enzyme for the LARGE-dependent functional glycosylation of alpha-dystroglycan. Elife. 2014; 3
32. Li T, et al. Divergent Chemoenzymatic Synthesis of Asymmetrical-Core-Fucosylated and Core-Unmodified N-Glycans. Chemistry. 2016; 22:18742–18746. [PubMed: 27798819]
33. 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. 2016; 138:11575–11582. [PubMed: 27541995]
34. Marcos NT, et al. Role of the human ST6GalNAc-I and ST6GalNAc-II in the synthesis of the cancer-associated sialyl-Tn antigen. Cancer Res. 2004; 64:7050–7057. [PubMed: 15466199]
35. 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. 2016; 2:e1600241. [PubMed: 27386547]
36. 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. 2003; 14:889–902. [PubMed: 12631711]
37. Vandersall-Nairn AS, Merkle RK, O'Brien K, Oeltmann TN, Moremen KW. Cloning, expression, purification, and characterization of the acid alpha-mannosidase from Trypanosoma cruzi. Glycobiology. 1998; 8:1183–1194. [PubMed: 9858640]
38. Tessier DC, Thomas DY, Khouri HE, Laliberte F, Vernet T. Enhanced secretion from insect cells of a foreign protein fused to the honeybee melittin signal peptide. Gene. 1991; 98:177–183. [PubMed: 2016060]
39. Zhang F, et al. Characterization of the interaction between Robo1 and heparin and other glycosaminoglycans. Biochimie. 2013; 95:2345–2353. [PubMed: 23994753]
40. Luft JR, et al. A deliberate approach to screening for initial crystallization conditions of biological macromolecules. Journal of structural biology. 2003; 142:170–179. [PubMed: 12718929]
41. Kwan EM, Boraston AB, McLean BW, Kilburn DG, Warren RA. N-Glycosidase-carbohydrate-binding module fusion proteins as immobilized enzymes for protein deglycosylation. Protein engineering, design & selection : PEDS. 2005; 18:497–501.
42. Evans PR. An introduction to data reduction: space-group determination, scaling and intensity statistics. Acta Crystallogr D Biol Crystallogr. 2011; 67:282–292. [PubMed: 21460446]
43. Sheldrick GM. A short history of SHELX. Acta Crystallogr A. 2008; 64:112–122. [PubMed: 18156677]
44. Terwilliger TC, Berendzen J. Automated MAD and MIR structure solution. Acta Crystallogr D Biol Crystallogr. 1999; 55:849–861. [PubMed: 10089316]

45. McRee DE. XtalView/Xfit--A versatile program for manipulating atomic coordinates and electron density. Journal of structural biology. 1999; 125:156–165. [PubMed: 10222271]
46. Otwinowski Z, Minor W. Processing of X-ray diffraction data collected in oscillation mode. Meth Enzymol. 1997; 276:307–326.
47. Jogl G, Tao X, Xu Y, Tong L. COMO: a program for combined molecular replacement. Acta Crystallogr D Biol Crystallogr. 2001; 57:1127–1134. [PubMed: 11468396]
48. Schroder GF, Levitt M, Brunger AT. Super-resolution biomolecular crystallography with low-resolution data. Nature. 2010; 464:1218–1222. [PubMed: 20376006]
49. Adams PD, et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D Biol Crystallogr. 2010; 66:213–221. [PubMed: 20124702]

