Crystal structure and sugar-binding ability of the C-terminal domain of N-acetylglucosaminyltransferase IV establish a new carbohydrate-binding module family

Crystal structure and sugar-binding ability of the C-terminal domain of *N*-acetylglucosaminyltransferase IV establish a new carbohydrate-binding module family

Nozomi Oka[1*], Sota Mori[1*], Marina Ikegaya[2], Enoch Y. Park[1,2,3], and Takatsugu Miyazaki[1,2,3**]

[1] Department of Agriculture, Graduate School of Integrated Science and Technology, Shizuoka University, Shizuoka, Japan.

[2] Department of Bioscience, Graduate School of Science and Technology, Shizuoka University, Shizuoka, Japan.

[3] Research Institute of Green Science and Technology, Shizuoka University, 836 Ohya, Suruga-ku, Shizuoka, 422-8529, Japan.

****Correspondence:** Takatsugu Miyazaki, Research Institute of Green Science and Technology, Shizuoka University, 836 Ohya, Suruga-ku, Shizuoka, 422-8529, Japan; Tel: +81-54-238-4886; E-mail: miyazaki.takatsugu@shizuoka.ac.jp

*These authors equally contributed to this study.

Running Title: C-terminal domain of GnT-IV establishes a new CBM family

Keywords: carbohydrate-binding module/ glycosyltransferase/ lectin/ *N*-glycan/ X-ray crystallography

Supplementary data: Supporting information includes Figures S1–S5 and Tables SI–SIII.

Abstract

N-glycans are modified by glycosyltransferases in the endoplasmic reticulum and Golgi apparatus. *N*-acetylglucosaminyltransferase IV (GnT-IV) is a Golgi-localized glycosyltransferase that synthesizes complex-type *N*-glycans in vertebrates. This enzyme attaches *N*-acetylglucosamine (GlcNAc) to the α-1,3-linked mannose branch of the *N*-glycan core structure *via* a β-1,4 linkage. Deficiency of this enzyme is known to cause abnormal cellular functions, making it a vital enzyme for living organisms. However, there has been no report on its three-dimensional structure to date. Here, we demonstrated that the C-terminal regions (named CBML) of human GnT-IVa and *Bombyx mori* ortholog have the ability to bind β-*N*-acetylglucosamine. Additionally, we determined the crystal structures of human CBML, *B. mori* CBML, and its complex with β-GlcNAc at 1.97, 1.47, and 1.15 Å resolutions, respectively, and showed that they adopt a β-sandwich fold, similar to carbohydrate-binding module family 32 (CBM32) proteins. The regions homologous to CBML (≥24% identity) were found in GnT-IV isozymes, GnT-IVb, and GnT-IVc (known as GnT-VI), and the structure of *B. mori* CBML in complex with β-GlcNAc indicated that the GlcNAc-binding residues were highly conserved among these isozymes. These residues are also conserved with the GlcNAc-binding CBM32 domain of β-*N*-acetylhexosaminidase NagH from *Clostridium perfringens* despite the low sequence identity (<20%). Taken together with the phylogenetic analysis, these findings indicate that these CBMLs may be novel CBM family proteins with GlcNAc-binding ability.

Introduction

Protein glycosylation is one of the most common posttranslational modifications in eukaryotic cells and is involved in variable physiological functions. *N*-Glycans are attached to asparagine residues in the Asn–X–Ser/Thr (X is any amino acid except Pro) sequon of proteins and perform various functions, including protein folding, stability, trafficking, and intracellular communication (Varki, 2017). Since the development of some diseases, such as cancer, causes specific changes in the glycan structures, they have recently been used as biomarkers (Kizuka and Taniguchi, 2016). An oligosaccharide precursor consisting of 14 monosaccharides ($Glc_3Man_9GlcNAc_2$) is transferred from a dolichol pyrophosphate carrier to asparagine side chains of nascent polypeptides as they are translocated into the lumen of the endoplasmic reticulum in the early step of *N*-glycan biosynthesis. The glycan is processed by glycoside hydrolases, including α-glucosidase I, α-glucosidase II, and α-mannosidases in the ER and further modified by Golgi-resident α-mannosidases and several glycosyltransferases. *N*-glycan structure varies among organisms and their tissue, and maturation of *N*-glycans is achieved through addition and branching with various monosaccharides, including *N*-acetylglucosamine (GlcNAc), galactose, fucose, and sialic acid (Helenius and Aebi, 2001). Hybrid- and complex-type *N*-glycans, mainly found in vertebrates, consist of these monosaccharides attached to the trimannosyl core of *N*-glycans by glycosyltransferases, such as *N*-acetylglucosaminyltransferases (GnTs) and galactosyltransferases. Although invertebrates, especially insects, mainly produce pauci-mannose-type glycans, recent studies have demonstrated that a small portion of *N*-glycan structures in invertebrates are more complex and elongated, with multiple monosaccharides and branches (Kimura et al., 2003; Koles et al., 2007; Aoki et al., 2007; Kurz et al., 2015; Stanton et al., 2017; Hykollari et al., 2018; Hykollari et al., 2019).

Golgi-resident GnTs are enzymes that transfer GlcNAc from a uridine diphosphate (UDP)-

GlcNAc donor to α-1,3 and α-1,6-mannosyl residues in *N*-glycans to form branches. GnT-I (MGAT1) is a key enzyme that catalyzes the transfer of GlcNAc to α-1,3-arm of the trimannosyl core via β-1,2 linkage, resulting in hybrid-type *N*-glycans (Brockhausen et al., 1988). This GlcNAc residue is essential for mannose-trimming by Golgi α-mannosidase II and addition of GlcNAc on α-1,6-arm by GnT-II (MGAT2) (Bendiak et al., 1987). Further branches of GlcNAc are formed by GnT-III (MGAT3), GnT-IVa,b (MGAT4A, B), GnT-IVc (MGAT4C, also known as GnT-VI and GnT-IV-H), GnT-V (MGAT5), and GnT-IX (MGAT5B) (reviewed in Nagae et al., 2020). Among them, the crystal structure of GnT-I was reported as the first case (Unligil et al., 2000), and those of GnT-II and GnT-V have been reported in recent years (Kadirvelraj et al., 2018; Nagae et al., 2018; Darby et al., 2020). The catalytic domains of all GnTs other than GnT-V are predicted to adopt the GT-A fold (Nagae et al., 2020), whereas that of GnT-V adopts the GT-B fold.

GnT-IV (also known as MGAT4) belongs to the glycosyltransferase family 54 (GT54) based on the CAZy classification, and this family contains homologs, GnT-IVa, GnT-IVb, GnT-IVc, and GnT-IVd (MGAT4D, also known as GnT-1IP). GnT-IVa and GnT-IVb are type II membrane proteins that catalyze the β-1,4 transfer of GlcNAc to the α-1,3-linked mannose branch of the *N*-glycan core structure (Oguri et al., 1997; Oguri et al., 2006). The glycosylation reaction of GnT-IVa and GnT-IVb requires a β-1,2-linked GlcNAc residue on the α-1,3-mannose arm (Fig. 1). Bovine GnT-IV and human GnT-IVa and GnT-IVb have been cloned and enzymatically characterized (Minowa et al., 1998; Yoshida et al., 1998). GnT-IVc has also been cloned from humans and chickens (Frukawa et al. 1999; Sakamoto et al. 2000), and the chicken enzyme exhibited the ability to transfer a residue on the α-1,6-mannose arm via a β-1,4 linkage (Taguchi et al., 2000; Sakamoto et al., 2000). In humans, GnT-IVa is expressed locally in the thymus, pancreas, and spleen, while GnT-IVb is expressed in various body tissue (Yoshida et al., 1998; Yoshida et al., 1999). Their products increase in chorionic gonadotropin in

choriocarcinoma patients (Mizuochi et al., 1983). GnT-IVd/MGAT4D has not been shown to possess glycosyltransferase activity and is primarily expressed in spermatogonium and spermatocytes, conferring heat stress tolerance to germ cells (Akintayo et al., 2020a; Akintayo et al., 2020b). This protein was reported to inhibit GnT-I function and to localize to the ER, the ER-Golgi intermediate compartment, and the *cis*-Golgi (Huang and Stanley, 2010; Au et al., 2015). In contrast to their physiological functions, their domain organization and three-dimensional structures remain unclear.

Herein, we hypothesize that GnT-IV has multiple domains with different functions and focused on its C-terminal region based on structural prediction. We identified the C-terminal regions of human GnT-IVa (HsGnT-IVa) and its insect ortholog as novel carbohydrate-binding modules, and their crystal structures were determined.

Results

Domain structure prediction

A structure prediction of the full length of HsGnT-IVa was initially performed using SWISS-MODEL (Waterhouse et al., 2018). Partial homology models were generated using rabbit GnT-I (PDB 1FO9, sequence identity = 14.5%) and a carbohydrate-binding module family 32 (CBM32) domain of *Clostridium perfringens* β-*N*-acetylglucosaminidase NagH (CpNagHCBM32-2, PDB 2WDB, sequence identity = 21.4%) as templates. The result indicated that the N-terminal region (residues 104–380) of the HsGnT-IVa luminal domain has a GT-A fold, as previously predicted, and the C-terminal region (residues 386–535) adopts a β-sandwich fold (Fig. S1). AlphaFold2 prediction showed that the luminal region of HsGnT-IVa is divided into the *N*-terminal GT-A fold and the C-terminal β-sandwich domains connected by a linker (Fig. 2A). The β-sandwich fold is observed in various CBMs, which are domains found in carbohydrate-active enzymes, such as glycoside hydrolases and glycosyltransferases, to facilitate substrate binding. Therefore, we hypothesized that the C-terminal CBM-like domain (CBML) acts as CBM. The sequence alignment of human GnT-IV homologs revealed that only GnT-IVd/MGAT4D lacks a CBML domain. GnT-IV orthologs are encoded in vertebrate and invertebrate genomes (Walski et al., 2016; Nishihara, 2018). A GnT-IV ortholog (BmGnT-IV) from *Bombyx mori*, a model lepidopteran insect, has 35.4% sequence identity and a similar domain organization as HsGnT-IVa.

CBML binds GlcNAc

Recombinant proteins of CBMLs of HsGnT-IVa and BmGnT-IV (named HsCBML and BmCBML, respectively) were expressed in *Escherichia coli* and purified to investigate the carbohydrate-binding ability of CBML (Fig. 2B). Isothermal titration calorimetry revealed that both CBMLs showed affinity toward GlcNAc but not mannose, glucose, and *N*-

acetylgalactosamine (Figs. 3 and S2). Due to the low affinity, the fitting process could not accurately calculate the binding stoichiometry. However, based on the CBML structures, which revealed a single binding site (see below), adjusting the stoichiometry to one for data fitting was deemed valid. The binding constants (K_a) of HsCBML and BmCBML for GlcNAc were 287 ± 71 M^{-1} and 217 ± 60 M^{-1}, respectively (Table I). We also tested *N*,*N*'-diacetylchitobiose (GlcNAc-β-1,4-GlcNAc) and *p*-nitrophenyl (pNP)-β-GlcNAc for HsCBML (Fig. 3) and discovered that GlcNAc-β-1,4-GlcNAc and pNP-β-GlcNAc had 4.4 and 6.9 times higher association constants than GlcNAc, respectively. In contrast, HsCBML and BmCBML had no binding activity to pNP-α-GlcNAc (Fig. 3), suggesting that CBMLs recognize the β anomer of GlcNAc. The Wiseman *c* parameter (Wiseman et al., 1989) was calculated to be 0.065–0.591. Generally, *c* values within the range of $10 < c < 500$ is considered to be meaningful for K_a calculation, but Turnbull and Daranas reported that binding parameters can be obtained even if *c* value is in the range of $0.01 < c < 10$ if analysis meets the proposed conditions: (1) a sufficient portion of the binding isotherm is used for analysis, (2) the binding stoichiometry is known, (3) the concentrations of both ligand and receptor are known with accuracy, and (4) there is an adequate level of signal-to-noise in the data (Turnbull and Daranas, 2003). Therefore, the parameters in the present study can be reliable (Table I). The ΔH values indicate that all CBML bindings were enthalpy-driven. CpNagHCBM32-2 binds GlcNAc and β-GlcNAc-linked disaccharides. Among the reported *N*-acetylhexosamine-binding members of CBM32, CpNagHCBM32-2 had the most similar association constants to the CBMLs (Table I).