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

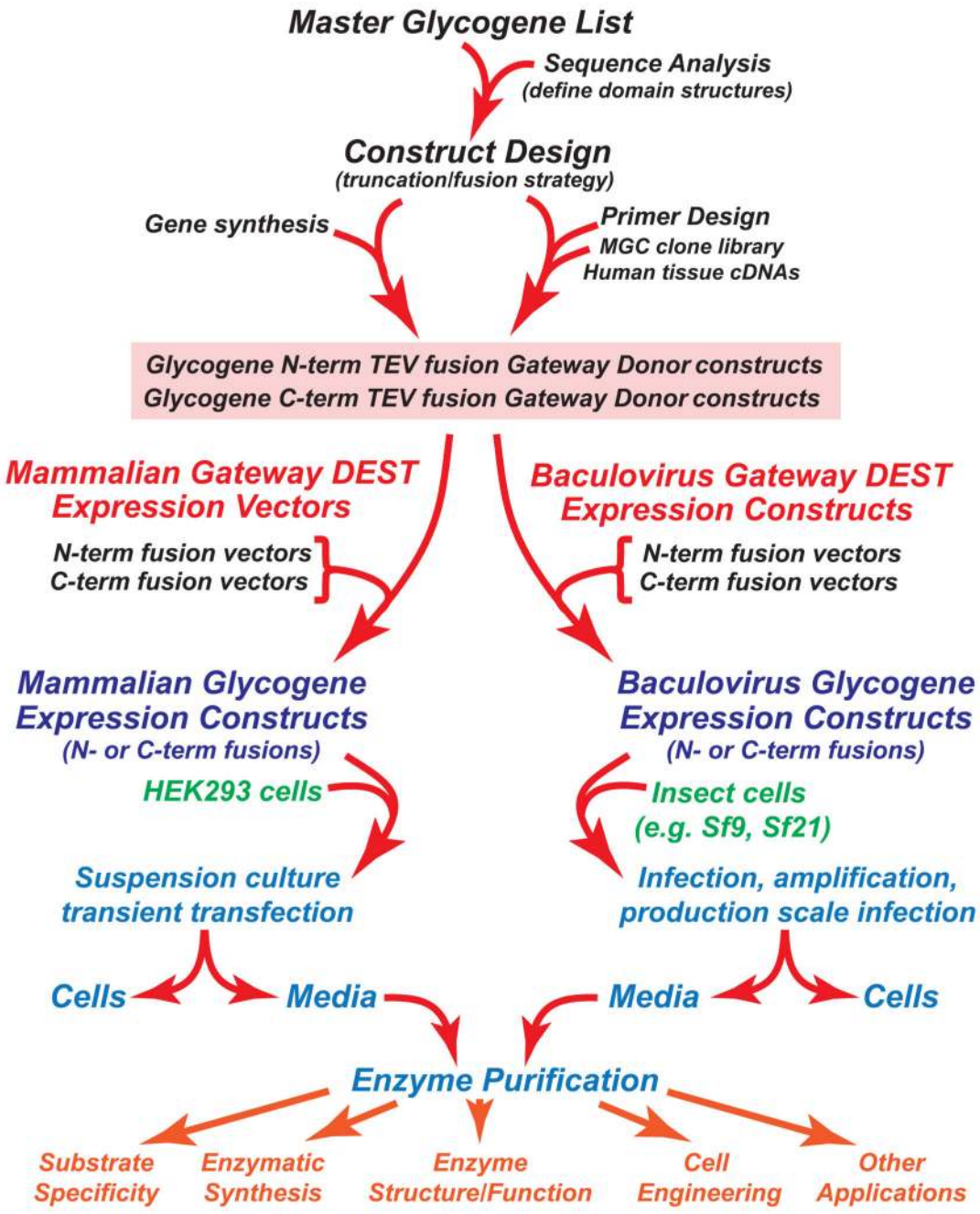

Figure 1. Flow chart for generation of glycoenzyme expression constructs.

The general workflow for the generation of human glycoenzyme expression constructs is depicted in the flow chart. A master glycogene list was developed as described in Experimental Procedures. Construct designs were generated based on the presence and location of TMD sequences and cleavable signal sequences and available knowledge regarding the location of the enzyme active sites, as described in Experimental Procedures. Truncated or full-length coding regions were incorporated into Gateway "donor" vectors by PCR from either Mammalian Gene Collection clone templates or human cDNA sources or,

in some cases, by gene synthesis. Each of the enzyme coding regions contained a TEV protease recognition site appended to either the NH_2- or COOH-terminal end of the open reading frame depending on the respective fusion protein strategy (Supplementary Fig. 1). Thus, a library of all human glycoenzyme constructs ("glycogenes") was captured in donor vectors that could be transferred to Gateway DEST vectors by Gateway LR recombination (see Experimental Procedures). Custom mammalian or baculovirus DEST vectors were generated 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) followed by separation into cell and media fractions. Secreted products were used for protein purification to produce enzyme preparations with utility for enzymology, structural studies, enzymatic glycan synthesis and numerous other applications.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

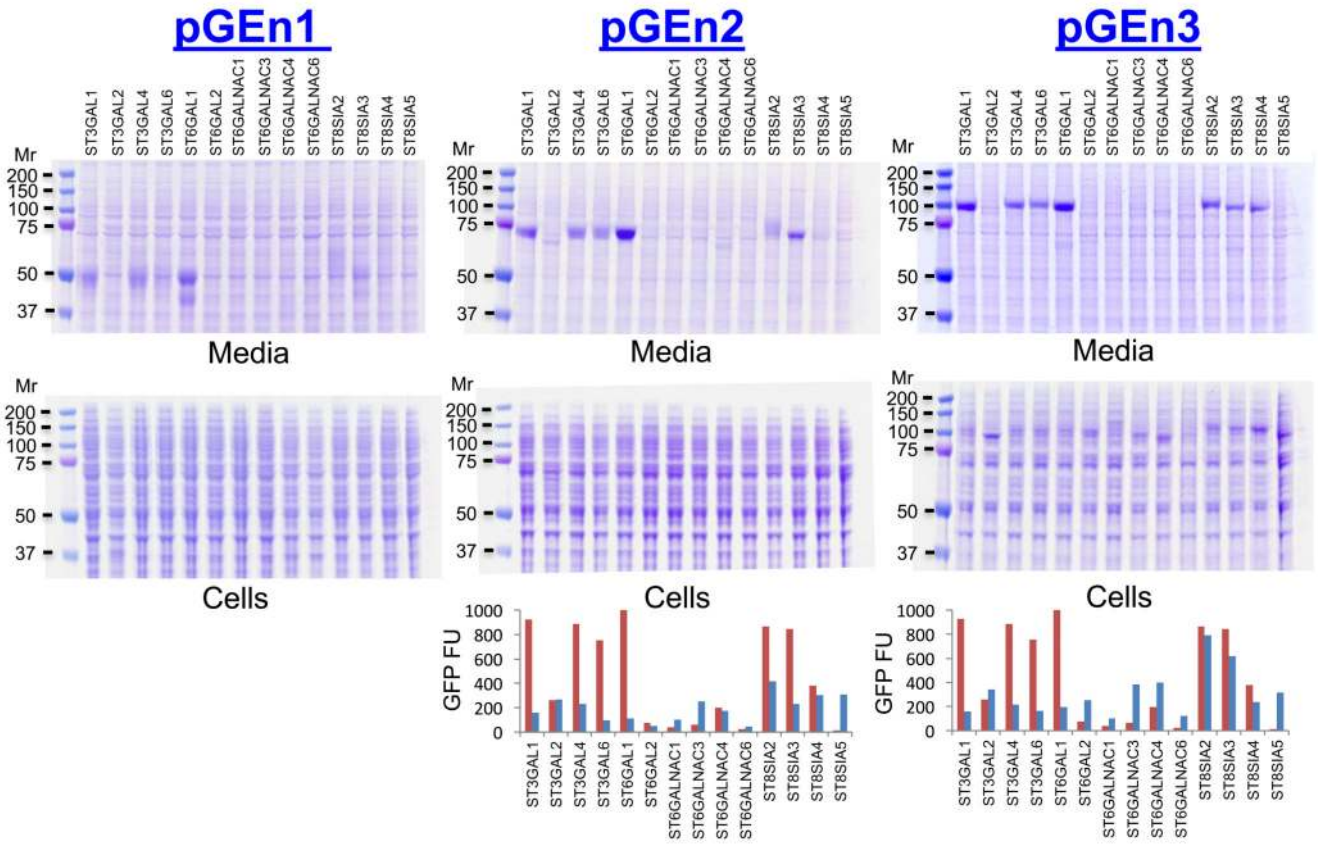

Figure 2. Expression and secretion of representative GTs in different NH_2-terminal fusion vectors.

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 HEK293 cells. Clarified conditioned media and cell extracts were collected and resolved by SDS-PAGE with subsequent protein staining. Additional samples of the conditioned media and extracts of cells transfected with the pGEn2 and pGEn3 expression constructs were tested for GFP fluorescence to quantify the recombinant products (bar charts for pGEn2 and pGEn3 constructs). 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.