Crystal structures of CBMLs

The crystal structures of HsCBML and BmCBML were determined at 1.97 and 1.47 Å resolution, respectively, by the molecular replacement method using their AlphaFold2 models (Table II). They belong to the space group $P2_1$ and contain two monomers in each asymmetric

unit. HsCBML and BmCBML adopt the β-sandwich fold, consisting of nine and eight β-strands, and have three and two short α-helices, respectively (Fig. 4A and B). The structural alignment of these CBMLs revealed that their core regions are almost identical [root mean squared deviation (RMSD) of Cα, 0.735 Å, (Fig. 4C)]. No metal ion was found in both structures. A structural homology search using the Dali server revealed that the C-terminal domain of *Streptococcus pneumoniae* GH29 α-fucosidase (PDB 6ORG, *Z* score = 12.6–12.7, RMSD = 2.4–2.6 Å, sequence identity = 12%–13%) and CpNagHCBM32-2 (PDB 2WDB, *Z* score = 12.4–12.7, RMSD = 2.4 Å, sequence identity = 19%–20%) are the most structurally homologous proteins for HsCBML and BmCBML (Fig. 4D and Tables SI and SII). The CBM47 proteins, similar to several CBM32 proteins, were also found to be structurally homologous (Tables SI and SII). All hit proteins comprised of a β-sandwich topology.

GlcNAc-binding site

Several attempts to co-crystallize HsCBML and GlcNAc failed because the predicted sugar-binding sites faced each other molecule in the crystals, making it difficult for the ligand to bind the site. Instead, the structure of BmCBML in complex with GlcNAc was determined at 1.15 Å resolution (Table II). Soaking and co-crystallization with GlcNAc-β-1,4-GlcNAc and pNP-β-GlcNAc were also unsuccessful to obtain diffractable crystals. The co-crystal belongs to the same space group ($P2_1$) and contains two molecules in the crystallographic asymmetric unit as the unliganded BmCBML. An electron density map for β-GlcNAc was found at the side surface of the β-sandwich structure in each monomer (Fig. 5A). Asp480 forms hydrogen bonds with two water molecules in the unliganded structure (Fig. 5B), while it forms hydrogen bonds with O3 and O4 of β-GlcNAc (Fig. 5C). The main chain N of Trp535 also directly forms a hydrogen bond with O4 of β-GlcNAc and the side chain of His477 interacts with O3 of β-GlcNAc *via* a water molecule. Trp535 forms π–π stacking interaction with the pyranose ring,

and Trp445 and Tyr429 may contribute to creating a hydrophobic environment to accept the sugar ring and the methyl group of GlcNAc. Superposition of the binding sites of HsCBML and BmCBML showed that all amino acid residues involved in the binding are identical (Fig. 5D). Among human GnT-IV paralogs, only GnT-IVd/MGAT4D lacks a domain corresponding to CBML (Fig. S3). Amino acid sequence alignment of vertebrate GnT-IVa, -IVb, and -IVc, and insect orthologs revealed that amino acid residues involved in GlcNAc-binding are well conserved (Figs. 5E, S3, and S4).

The structure of CpNagHCBM32-2, a GlcNAc-binding member of the CBM32 family, in complex with GlcNAc-β-1,2-Man (PDB 2WDB) was reported (Ficko-Blean and Boraston 2009). The disaccharide is the nonreducing-end component of the GnT-IV substrate and product (Fig. 6A). Figure 6B depicts the superimposition of the ligand-binding sites of BmCBML and CpNagHCBM32-2. Despite having a low whole sequence identity, the amino acid residues involved in GlcNAc-binding are well conserved. Furthermore, the orientations of β-GlcNAc pyranose rings are almost identical, and the α-face of GlcNAc is held by the tryptophan side chain. These observations indicate that the binding site is specific for β-GlcNAc, which is consistent with the ITC results. In contrast to GlcNAc, the reducing-end mannose residue is exposed to solvent (Fig. 6C), suggesting that longer oligosaccharides with the non-reducing end β-GlcNAc could be accepted in the binding site. This agrees with the fact that HsCBML showed affinity toward both GlcNAc-β-1,4-GlcNAc and pNP-β-GlcNAc. Unfortunately, we could not obtain crystal structures in complex with disaccharides, and it is not clear whether and how the reducing-end sugar interacts with the CBML. The reducing end sugar may be specifically recognized by the CBML because CpNagHCBM32-2 showed the higher K_a value for GlcNAc-β-1,2-Man than that for GlcNAc-β-1,4-GlcNAc (Table I).

Discussion

In this study, we demonstrated that the C-terminal domains of HsGnT-IVa and insect ortholog, which were initially predicted to have a similar fold as CBM32 proteins, had specific binding ability toward β-GlcNAc using the ITC binding analysis, supported by X-ray crystallography. CBMs are noncatalytic modules with sugar-binding ability found in carbohydrate-active enzymes and play various roles in assisting the enzyme function, such as proximity effect and targeting function (Boraston et al., 2004). CBMs are divided into three types: Type A, surface-binding CBM; Type B, glycan-chain-binding B; and Type C, small-sugar-binding. Type A CBMs bind insoluble crystalline polysaccharides and they have platform-like binding sites with multiple aromatic residues. Type B CBMs have a long cleft in common that recognizes an oligosaccharide unit (four or more residues) in polysaccharides. Type C CBMs typically recognize shorter oligosaccharides (three or less residues) compared to Type A and B CBMs. This classification is useful to understand structure-function relationships of CBMs. The functional and structural analyses in this study revealed that CBMLs specifically recognize the nonreducing-end β-GlcNAc residue of oligosaccharides (Table I and Fig. 5). CBM32 domains have been found in various enzymes, including GHs, polysaccharide lyases, and peptidases, and their binding specificities vary, including galactose (Ficko-Blean and Boraston, 2006; Boraston et al., 2007; Ficko-Blean et al., 2012; Grondin et al., 2014), GalNAc (Ficko-Blean and Boraston, 2006; Ficko-Blean et al., 2012; Grondin et al., 2014; Grondin et al., 2017; Pluvinage et al., 2021), GlcNAc (Ficko-Blean and Boraston, 2009), chitosan (Shinya et al., 2013; Das et al., 2016; Shinya et al., 2016), and polygalacturonic acid (Abbott et al., 2007). Because all CBM32 proteins have been reported to recognize the nonreducing end of substrate oligosaccharides, CBM32 are classified into Type C (Boraston et al., 2004). Additionally, the CBMLs have similar binding and structural characteristics as the CBM32 members, especially CpNagHCBM32-2, indicating that CBML can be classified as

Type C CBM. HsCBML shows 34% sequence identity with BmCBML; the corresponding regions of human GnT-IVb and GnT-IVc share 57% and 26% sequence identities with HsCBML, respectively. In contrast, HsCBML and BmCBML show 13% and 14% identities with CpNagHCBM32-2, respectively. Moreover, the phylogenetic analysis showed that CBML proteins form the distinct clades from CBM32 proteins (Fig. S5), indicating that the CBMLs define a novel CBM family (CBM94, B. Henrissat and N. Terrapon, personal communication).

The GlcNAc-bound structure indicates that three direct hydrogen bonds, one indirect hydrogen bond *via* a water molecule, and hydrophobic interaction are involved in GlcNAc-binding (Fig. 5C). Comparison between the unliganded form and GlcNAc-complex of BmCBML suggests that the hydrogen bonds between Asp480 and water molecules are dissolved and the water molecules are released to bulk when GlcNAc binds to the protein. The magnitude of binding enthalpy reflects the loss of protein–solvent hydrogen bonds, van der Waals interactions, and the formation of protein–ligand bond and salt bridges (Perozzo et al., 2004). Therefore, the enthalpically-driven bindings observed by the ITC analysis are consistent with the interactions observed in the structure of BmCBML-GlcNAc complex. An enthalpically-driven process is commonly seen in protein–carbohydrate interactions (Dam and Brewer, 2002; Boraston et al. 2004; Abbott and Boraston, 2012).

The structure of BmCBML in complex with β-GlcNAc suggested that these proteins do not strictly recognize the second sugar of the ligand compared with the non-reducing β-GlcNAc. The ITC results for HsCBML showed a slight difference in the binding constants between GlcNAc-β-1,4-GlcNAc, and pNP-β-GlcNAc. Also, the K_a values of CpNagHCBM32-2 toward disaccharides with non-reducing end β-GlcNAc were reported to vary and the highest for GlcNAc-β-1,2-Man (Table I). The side chain of Tyr429 of BmCBML is predicted to be relatively close to the second sugar (Fig. 6). Tyr429 is well conserved among CBML homologs, but it is substituted to phenylalanine in GnT-IVc (Figs. S3 and S4). Therefore, CBML proteins

may have slightly different binding strength and specificity toward disaccharides.

GnT-IVa and -IVb have been shown to act only on *N*-glycans with β-1,2-GlcNAc at the α-1,3-mannose branch attached by GnT-I (Oguri et al., 1997; Oguri et al., 2006); GnT-IVc requires both β-1,2-GlcNAc and β-1,6-GlcNAc at the α-1,6-mannose branch attached by GnT-II and GnT-V (Taguchi et al., 2000). These specificities indicate that these enzymes specifically recognize the nonreducing end β-1,2-GlcNAc residues. Although it is unclear whether the CBML recognizes the sugar residue in concert with the active site of the catalytic domain, CBMLs may be involved in targeting and proximity to the substrate GlcNAc-terminated *N*-glycans to act on. Some CBM32 proteins have been reported to be involved in efficient catalysis via substrate binding (Mizutani et al., 2014; Das et al., 2016). Similarly, sugar-binding domains have been discovered in some glycosyltransferases and they are often called 'lectin domains' or 'lectin-like domains.' The sugar-binding domain of polypeptide *N*-acetylgalactosaminyltransferases, which is classified into the CBM13 family and is involved in the biosynthesis of mucin-type *O*-glycan, affects catalytic efficiency (Gerken et al., 2013). The stem region of protein *O*-linked mannose β1,2-*N*-acetylglucosaminyltransferase 1 comprises two β-sheets and two α-helices and recognizes β-linked GlcNAc of *O*-mannosyl glycan (Kuwabara et al., 2016) although the stem regions of many Golgi-localized GTs have not been reported to show carbohydrate-binding ability. Therefore, CBML domains in GnT-IV homologs may have similar functions as the 'lectin-like domains' in the reported glycosyltransferases. Further investigation is necessary to reveal the relationship between the sugar-binding ability of CBML and the GnT-IV catalytic reaction.

Conclusively, we identified the C-terminal domains (CBML) of GnT-IVa and its insect ortholog as novel CBM that binds β-GlcNAc. The crystal structures of the CBMLs clarified the GlcNAc-binding mechanism and its commonality with the CBM32 proteins, especially CpNagHCBM32-2. Considering the low sequence identity with known CBMs, CBMLs are

now classified into a new CBM family, CBM94.

Materials and methods

Structure prediction and bioinformatics

SWISS-MODEL (https://swissmodel.expasy.org/, Waterhouse et al., 2018) and AlphaFold2/ColabFold (Jumper et al., 2021; Mirdita et al., 2022) were used to predict structures of HsGnT-IVa and BmGnT-IV. The transmembrane regions of HsGnT-IVa and BmGnT-IV were predicted by the TMHMM server (https://services.healthtech.dtu.dk/service.php?TMHMM-2.0) (Krogh et al., 2001). Amino acid sequence alignment was performed using Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/, Sievers et al., 2011) and visualized using ESPript 3 (Robert and Gouet, 2014).

Cloning and plasmid construction

DNA encoding HsCBML was amplified by polymerase chain reaction (PCR) using primers, HsCBML_NheI_ F and HsCBML_XhoI_R, cDNA of HEK293 cell as a template, and a KOD ONE DNA polymerase (Toyobo, Osaka, Japan). DNA fragment was treated with NheI (New England Biolabs, Massachusetts, U.S.A) and XhoI (New England Biolabs) and ligated with a pET-28a(+) vector (Merck, Darmstadt, Germany) pre-treated with the same restriction enzymes.

The first-strand cDNA was synthesized from a fifth-instar larva using a PrimeScript RT reagent kit (Takara Bio, Kusatsu, Japan), as previously described (Miyazaki et al., 2019a; Miyazaki et al., 2019b). A DNA fragment encoding full-length BmGnT-IV was amplified by

PCR using ExTaq DNA polymerase (Takara Bio) and primer pair, BmGnT-IV_F and BmGnT-IV_R. The resultant DNA was ligated into a pMD19 T vector (Takara Bio) and sequenced. Subsequently, DNA encoding BmCBML was amplified using a KOD One DNA polymerase and primer pair, BmCBML_NdeI_F and BmCBML_HindIII_R, then ligated into a pET-28a(+) vector pre-digested with NdeI and HindIII restriction enzymes.

Expression and purification

E. coli BL21 (DE3) transformed with each expression plasmid was cultured at 37°C in 1 L of Luria–Bertani medium containing 50 μg/mL kanamycin. When the culture reached an optical density of 0.6–0.8 at 600 nm, the culture was ice-cooled, induced with 1.0 mM isopropyl-β-D-1-thiogalactopyranoside, and further incubated for 20 h at 20°C. After the incubation, bacterial cells were harvested by centrifugation at 10,000 × g for 15 min and stored at −30ºC until further experiments. Then, the cells were resuspended in 50-mM Tris-HCl buffer (pH 7.5) containing 300-mM NaCl and disrupted by sonication. Next, the insoluble materials were removed by centrifugation at 10,000 × g for 15 min.

The supernatant was added to 3 mL of Ni-sepharose 6 Fast Flow (Cytiva, Tokyo, Japan) in a gravity column equilibrated with 50 mM Tris-HCl buffer (pH 7.5) containing 300 mM NaCl. The resin was washed with 50 mM Tris-HCl (pH 7.5) containing 300 mM NaCl, and the recombinant proteins were eluted with step-wise increasing concentrations (50, 100, and 250 mM) of imidazole. Protein purity was confirmed by SDS-PAGE with Coomassie brilliant blue staining. The absorbance of purified protein was measured at 280 nm wavelength by ultraviolet–visible spectrophotometry using a spectrophotometer UV-1800 (Shimadzu, Kyoto, Japan). Sample concentrations were determined from the molar absorption coefficients, 18,450 cm^{-1} M^{-1} for HsCBML and 28,420 cm^{-1} M^{-1} for BmCBML, calculated by ProtParam (Gasteiger et al., 2005).

Isothermal titration calorimetry

The purified recombinant proteins were concentrated to 0.3 mM using 3 kDa-cut off Amicon Ultra-15 (Merck) and dialyzed against 50 mM 4-(2-hydroxyethyl)piperazine-1-ethanesulfonicacid (HEPES)-NaOH (pH 7.0). D-GlcNAc, *N*-acetyl-D-galactosamine, D-mannose, and D-glucose (30 mM, 0.5 μL each) were titrated into 0.3 mM HsCBML and BmCBML. HsCBML was also measured for affinity to pNP-β-GlcNAc (10 mM), *N*,*N*'-diacetylchitobiose (10 mM), and pNP-α-GlcNAc (5 mM). Titrations were performed by MicroCal iTC200 (Malvern Panalytical Ltd, Enigma Business Park, UK) at 25°C. The titration results were analyzed using MicroCal Origin ITC (Malvern Panalytical Ltd.). Thermodynamic parameters were obtained by nonlinear least-squares fitting using the one binding site model, and the average and standard error were calculated using data from at least three ITC measurements.

X-ray crystallography

HsCBML and BmCBML were concentrated in 50-mM HEPES-NaOH buffer (pH 7.0) containing 50-mM NaCl to 20 mg/mL. Crystallization was performed at 20°C using the hanging-drop vapor diffusion method, where 1 μL of protein was mixed with the same volume of a reservoir solution. HsCBML was crystallized in a reservoir solution comprised of 0.2-M lithium nitrate mixed with 20% PEG 3350. BmCBML was crystallized in a reservoir solution containing 12% (w/v) PEG20000 and 0.1-M 2-(*N*-morpholino)ethanesulfonic acid-NaOH buffer (pH 7.0). For co-crystallization with GlcNAc, 0.2 μL of 100 mM GlcNAc was added to a 2-μL drop where BmCBML and the reservoir solution were mixed as described above. Crystals were cryo-protected with 20% (v/v) ethylene glycol in each reservoir solution and flash-frozen in liquid nitrogen. Diffraction data were collected at Photon Factory beamlines

AR-NW12A and BL5A (Tsukuba, Japan). Data collected were processed by XDS (Kabsch 2010) or Mosflm (Battye et al., 2011) and scaled using SCALA (Evans, 2006) in the CCP4 suite (Winn et al., 2011). The phase was determined by the molecular replacement method using MOLREP (Vagin et al., 2010), and each model was created with ColabFold/AlphaFold2 as a search model. Manual model building was performed using COOT (Emsley et al., 2010), and refinement was performed using REFMAC5 (Murshudov et al., 2011) and Translation/Libration/Screw Motion Determination (Painter and Merritt. 2006). The final models were evaluated using the MolProbity server (Chen et al., 2010). Structure images were prepared using PyMOL (Schrödinger, LLC). Structural similarity searches were performed using the Dali server (Holm, 2020).

Acknowledgments

We acknowledge the staff of the Photon Factory for X-ray data collection. This research was approved by the Photon Factory Program Advisory Committee (proposal No. 2021G013). We thank Enago (www.enago.jp) for the English language review.

Abbreviations: CBM, carbohydrate-binding module; CBM32, carbohydrate-binding module family 32; CBML, carbohydrate-binding module-like; CpNagHCBM32-2, CBM32-2 domain of *Clostridium perfringens* *N*-acetylglucosaminidase NagH; GnT, *N*-acetylglucosaminyltransferase; GnT-IV, *N*-acetylglucosaminyltransferase IV; pNP, *p*-nitrophenyl; RMSD, root mean square deviation.

Data availability statement

The nucleotide sequence of BmGnT-IV was submitted to the DDBJ/EMBL/GenBank databases under the accession number LC710141. The atomic coordinates and structure factors of HsCBML, BmCBML, and BmCBML complexed with GlcNAc have been deposited in the Worldwide Protein Data Bank (wwPDB, http://wwpdb.org/) under the accession codes 7XTL, 7XTM, and 7XTN, respectively.

Funding

This work was supported by a grant-in-aid for scientific research by the Japan Society for the Promotion of Science KAKENHI (grant number: 19K15748).

Conflict of interest

None declared.

References

Abbott DW, Boraston AB. 2012. Quantitative approaches to the analysis of carbohydrate-binding module function. *Methods Enzymol*. 510, 211–231

Abbott DW, Hrynuik S, Boraston AB. 2007. Identification and characterization of a novel periplasmic polygalacturonic acid binding protein from *Yersinia enterolitica*. *J Mol Biol*. 367:1023–1033

Akintayo A, Liang M, Bartholdy B, Batista F, Aguilan J, Prendergast J, Sabrin A, Sundaram S, Stanley P. 2020a. The Golgi Glycoprotein MGAT4D is an Intrinsic Protector of Testicular Germ Cells From Mild Heat Stress. *Sci Rep*. 10:2135

Akintayo A, Mayoral J, Asada M, Tang J, Sundaram S, Stanley P. 2020b. Point mutations that inactivate MGAT4D-L, an inhibitor of MGAT1 and complex *N*-glycan synthesis. *J Biol Chem*. 295:14053–14064.

Aoki K, Perlman M, Lim J-M, Cantu R, Wells L, Tiemeyer M. 2007. Dynamic developmental elaboration of *N*-linked glycan complexity in the *Drosophila melanogaster* embryo. *J Biol Chem*. 282:9127–9142

Au CE, Hermo L, Byrne E, Smirle J, Fazel A, Simon PH, Kearney RE, Cameron PH, Smith CE, Vali H, Fernandez-Rodriguez J, Ma K, Nilsson T, Bergeron JJ. 2015. Expression, sorting, and segregation of Golgi proteins during germ cell differentiation in the testis. *Mol Biol Cell*. 26:4015–4032

Battye TG, Kontogiannis L, Johnson O, Powell HR, Leslie AGW. 2011. iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM. *Acta Crystallogr D Biol Crystallogr*. 67:271–81

Bendiak B, Schachter H. 1987. Control of glycoprotein synthesis. Purification of UDP-N-acetylglucosamine:α-D-mannoside β1-2 N-acetylglucosaminyltransferase II from rat liver. *J Biol Chem*. 1987. 262:5775–5583

Boraston AB, Bolam DN, Gilbert HJ, Davies GJ. 2004. Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochem J*. 382:769–781.

Boraston AB, Ficko-Blean E, Hearley M. 2007. Carbohydrate recognition by a large sialidase toxin from *Clostridium perfringens*. *Biochemistry*. 46:11352–11360

Boraston AB, Wang D, Burke RD. 2006. Blood Group Antigen Recognition by a *Streptococcus Pneumoniae* Virulence Factor. *J Biol Chem*. 281:35263–35271

Brockhausen I, Narasimhan S, Schachter H. 1988. The biosynthesis of highly branched N-glycans: studies on the sequential pathway and functional role of N-acetylglucosaminyltransferases I, II, III, IV, V and VI. *Biochimie*. 70:1521–1533

Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormono RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC. 2010. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr*. 66:12–21

Dam TK, Brewer CF. 2002. Thermodynamic studies of lectin−carbohydrate interactions by isothermal titration calorimetry. *Chem Rev*. 102:387–430

Darby JF, Gilio AK, Piniello B., Roth C, Blagova E, Hubbard RE, Rovira C, Davies GJ, Wu L. 2010. Substrate Engagement and Catalytic Mechanisms of *N*-Acetylglucosaminyltransferase V. *ACS Catal*. 10:8590−8596

Das SN, Wagenknecht M, Nareddy PK, Bhuvanachandra B, Niddana R, Balamurugan R, Swamy MJ, Moerschbacher BM, Podile AR. 2016. Amino Groups of Chitosan Are Crucial for Binding to a Family 32 Carbohydrate Binding Module of a Chitosanase from *Paenibacillus elgii*. *J Biol Chem*. 291:18977–18990

Emsley P, Lohkamp B, Scott WG, Cowtan K. 2010. Features and development of Coot. *Biological Crystallography*. 66:486–501

Evans P. 2006. Scaling and assessment of data quality. *Acta Crystallogr D Biol Crystallogr*. 62:72–82.

Ficko-Blean E, Boraston AB. 2006. The Interaction of a Carbohydrate-binding Module from a *Clostridium perfringens N*-Acetyl-β-hexosaminidase with Its Carbohydrate Receptor. *J Biol Chem*. 281:37748–37757

Ficko-Blean E, Boraston AB. 2009. *N*-acetylglucosamine recognition by a family 32 carbohydrate-binding module from *Clostridium perfringens* NagH. *J Mol Biol*. 390:208–220

Ficko-Blean E, Stuart CP, Suits MD, Cid M, Tessier M, Woods RJ, Boraston AB. 2012. Carbohydrate recognition by an architecturally complex α-*N*-acetylglucosaminidase from *Clostridium perfringens*, *PLoS one*, **7**: e33524

Frukawa T, Youssef EM, Yatsuoka T, Yokoyama T, Makino N, Inoue H, Fukushige S, Hoshi M, Hayashi Y, Sunamura M, Horii A. 1999. Cloning and characterization of the human UDP-N-acetylglucosamine: α-1,3-D-mannoside β-1,4-*N*-acetylglucosaminyltransferase IV-homologue (hGnT-IV-H) gene. *J Hum Genet*. 44:397–401.

Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A. 2005. Protein Identification and Analysis Tools on the ExPASy Server. In Walker JM editor: The Proteomics Protocols Handbook, Humana Press, p. 571–607.

Gerken TA, Revoredo L, Thome JJC, Tabak LA, Vester-Christensen MB, Clausen H, Gahlay GK, Jarvis DL, Johnson RW, Moniz HA, Moremen K. 2013. 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.