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

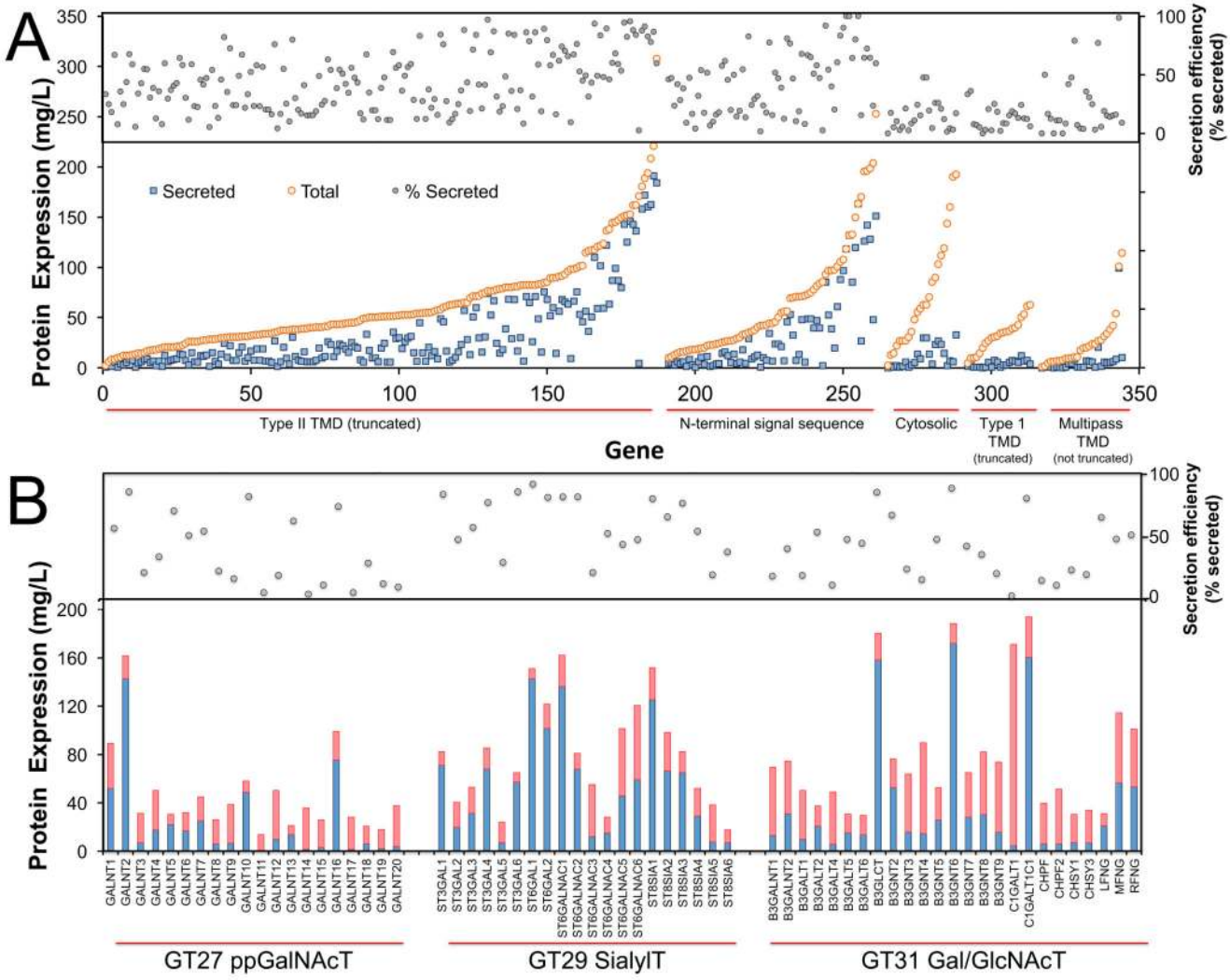

Figure 3. Expression and secretion of the glycoenzyme constructs in mammalian cells.

The full collection of all 339 human glycoenzymes as NH_2- or COOH-terminal fusion proteins was expressed in HEK293 cells and fusion protein production and secretion were profiled (*Panel A*). 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 Experimental Procedures, 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 determined to assess the efficiency of fusion protein secretion (% secretion). Fluorescence values were converted into milligram quantities of fusion protein per liter as described in Experimental Procedures. 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 (grey circles). To examine the relative expression and secretion levels within individual GT sequence families (*Panel B*) the values of secreted (blue bars) and cell-associated (pink bars) fusion proteins for individual GT27, GT29, and GT31 glycosyltransferase family members are displayed as stacked bar graphs with the respective gene name listed below each bar. The upper panel shows the efficiency of fusion protein secretion (% secretion) for each coding region (grey circles).