Grondin JM, Chitayat S, Ficko-Blean E, Czjzek M, Boraston AB, Smith SP. 2014. Solution Structure and Dynamics of Full-length GH84A, a multimodular β-*N*-acetylglucosaminidase from *Clostridium perfringens*. *Journal of Molecular Biology*. 426:869–880

Grondin JM, Duan D, Kirlin AC, Abe KT, Chitayat S, Spencer HL, Spencer C, Campigotto A, Houliston S, Arrowsmith CH, Allingham JS, Boraston AB, Smith SP. 2017. Diverse modes of

galacto-specific carbohydrate recognition by a family 31 glycoside hydrolase from *Clostridium perfringens*. *PLoS One*. 12:e0171606

Helenius A and Aebi M. 2001. Intracellular Functions of *N*-Linked Glycans. *Science*. 291: 2364–2369

Holm L. 2020. Using Dali for Structural Comparison of Proteins. *Methods Mol Biol*. 2112: 29–42

Huang HH, Stanley P. 2010. A testis-specific regulator of complex and hybrid *N*-glycan synthesis. *J Cell Biol*. 190:893–910

Hykollari A, Malzl D, Eckmair B, Vanbeselaere J, Scheidl P, Jin C, Karlsson NG, Wilson IBH, Paschinger K. 2018. Isomeric Separation and Recognition of Anionic and Zwitterionic N-glycans from Royal Jelly Glycoproteins. *Mol Cell Proteomics*. 17:2177–2196

Hykollari A, Malzl D, Stanton R, Eckmair B, Paschinger K. 2019. Tissue-specific glycosylation in the honeybee: Analysis of the N-glycomes of *Apis mellifera* larvae and venom. *Biochim Biophys Acta Gen Subj*. 1863:129409

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. 2021. Highly accurate protein structure prediction with

AlphaFold. *Nature*. 596:583–589.

Kabsch W. 2010. XDS. *Acta Crystallogr D Biol Crystallogr*. 66:125–132.

Kadirvelraj R, Yang JY, Sanders JH, Liu L, Ramiah A, Prabhakar PK, Boons GJ, Wood ZA, Moremen KW. 2018. Human *N*-acetylglucosaminyltransferase II substrate recognition uses a modular architecture that includes a convergent exosite. *Proc Natl Acad Sci U S A*. 115:4637–4642

Kawamoto M, Jouraku A, Toyoda A, Yokoi K, Minakuchi Y, Katsuma S, Fujiyama A, Kiuchi T, Yamamoto K, Shimada T. 2019. High-quality genome assembly of the silkworm, *Bombyx mori*. *Insect Biochem Mol Biol*. 107:53–62

Kimura Y, Tsumura K, Kimura M, Okihara K, Sugimoto H, Yamada H. 2003. First Evidence for Occurrence of Galβ1-3GlcNAcβ1-4Man Unit in *N*-Glycans of Insect Glycoprotein: β1-3Gal and β1-4GlcNAc Transferases Are Involved in *N*-Glycan Processing of Royal Jelly Glycoproteins. *Biosci Biotechnol Biochem*. 67: 1852–1856

Kizuka Y, Taniguchi N. 2016. Enzymes for *N*-Glycan Branching and Their Genetic and Nongenetic Regulation in Cancer. *Biomolecules*. 6:25

Koles K, Lim J-M, Aoki K, Porterfield M, Tiemeyer M, Wells L, Panin V. 2007. Identification of *N*-glycosylated proteins from the central nervous system of *Drosophila melanogaster*. *Glycobiology.* 17:1388–1403

Krogh A, Larsson B. Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol*. 305:567–580

Kurz S, Aoki K, Jin C, Karlsson NG, Tiemeyer M, Wilson IB, Paschinger K. 2015. Targeted release and fractionation reveal glucuronylated and sulphated N- and O-glycans in larvae of dipteran insects. *J Proteomics*. 126:172–188

Kuwabara N, Manya H, Yamada T, Tateno H, Kanagawa M, Kobayashi K, Akasaka-Manya K, Hirose Y, Mizuno M, Ikeguchi M, Toda T, Hirabayashi J, Senda T, Endo T, Kato R. 2016. Carbohydrate-binding domain of the POMGnT1 stem region modulates *O*-mannosylation sites of α-dystroglycan. *Proc Natl Acad Sci U S A*. 113:9280–9285.

Minowa M, Oguri S, Yoshida A, Hara T, Iwamatsu M, Ikenaga H, Takeuchi M. 1998. cDNA Cloning and Expression of Bovine UDP-*N*-acetylglucosamine:α1,3-D-Mannoside β1,4-*N*-Acetylglucosaminyltransferase IV. *J Biol Chem*. 273:11556–11562

Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. 2022. ColabFold: making protein folding accessible to all. *Nat Methods*. 19:679–682

Miyazaki T, Miyashita R, Mori S, Kato T, Park EY. 2019a. Expression and characterization of silkworm *Bombyx mori* β-1,2-*N*-acetylglucosaminyltransferase II, a key enzyme for complex-type *N*-glycan biosynthesis. *J Biosci Bioeng*. 127:273–280

Miyazaki T, Miyashita R, Nakamura S, Ikegaya M, Kato T, Park EY. 2019b. Biochemical

characterization and mutational analysis of silkworm *Bombyx mori* β-1,4-*N*-acetylgalactosaminyltransferase and insight into the substrate specificity of β-1,4-galactosyltransferase family enzymes. *Insect Biochem Mol Biol*. 115:103254.

Mizutani K. Fernandes VO, Karita S, Luís AS, Sakka M, Kimura T, Jackson A, Zhang X, Fontes CMGA, Gilbert HJ, Sakka K. 2012. Influence of a Mannan Binding Family 32 Carbohydrate Binding Module on the Activity of the Appended Mannanase. *Appl Environ Microbiol*. 78: 4781–4787

Mizuochi T, Nishimura S, Derappe C, Taniguchi T, Hamamoto T, Mochizuki M, Kobata A. 1983. Structures of the asparagine-linked sugar chains of human chorionic gonadotropin produced in choriocarcinoma. Appearance of triantennary sugar chains and unique biantennary sugar chains. *J. Biol. Chem*. 258:14126–14129

Murshudov GN, Skubak P, Lebedev AA, Pannu NS, Steiner RA, Nicholls RA, Winn MD, Long F, Vagin AA. 2011. REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr D Biol. Crystallogr*. 67:355–367

Nagae M, Yamaguchi Y, Taniguchi N, Kizuka Y. 2020. 3D Structure and Function of Glycosyltransferases Involved in *N*-glycan Maturation. *Int J Mol Sci*. 21: 437

Nagae M, Kizuka Y, Mihara E, Kitago Y, Hanashima S, Ito Y, Takagi J, Taniguchi N, Yamaguchi Y. 2018. Structure and mechanism of cancer-associated N-acetylglucosaminyltransferase-V. *Nat Commun*. 9:3380

Nishihara S. 2018. Glycans in stem cell regulation: from Drosophila tissue stem cells to mammalian pluripotent stem cells. *FEBS Lett*. 592:3773–3790

Oguri S, Minowa T, Ihara Y, Taniguchi N, Ikenaga H, Takeuchi M. 1997. Purification and characterization of UDP-*N*-acetylglucosamine: alpha1,3-D-mannoside beta1,4-*N*-acetylglucosaminyltransferase (*N*-acetylglucosaminyltransferase-IV) from bovine small intestine. *J Biol Chem*. 272:22721–22727

Oguri S, Yoshida A, Minowa M, Takeuchi M. 2006. Kinetic properties and substrate specificities of two recombinant human *N*-acetylglucosaminyltransferase-IV isozymes. *Glycoconj J*. 23:473–480

Painter J, Merritt EA. 2006. Optimal description of a protein structure in terms of multiple groups undergoing TLS motion. *Acta Crystallogr D Biol Crystallogr*. 62:439–450

Perrozzo R, Folkers G, Scapozza L. 2009. Thermodynamics of protein–ligand interactions: history, presence, and future aspects. *J Recept Signal Tranduct Res*. 24:1–52

Pluvinage B, Ficko-Blean E, Noach I, Stuart C, Thompson N, McClure H, Buenbrazo N, Wakarchuk W, Boraston AB. 2021. Architecturally complex *O*-glycopeptidases are customized for mucin recognition and hydrolysis. *Proc Natl Acad Sci U S A*. 118:e2019220118

Robert X and Gouet P. 2014. Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res*. 42:320–324

Sakamoto Y, Taguchi T, Honke K, Korekane H, Watanabe H, Tano Y, Dohmae N, Takio K, Horii A, Taniguchi N. 2000. Molecular cloning and expression of cDNA encoding chicken UDP-N-acetyl-D-glucosamine (GlcNAc): GlcNAcβ1-6(GlcNAcβ1-2)-Manα1-R[GlcNAc to Man]β 1,4*N*-acetylglucosaminyltransferase VI. *J Biol Chem*. 275:36029–36034

Shinya S, Nishimura S, Kitaoku Y, Numata T, Kimoto H, Kusaoke H, Ohnuma T, Fukamizo T. 2016. Mechanism of chitosan recognition by CBM32 carbohydrate-binding modules from a Paenibacillus sp. IK-5 chitosanase/glucanase. *Biochem J*. 473:1085–1095

Shinya S, Ohnuma T, Yamashiro R, Kimoto H, Kusaoke H, Anbazhagan P, Juffer AH, Fukamizo T. 2013. The first identification of carbohydrate binding modules specific to chitosan. *J Biol Chem*. 288:30042–30053

Sievers F, Wilm A, Dineen D, Gibson TG, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol*. 11:539

Stanton R, Hykollari A, Eckmair B, Malzl D, Dragosits M, Palmberger D, Wang P, Wilson IB, Paschinger K. 2017. The underestimated N-glycomes of lepidopteran species. *Biochim Biophys Acta Gen Subj*. 1861:699–714

Taguchi T, Ogawa T, Inoue S, Inoue Y, Sakamoto Y, Korekane H, Taniguchi N. 2000. Purification and characterization of UDP-GlcNAc: GlcNAcβ1-6(GlcNAcβ1-2)Manα1-R [GlcNAc to Man]-β1,4-N-acetylglucosaminyltransferase VI from hen oviduct. *J Biol Chem*. 275:32598–32602

Turnbull WB, Daranas AH. 2003. On the value of *c*: Can low affinity aystems be studied by isothermal titration calorimetry? *J Am Chem Soc*. 125:14859–14866.

Unligil UM, Zhou S, Yuwaraj S, Sarkar M, Schachter H, Rini JM. 2000. X-ray crystal structure of rabbit N-acetylglucosaminyltransferase I: catalytic mechanism and a new protein superfamily. *EMBO J.* 19:5269–5280

Vagin A, Teplyakov A. 2010. Molecular replacement with MOLREP. *Acta Crystallogr D Biol Crystallogr.* 66:22–25

Varki A, Cummings RD, Esko JD, Freeze HH, Stanley P, Marth JD, Bertozzi CR, Hart GW, Etzler ME. 2009. Symbol nomenclature for glycan representation. *Proteomics*. 9:5398–5399.

Varki A. 2017. Biological roles of glycans. *Glycobiology* 27:3–49

Walski T, Damme EJMV, Smargiasso N, Christiaens O, Pauw ED, Smagghe G. 2016. Protein *N*-glycosylation and *N*-glycan trimming are required for postembryonic development of the pest beetle *Tribolium castaneum. Sci Rep*. 6:35151

Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, de Beer TAP, Rempfer C, Bordoli L, Lepore R, Schwede T. 2018. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res*. 46:W296–W303

Winn MD, Ballard CC, Cowtown KD, Dodson EJ, Emsley P, Evans PR, Keegan RM, Krissinel

EB, Leslie AGW, McCoy A, McNicholas SJ, Murshudov GN, Pannu NS, Potterton EA, Powell HR, Read RJ, Vagin A, Wilson KS. 2011. Overview of the CCP4 suite and current developments. *Acta Crystallogr D Biol Crystallogr*. 67:235–242

Wiseman T, Williston S, Brandts JF, Lin LN. 1989. Rapid measurement of binding constants and heats of binding using a new titration calorimeter. *Anal Biochem*. 179: 131−137

Yoshida A, Minowa MT, Takamatsu S, Hara T, Ikenaga H, Takeuchi M. 1998. A novel second isoenzyme of the human UDP-N-acetylglucosamine:α1,3-D-mannoside β1,4-N-acetylglucosaminyltransferase family: cDNA cloning, expression, and chromosomal assignment. *Glycoconj J*. 15:1115–1123

Yoshida A, Minowa MT, Takamatsu S, Hara T, Oguri S, Ikenaga H, Takeuchi M. 1999. Tissue specific expression and chromosomal mapping of a human UDP-N-acetylglucosamine: α1,3-D-mannoside β1, 4-*N*-acetylglucosaminyltransferase. *Glycobiology*. 9:303–310

Figure legends

Fig. 1. Schematic of GnT-IV reaction.