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

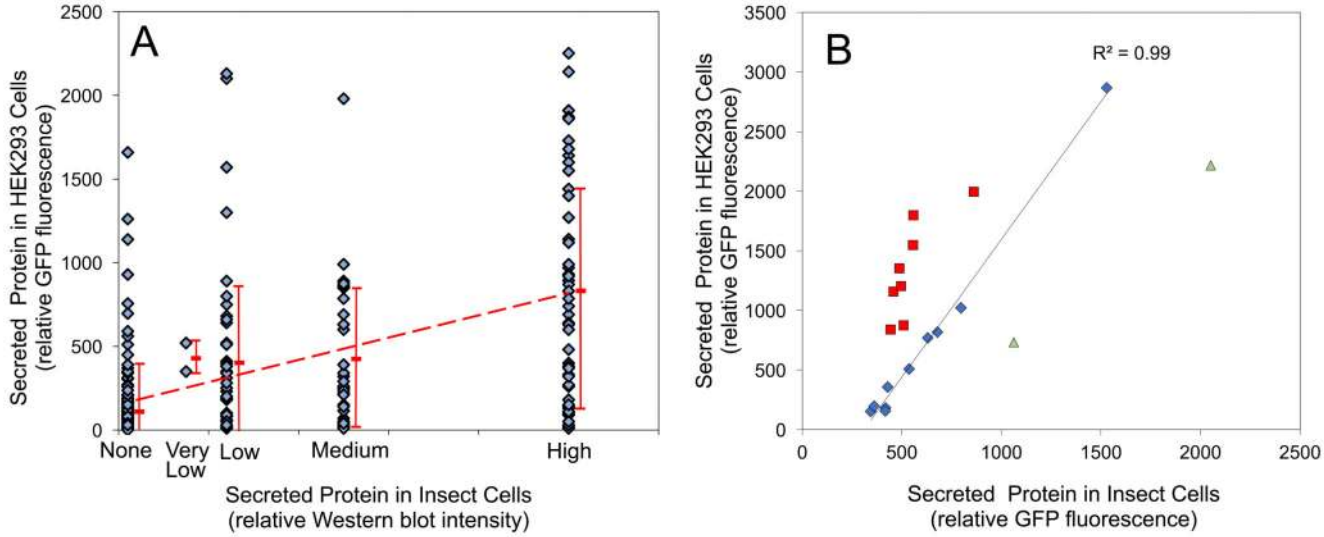

Figure 4. Comparison of secreted enzyme expression in transfected mammalian cells versus baculovirus infected insect cells.

The collection of human glycoenzymes was expressed using mammalian expression vectors or BEVs in each respective host (*Panel A*) 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. For BEV expression, the shorter His/StrepII tagged constructs were employed and the respective secreted fusion proteins were quantified using anti-His tag immunoblots (see Experimental Procedures). The semi-quantitative immunoblot intensities (Supplementary Table 2) were converted into arbitrary relative intensity values and compared with the values for enzyme secretion in mammalian cells in a 2-dimensional dot plot. The mean and standard error for the GFP fluorescence values are indicated by the solid red box and error bars. A trend line is shown for the mean values for GFP fluorescence (red dotted line). While the general trend for fusion protein secretion in each host system shows a positive correlation, there was significant scatter of the expression values in each recombinant host. Since 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 baculovirus large N-terminal tag vector for direct comparison with the equivalent pGEn2 mammalian GFP fusion constructs (*Panel B*). Comparison of secreted GFP fluorescence in the baculovirus and mammalian expression vectors indicated a close correlation for a subset of ten sialyltransferases (blue diamonds, R^2=0.99), while eight were more highly secreted in mammalian cells (red squares) and two were more highly secreted in BEVS (green triangles).