GnT-IV transfers GlcNAc from a uridine diphosphate (UDP)-GlcNAc donor to the α-1,3 mannose branch of the *N*-glycan core structure pre-attached with β-1,2-GlcNAc, forming a β-1,4 linkage. Sugar symbols are used according to Varki et al. (2009).

Fig. 2. HsGnT-IVa and BmGnT-IV domain organization and recombinant expression.

(A) The domains of HsGnT-IVa and BmGnT-IV are based on the AlphaFold2 models (Fig. S1). The transmembrane regions were predicted using TMHMM. (B) SDS-PAGE analysis of HsCBML and BmCBML. Lane M, PM1700 All Blue Broad Excel Band Range Protein Marker (SMOBIO, Tokyo, Japan); lane 1, HsCBML; lane 2, BmCBML.

Fig. 3. Sugar-binding analysis of CBML using ITC.

HsCBML in 50-mM HEPES-NaOH buffer (pH 7.0) was titrated with GlcNAc (A), pNP-β-GlcNAc (B), *N*,*N*'-diacetylchitobiose (C), and pNP-α-GlcNAc (D) using MicroCal iTC200. BmCBML in the same buffer was titrated with GlcNAc (E) and pNP-α-GlcNAc (F). The results were calculated and plotted using MicroCal Origin ITC.

Fig. 4. Crystal structures of HsCBML and BmCBML

(A and B) Ribbon models of HsCBML (A) and BmCBML (B). The β-sheets and α-helices are shown in cyan and red, respectively. GlcNAc bound to BmCBML is shown as a magenta stick model. (C) Superimposition of HsCBML (pink) and BmCBML (green). (D) Superimposition of BmCBML (green) and CpNagHCBM32-2 (orange). GlcNAcβ-1,2-Man and Ca^{2+} ion bound to CpNagHCBM32-2 are shown in slate blue-stick and gray-sphere models, respectively.

Fig. 5. GlcNAc-binding site of CBML.

(A) F_o-F_c omit map for GlcNAc in BmCBML. (B and C) GlcNAc-binding sites of unliganded BmCBML (B) and GlcNAc-bound BmCBML (C). (D) Comparison of the sugar-binding sites of HsCBML and BmCBML. Side chains interacting with GlcNAc (magenta) in BmCBML and the corresponding residues in HsCBML are represented as green and pick-stick models, respectively. (E) Multiple sequence alignment of GnT-IV homologs in various animals and CpNagHCBM32-2 generated using ESPript 3. First two letters in each protein name mean acronyms of scientific names: Hs, *Homo sapiens*; Bt, *Bos taurus*; Mm, *Mus musculus*; Gg, *Gallus gallus*; Xl, *Xenopus laevis*; Dr, *Danio rerio*; and Bm, *Bombyx mori*. Green arrows indicate the amino acid residues involved in the GlcNAc-binding of BmCBML.

Fig. 6. Comparison of sugar-binding sites of BmCBML and CpNagHCBM32-2.

(A) Structure of glycan product of GnT-IVa and GnT-IVb. (B) Superimposition of ligand-binding sites of BmCBML (green) in complex with GlcNAc (magenta) and CpNagHCBM32-2 (orange) in complex with GlcNAcβ-1,2-Man (slate blue) (PDB 2WDB). Symbols of GlcNAcβ-1,2-Man are shown in a dotted box. (C) Surface model of BmCBML. GlcNAcβ-1,2-Man was superimposed with PDB 2WDB. The molecular surfaces of aromatic residues involved in GlcNAc recognition are highlighted in pink. Dashed arrows in the right panel of (C) indicate the anomer orientations of the reducing-end mannose residue of GlcNAcβ-1,2-Man.

Table I. Binding constants of CBML and CBM32 proteins toward sugars determined by ITC

Name	Ligand	K_a (M^{-1})	ΔG (kcal/mol)	ΔH (kcal/mol)	$-T\Delta S$ (kcal/mol)	c value
HsCBML	GlcNAc	2.87 (± 0.58) $\times 10^2$	−1.43 ± 0.40	−1.26 ± 0.45	−0.173 ± 0.05	0.086
	pNP-β-GlcNAc	1.97 (± 0.15) $\times 10^3$	−5.21 ± 0.45	−5.28 ± 0.50	0.066 ± 0.050	0.591
	GlcNAc-β-1,4-GlcNAc	1.27 (± 0.07) $\times 10^3$	−2.84 ± 0.34	−2.71 ± 0.37	−0.127 ± 0.03	0.381
BmCBML	GlcNAc	2.42 (± 0.74) $\times 10^2$	−0.961 ± 0.26	−0.802 ± 0.29	−0.198 ± 0.04	0.065
CpNagHCBM32-2 [a]	GlcNAc	1.88×10^3				
	GlcNAc [e]	2.22×10^3				
	GlcNAc-β-1,4-GlcNAc [e]	1.25×10^3				
	GlcNAc-β-1,2-Man [e]	7.86×10^3				
	GlcNAc-β-1,3-Man [e]	3.98×10^3				
	GlcNAc-β-1,3-GalNAc	5.91×10^3				
PeCBM32 [b]	GlcNAc-β-1,4-GlcNAc	9.0×10^3				
	GlcN-β-1,4-GlcNAc	1.4×10^5				
EfCBM32 [c]	GalNAc	2.66×10^2				
BmCBM32 [d]	GalNAc	2.60×10^2				

Each experiment was performed in triplicate.

[a] CBM32-2 domain of *Clostridium perfringens* GH84 β-*N*-acetylglucosaminidase NagH (Ficko-Blean et al. 2009)

[b] CBM32 domain of *Paenibacillus elgii* GH8 chitosanase (Das et al. 2016)

[c] CBM32 domain of *Enterococcus faecalis* GH31 α-*N*-acetylgalactosaminidase (Ikegaya et al. 2021)

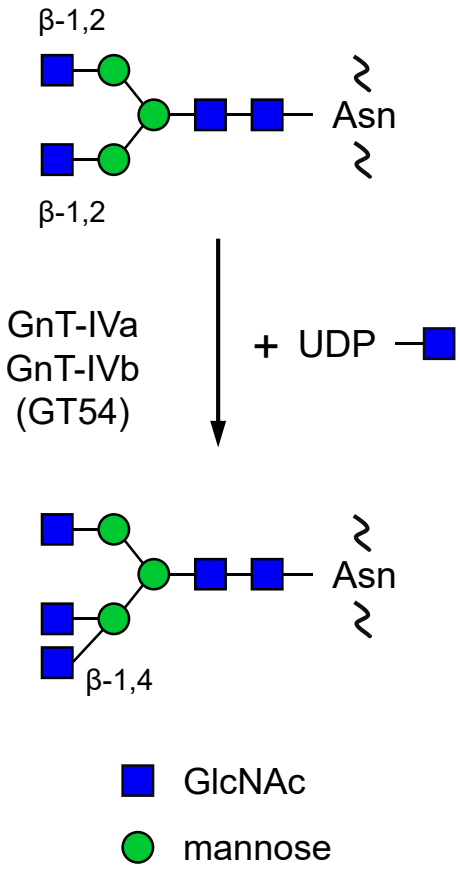
[d] CBM32 domain of *B. mori* GH31 α-*N*-acetylgalactosaminidase (Ikegaya et al. 2021)

[e] These constants were determined by UV difference titration.

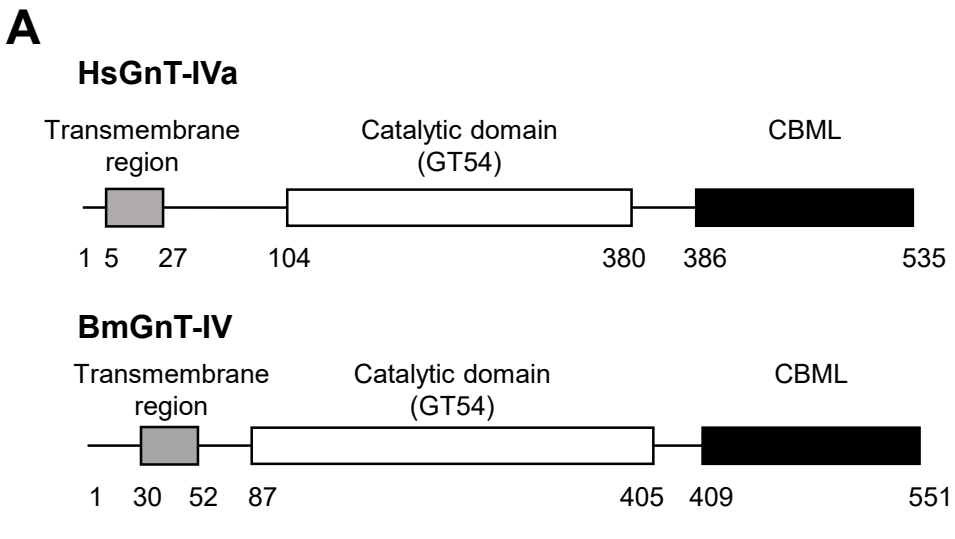
Table II. Data collection and refinement statistics of CBMLs

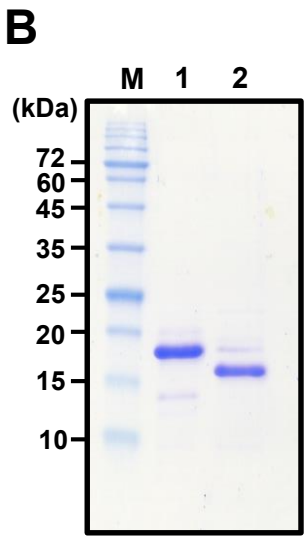
	HsCBML	BmCBML	BmCBML-GlcNAc
Data collection			
Beamline	BL-5A	AR-NW12A	AR-NW12A
Wavelength (Å)	1.0000	1.0000	1.0000
Space group	$P2_1$	$P2_1$	$P2_1$
Cell dimensions			
a, b, c (Å)	50.7, 32.5, 85.6	32.7, 64.5, 78.8	32.7, 64.1, 79.2
β (º)	94.7	95.1	95.9
Resolution range (Å)	50–1.97 (2.08–1.97)	39.25–1.47 (1.55–1.47)	39.37–1.15 (1.21–1.15)
Measured reflections	256528	343484	755968
Unique reflections	20101	53654	113597
Completeness (%)	100 (100)	96.6 (83.0)	98.6 (94.9)
Redundancy	12.8 (13.0)	6.4 (5.0)	6.7 (6.6)
Mean $I/\sigma(I)$	8.4 (2.4)	18.2 (2.7)	18.1 (2.3)
R_{merge}	0.207 (0.822)	0.052 (0.526)	0.046 (0.810)
$CC_{1/2}$	0.990 (0.951)	0.999 (0.841)	0.999 (0.782)
Refinement statistics			
R_{work}/R_{free}	0.220/0.266	0.224/0.248	0.193/0.209
RMSD			
Bond length (Å)	0.014	0.012	0.011
Bond angles (º)	1.704	1.729	1.677
Number of atoms			
Protein	A 1179, B 1158	A 1167, B 1190	A 1303, B 1233
Ligand/Ion	–	A 16, B 8	A 27, B 19
Water	22	152	214
Average B (Å)			
Protein	A 57.7, B 73.6	A 21.7, B 23.7	A 19.4, B 19.9
Ligands	–	A 34.2, B 38.6	A 29.3, B 25.4
Water	55.6	28.4	27.8
Ramachandran plot			
Favored (%)	97.2	97.1	97.0
Outliers (%)	0.7	0	0
Clashscore	4.09	3.88	4.45
PDB code	7XTL	7XTM	7XTN

Oka, Mori et al. Fig. 1

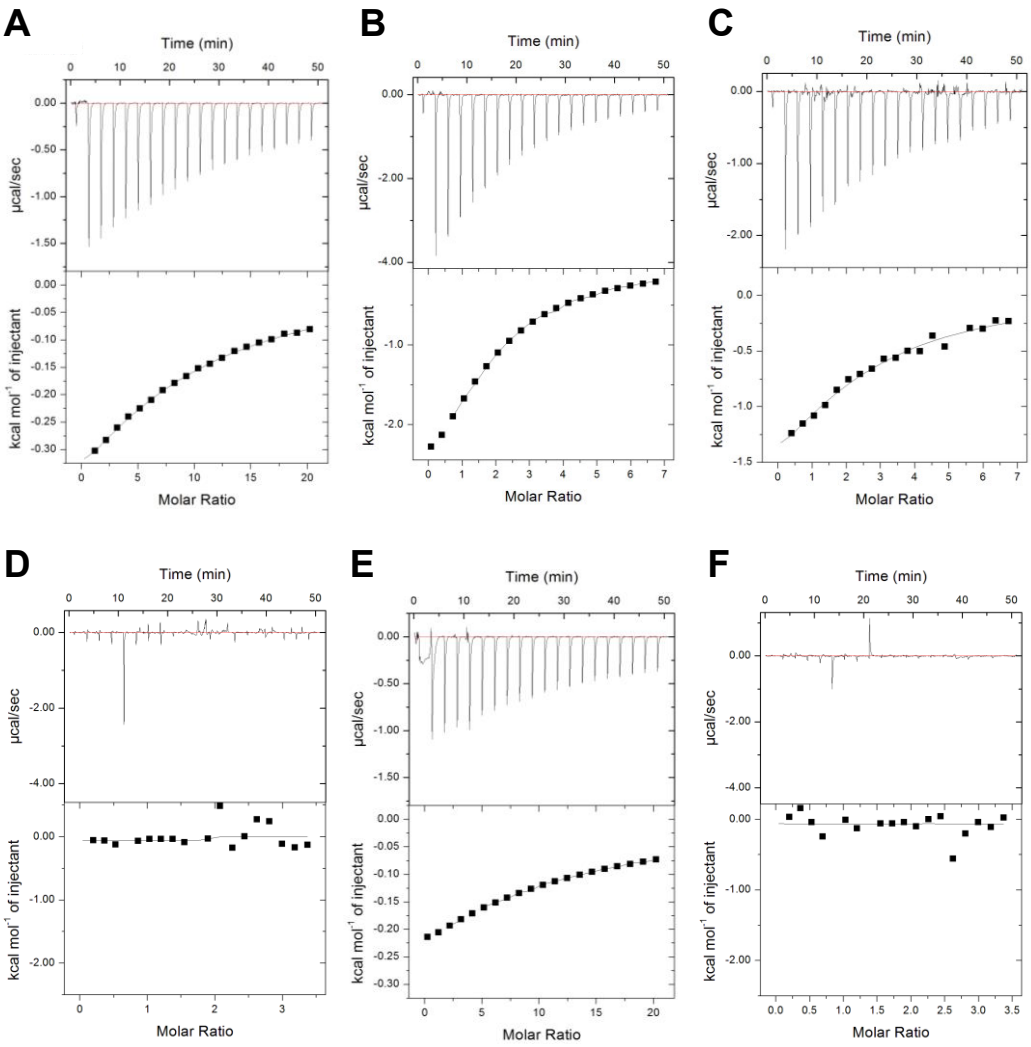


Oka, Mori et al. Fig. 2

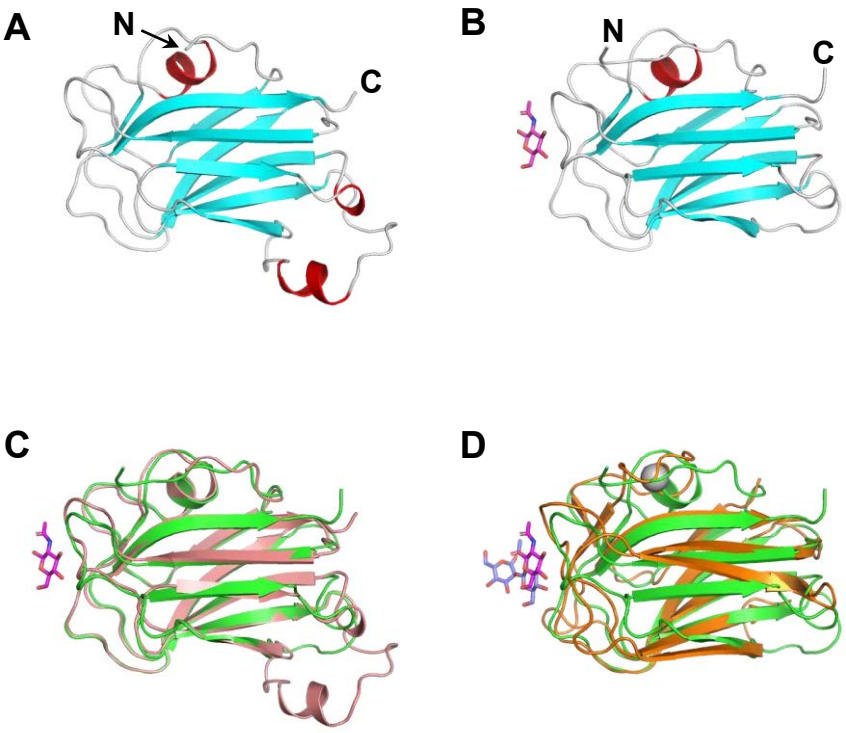




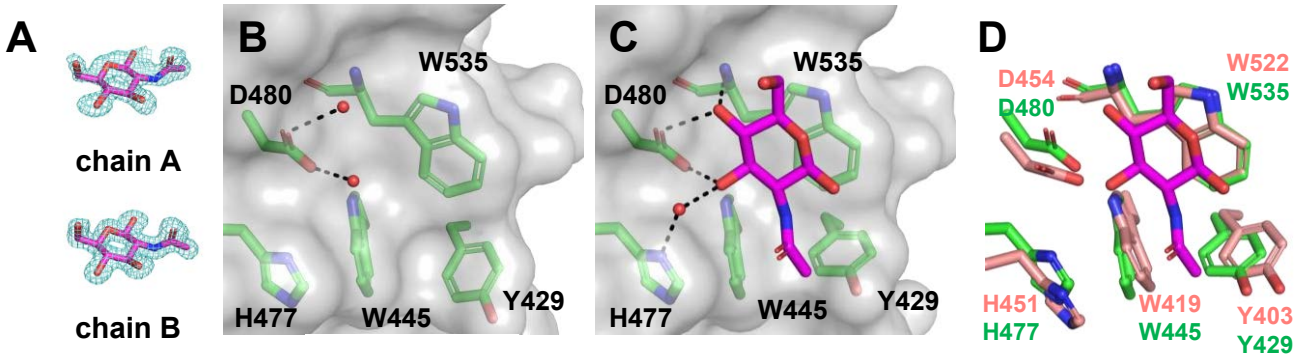
Oka, Mori et al. Fig. 3



Oka, Mori et al. Fig. 4



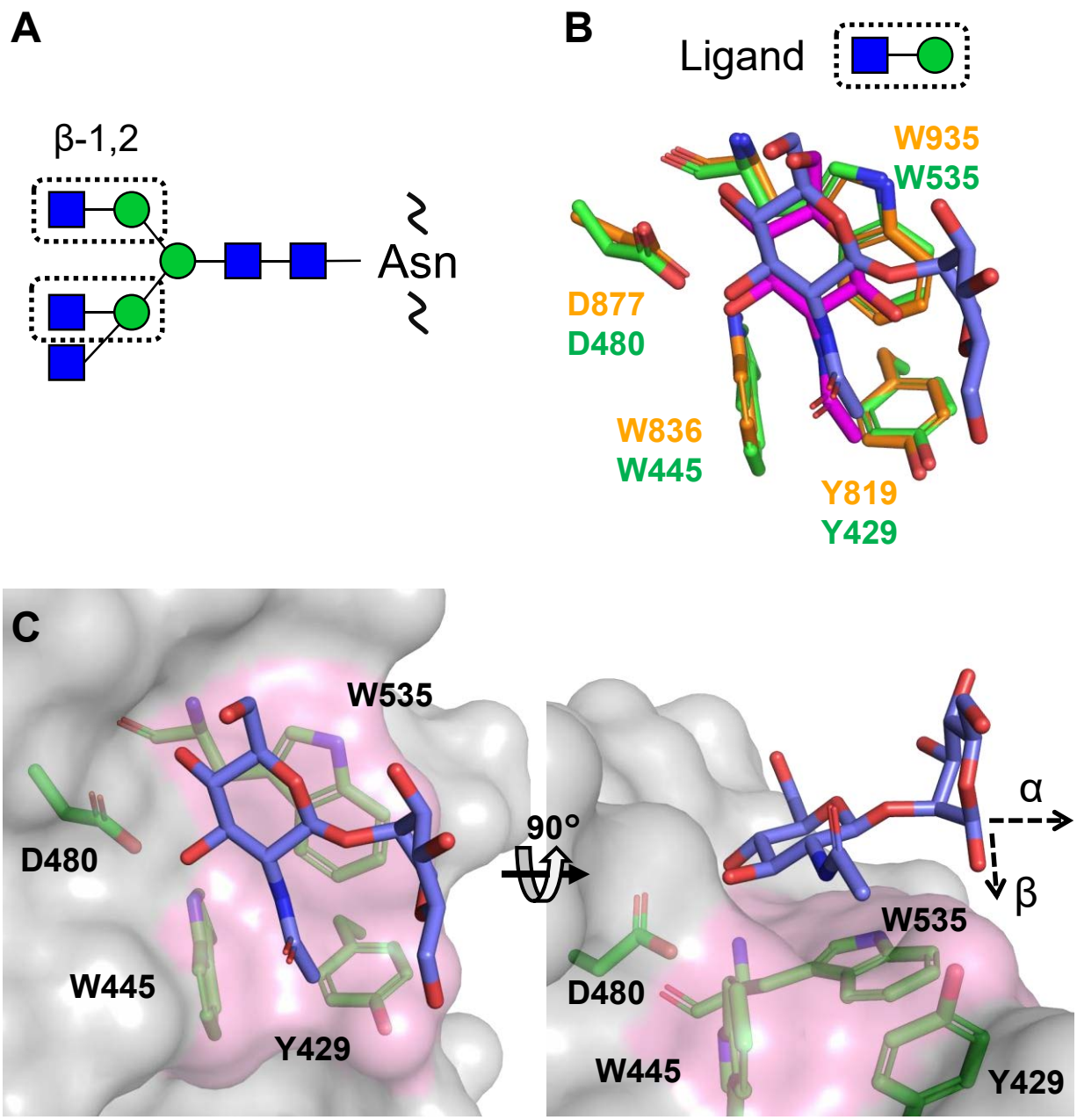
Oka, Mori et al. Fig. 5



E

```
HsGnT-IVa                  TT   α1      TT  β2      TT      TT TT                  β
                               410           420           450           520
HsGnT-IVa      402 VYQGHTLEKTYM 418 FWAITPIAGD 448 NQEHPGDILL 517 QNSAVWAILNE
HsGnT-IVb      425 TYQHFTLEKAYL 441 FWAFTPAAGD 470 NIEHPEDKLF 504 TDSPVWVILSE
BtGnT-IVa      403 VYQGHTLEKTYM 419 FWAITPVAGD 448 NQDHPGDILL 517 QNSAVWAILNE
MmGnT-IVa      403 VYQGHTLEKTYM 419 FWAITPTAGD 448 NQEHPGDILL 517 QNSAVWAILNE
GgGnT-VI       339 VFENYEPSKAYS 355 FWGKNPAVGS 384 SSERPGDFLH 418 QSQSEWLIIQS
XlGnT-IVa      403 VYQGHTLEKTYL 419 FWAITPVAGD 448 NPEHPGDQLW 517 QNSPVWVILNE
DrGnT-IVa      402 VYQGHTLEKTYM 418 FWAITPMAGD 447 NQEHPGDKIE 511 KDSAVWAIISE
BmGnT-IV       428 DYYQHSIQNAYE 444 FWGKKPKKGD 474 NAEHITDQFY 530 ADSKYWVILSE
CpNagHCBM32-2  818 VY.EGNEANLLD 835 VW.YKTLNGD 871 NGGGSSDKWN 930 ENINKWLTFSE
```

Oka, Mori et al. Fig. 6



Supporting information

Crystal structure and sugar-binding ability of the C-terminal domain of *N*-acetylglucosaminyltransferase IV establish a new carbohydrate-binding module family.