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

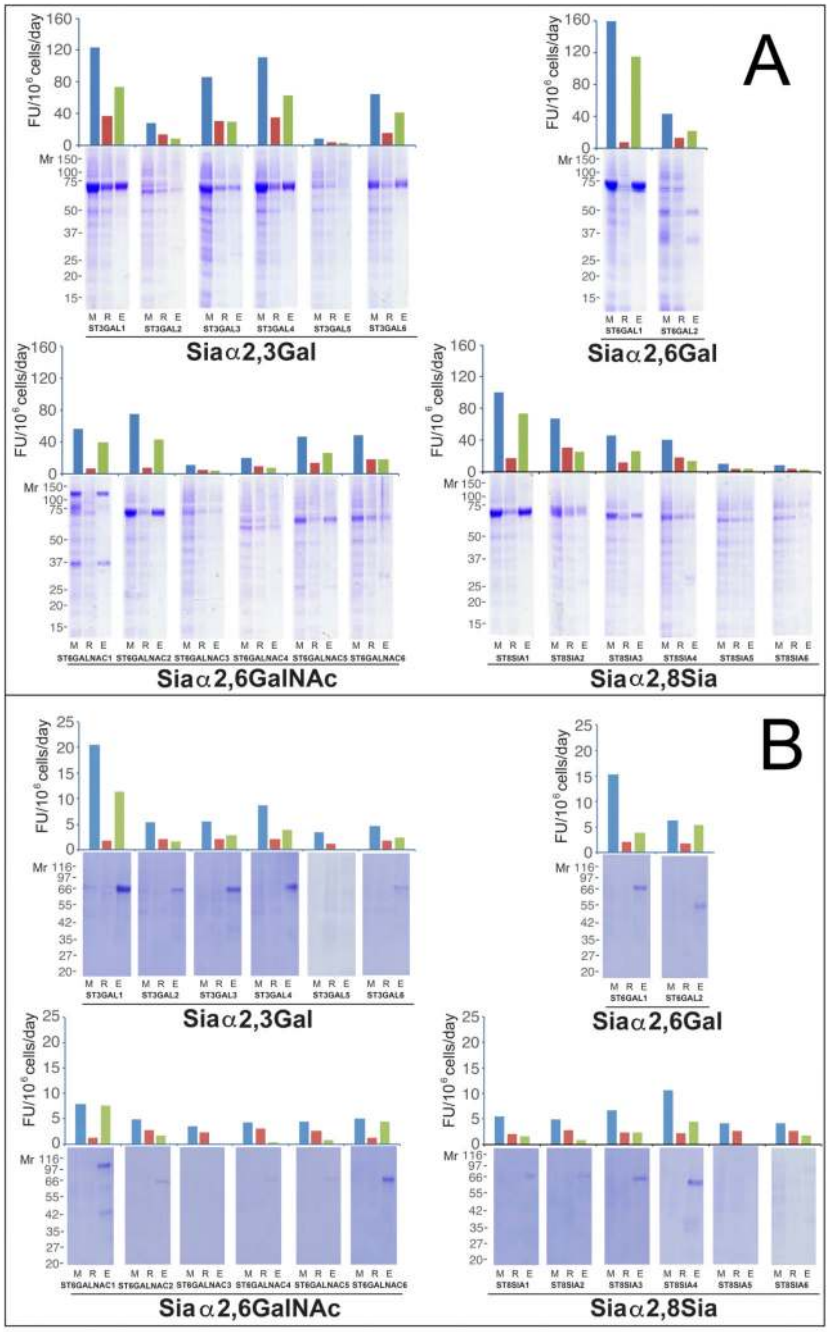

Figure 5. Purification of secreted recombinant fusion proteins.

The collection of recombinant 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 subfamily17. 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 ST6GALNAC1, which was expressed as a ~110 kDa glycosylated fusion protein.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