Nozomi Oka[1*], Sota Mori[1*], Marina Ikegaya[2], Enoch Y. Park[1,2,3], and Takatsugu Miyazaki[1,2,3**]

[1] Department of Agriculture, Graduate School of Integrated Science and Technology, Shizuoka University, Shizuoka, Japan.
[2] Department of Bioscience, Graduate School of Science and Technology, Shizuoka University, Shizuoka, Japan.
[3] Research Institute of Green Science and Technology, Shizuoka University, 836 Ohya, Suruga-ku, Shizuoka, 422-8529, Japan.

* These authors contributed equally.

****Correspondence:** Takatsugu Miyazaki, Research Institute of Green Science and Technology, Shizuoka University, 836 Ohya, Suruga-ku, Shizuoka, 422-8529, Japan; Tel: +81-54-238-4886; E-mail: miyazaki.takatsugu@shizuoka.ac.jp

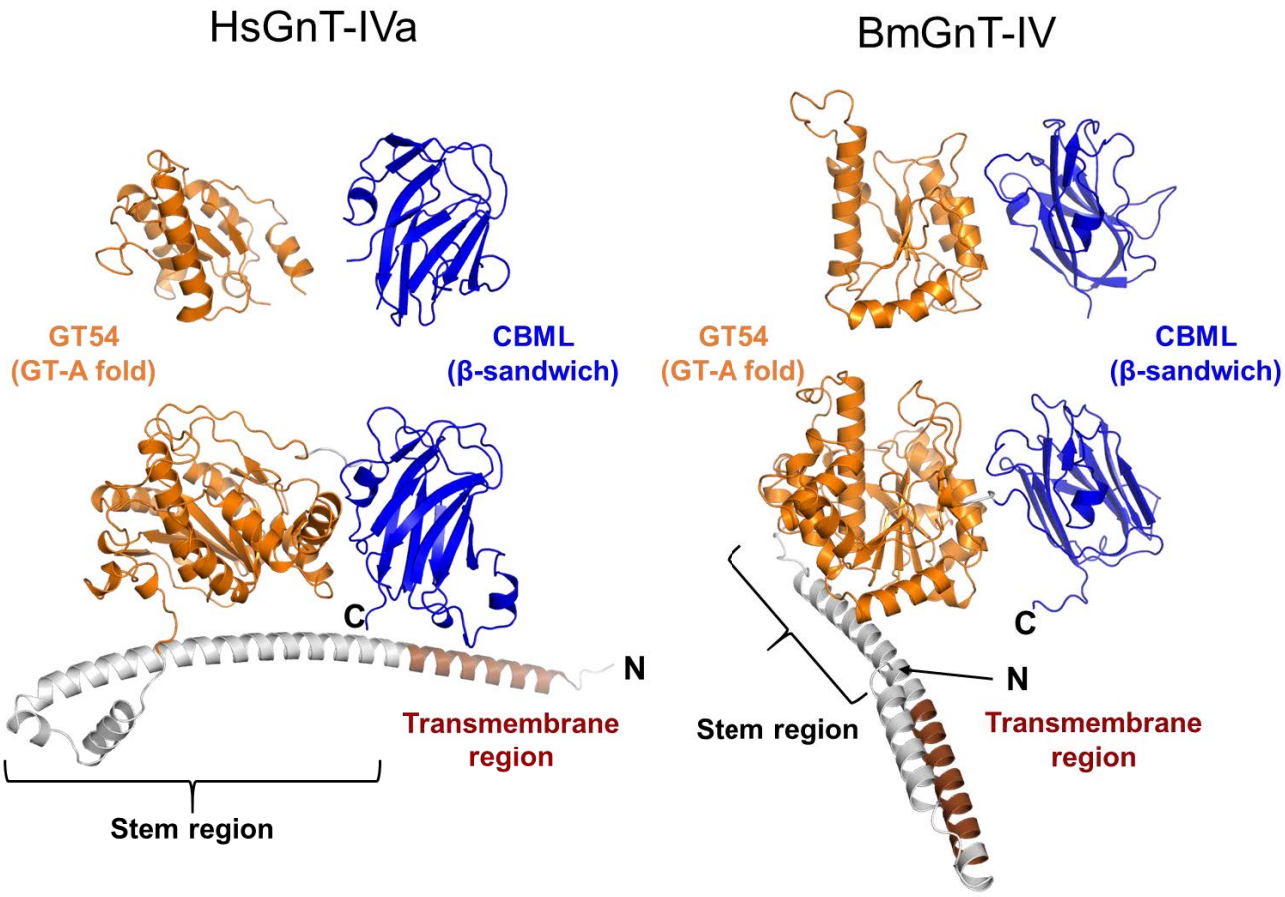


Figure S1. Structure modeling of HsGnT-IVa and BmGnT-IV using SWISS-MODEL and AlphaFold2.

Each partial homology model generated using SWISS-MODEL is shown in the upper row, and each full-length AlphaFold2 model is shown in the lower row. The predicted transmembrane regions, catalytic domains, and CBML domains are shown in brown, orange, and blue, respectively.

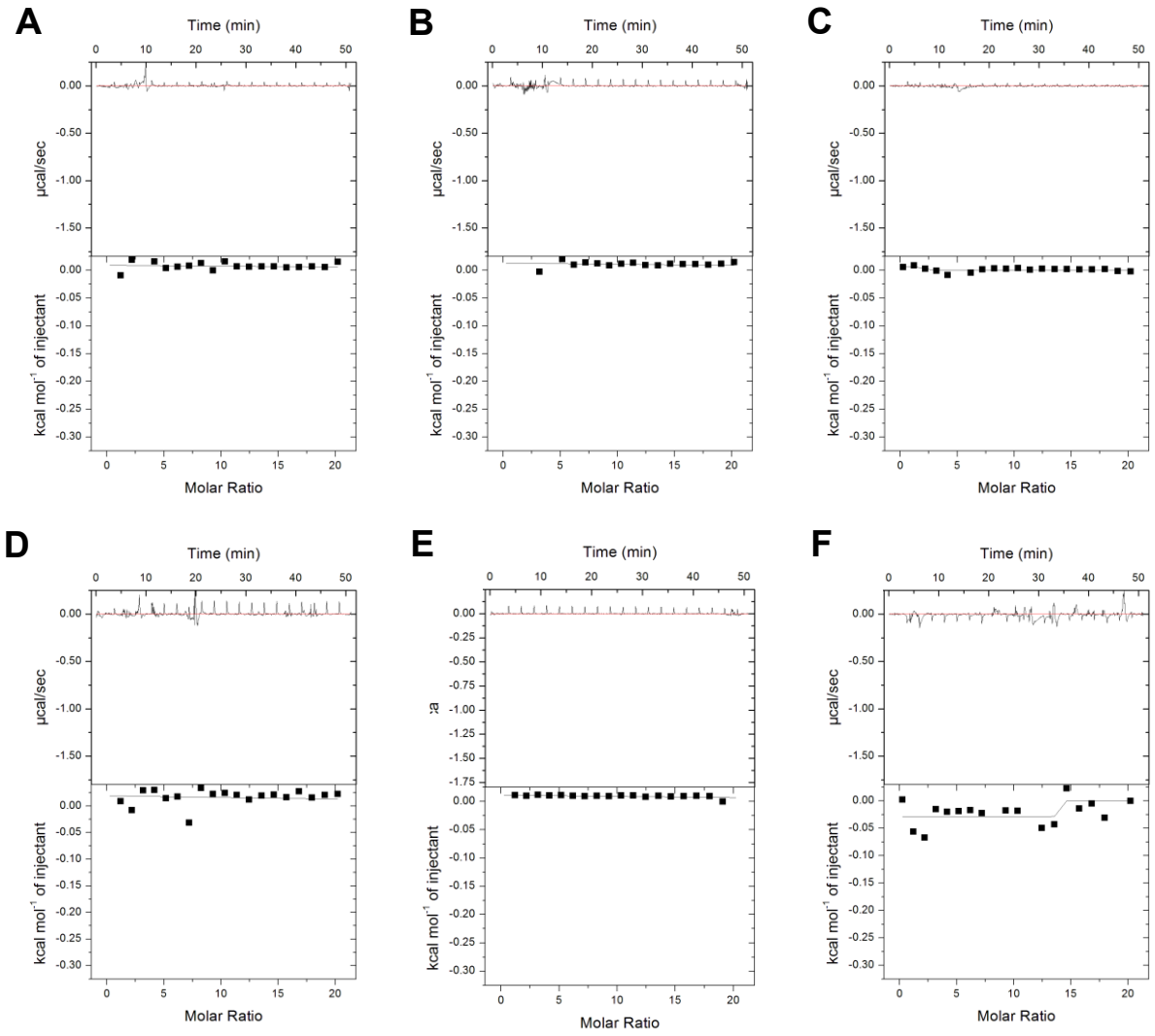


Figure S2. ITC results of sugars which CBMLs did not bind.

(a–c) HsCBML was titrated with mannose (a), glucose (b), and *N*-acetylgalactosamine (c). (d–f) BmCBML was titrated with mannose (d), glucose (e), and *N*-acetylgalactosamine (f).

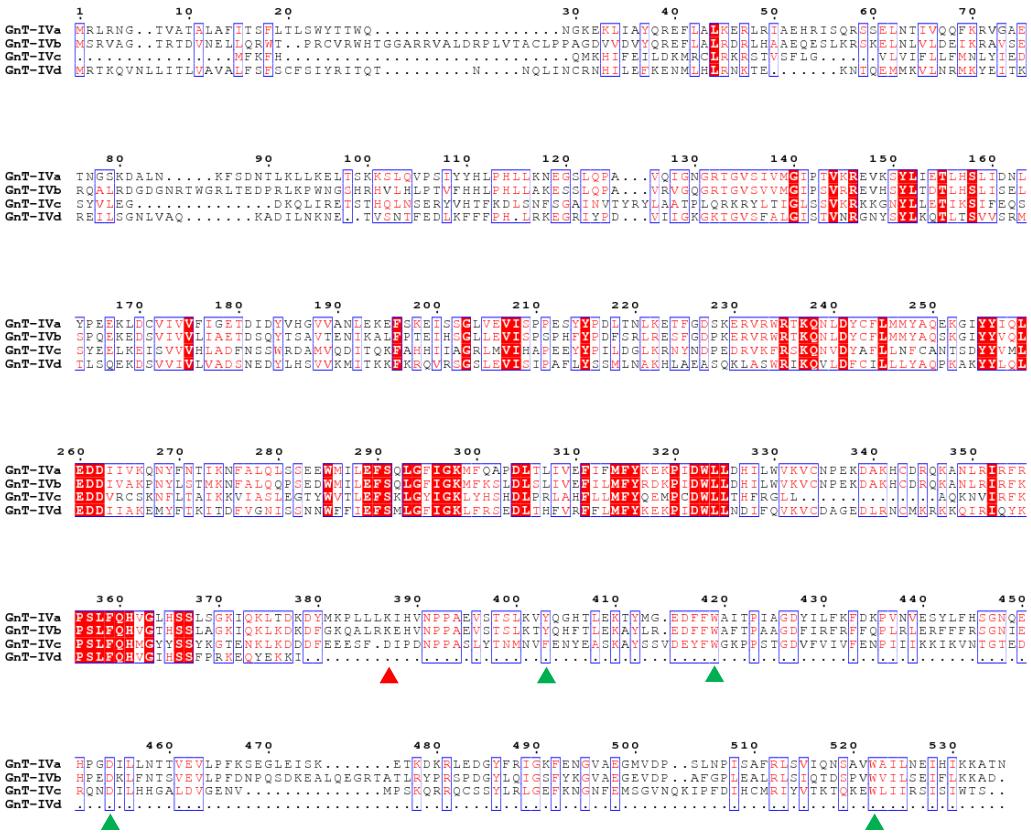


Figure S3. Sequence alignment of human GnT-IVa, -IVb, -IVc, and -IVd.

The alignment was performed using Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/) and visualized using ESPript 3. The red triangle indicates the amino acid residue at the start of CBML domain, and the green triangle indicates the amino acid residue involved in GlcNAc-binding.

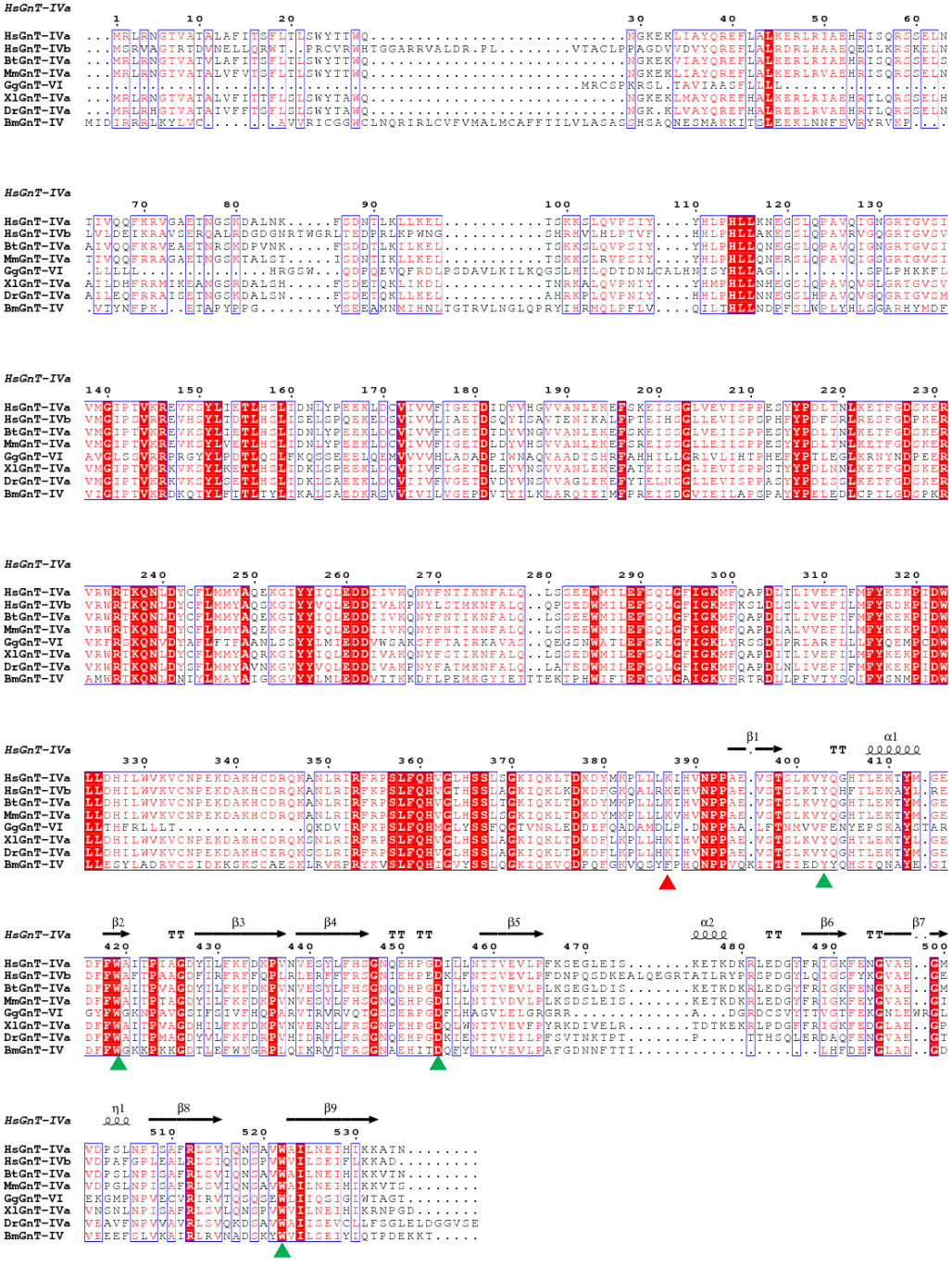


Figure S4. Sequence alignment of GnT-IV homologs in various species.

The alignment was performed using Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/) and visualized using ESPript 3. The red triangle indicates the amino acid residue at the start of CBML domain, and the green triangle indicates the amino acid residue involved in GlcNAc-binding. First two letters in each protein name mean acronyms of scientific names: Hs, *Homo sapiens*; Bt, *Bos taurus*; Mm, *Mus musculus*; Gg, *Gallus gallus*; Xl, *Xenopus laevis*; Dr, *Danio rerio*; and Bm, *Bombyx mori*.

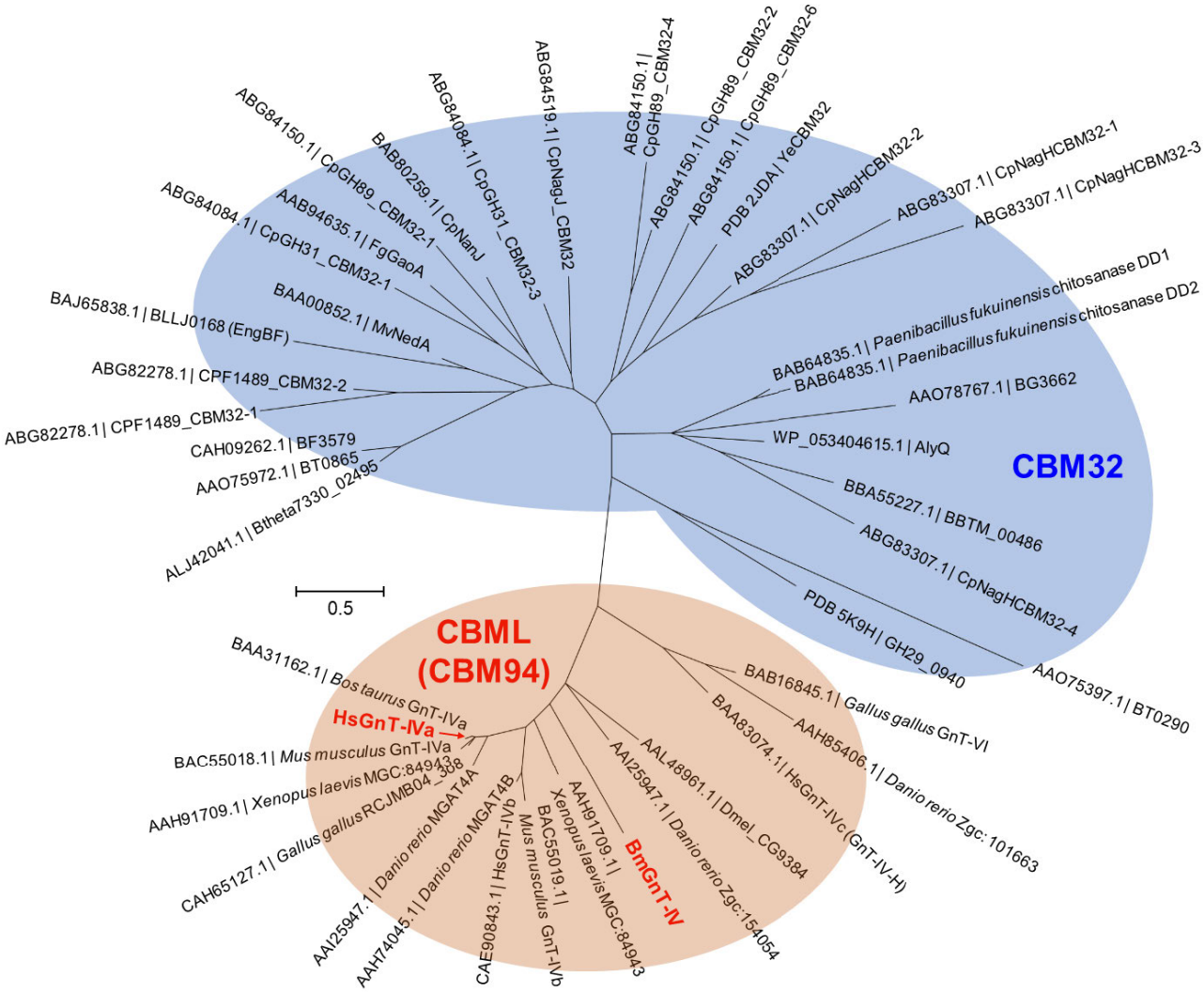


Figure S5. Phylogenetic tree of CBML domains of GnT-IV homologs and CBM32 proteins.

The sequence alignment was performed using the MUSCLE algorithm with amino acid sequences of CBML domains of GnT-IV homologs and the structure-determined CBM32 proteins listed in the CAZy database. The tree was generated using the maximum likelihood method conducted by MEGA X. GenBank IDs (or PDB IDs) and protein names are labeled in each branch. CBML domains of HsGnT-IVa and BmGnT-IV are highlighted in red.

Table S1. Structural homologs of HsCBML found using Dali server.

Protein	Origin	Sugar ligand	PDB	Z score	RMSD (Å)	Sequence identity (%)
C-terminal domain of GH29 α-fucosidase	*Streptococcus pneumoniae*	−	6ORG	12.6	2.6	12
CBM32 of GH84 *N*-acetylglucosaminidase NagH	*Clostridium perfringens*	GlcNAc-Man	2W1U	12.4	2.4	19
CBM32 DD1 of GH8 chitosanase/glucanase	*Paenibacillus fukuinensis*	GlcN	4ZY9	12.2	2.8	10
C-terminal β-sandwich domain of GH29 α-fucosidase	*Bacteroides thetaiotaomicron*	−	4OZO	12.1	2.5	14
CBM32 of BT3015	*Bacteroides thetaiotaomicron*	Gal	7BLG	12.1	2.5	15
Intraflagellar transport 25 (GTPase)	*Chlamydomonas reinhardtii*	−	2YC4	11.9	2.8	13
CBM32 DD2 of GH8 chitosanase/glucanase	*Paenibacillus fukuinensis*	GlcN-GlcN-GlcN	4ZZ8	11.9	2.6	12
CBM47 SpX-3 of GH98 Lewis^{Y}-specific endo-β-1,4-galactosidase	*Streptococcus pneumoniae*	Fuc	2J22	11.8	2.6	7
CBM47 of lectinolysin	*Streptococcus mitis*	α-L-Fucp-(1-2)-β-D-Galp-(1-4)[α-L-Fucp-(1-3)]-β-D-GlcpNAc	3LEG	11.8	2.6	13
CBM6 of xylanase	*Clostridium thermocellum*	xylan	1GMM	5.4	3.0	13

Table S2. Structural homologs of BmCBML found using Dali server.

Protein	Origin	Sugar ligand	PDB	Z score	RMSD (Å)	Sequence identity (%)
CBM32 of GH84 *N*-acetylglucosaminidase	*Clostridium perfringens*	GlcNAc-Man	2WDB	12.7	2.4	20
C-terminal domain of GH29 α-fucosidase	*Streptococcus pneumoniae*	−	6ORF	12.6	2.4	13
CBM32 of chitosanase/glucanase	*Paenibacillus fukuinensis*	GlcN	4ZZ5	12.6	2.6	16
CBM32 of chitosanase/glucanase	*Paenibacillus fukuinensis*	GlcN	4ZY9	12.5	2.9	16
C-terminal β-sandwich domain of GH29 α-fucosidase	*Bacteroides thetaiotaomicron*	−	4OZO	12.5	2.7	15
CBM32-1 of GH31 protein	*Clostridium perfringens*	GalNAc	4LPL	12.3	2.6	8