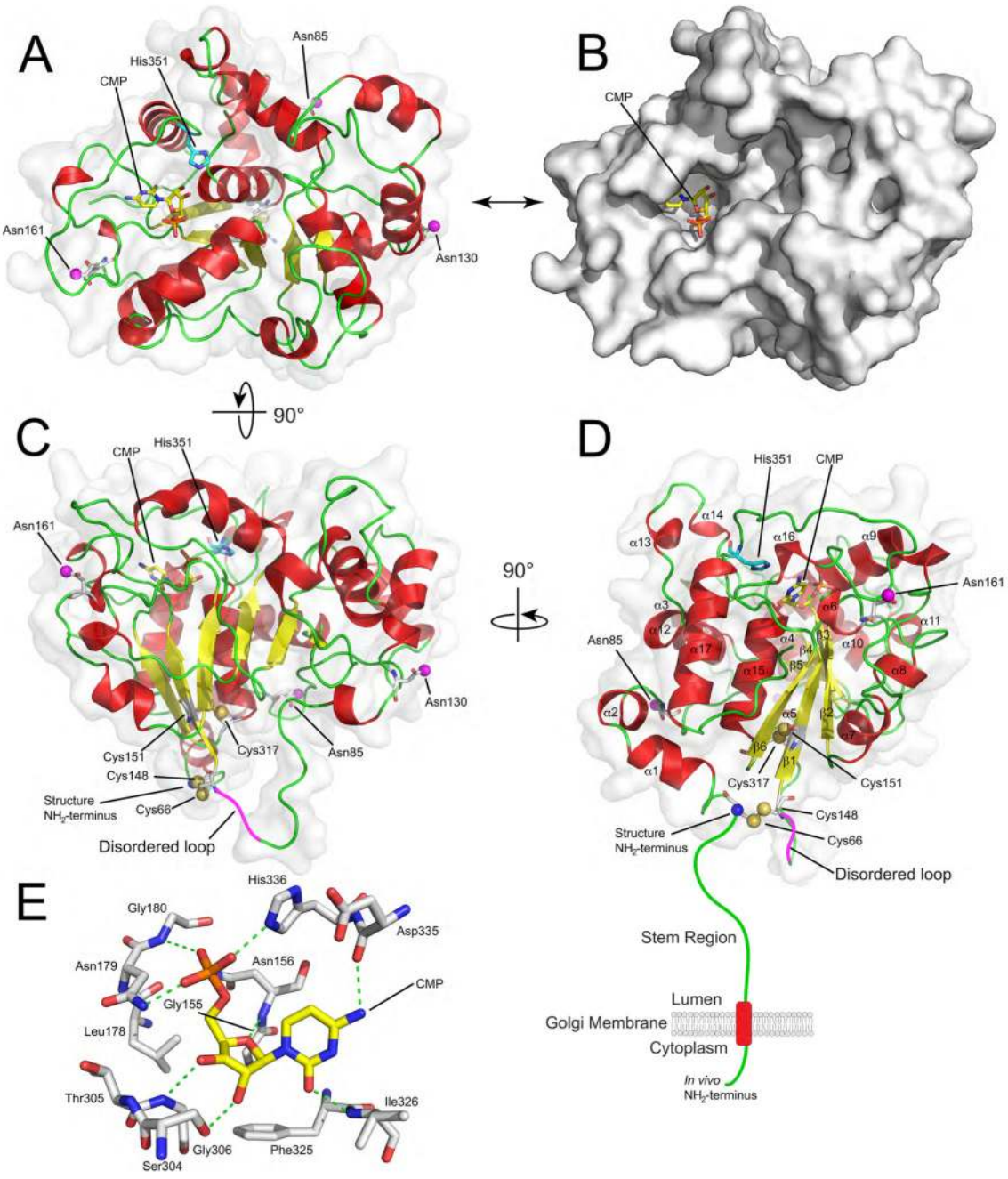

Figure 6. Structure of human ST6GALNAC2.

The structure of human ST6GALNAC2 was solved by X-ray diffraction using a recombinant enzyme preparation generated by transient transfection of HEK293S (GnTI-) cells as described in Experimental Procedures. The structure represents a single Rossmann-like (GT-A variant 2) fold17 with 6 β-strands in the domain core and 17 α-helical segments and loop regions (α-helices and β-sheets are labeled in *Panel D* numbered from the NH_2-terminus). There are six protomers in the crystallographic asymmetric unit and some of the units contain disordered loops of varying lengths. The images in *Panels A-E* represent Chain D,

which contains a single 3 residue disordered loop between helix α4 and strand β1 (magenta line in *Panels C and D*). *Panels A and B* represent a view from the top of the active site, where the bound donor analog, CMP, is shown in yellow stick representation. *Panels C and D* show a cartoon representation of the protein structure at 90° rotations relative to *Panel A*, while *Panel B* shows a surface representation in the same orientation as *Panel A*. The position of the His351 residue, the predicted catalytic base in the structure by comparison to structures of other GT29 sialyltransferases17, is shown in cyan stick representation. Weak electron densities for monosaccharide residues were found extended from the amide side chains of Asn85, Asn130, and Asn161 as predicted for each residue being found within an NxS glycosylation sequon (where x is any amino acid except Pro). The position of the respective Asn side chain is indicated by a white stick representation and a magenta sphere for the amide nitrogen 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 residues shown as spheres). The full length ST6GALNAC2 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 in *Panel D*. Extensive interactions were identified between the bound CMP donor analog and the binding site (*Panel E*), including polar interactions (green dotted lines) and a hydrophobic stacking with Phe325.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts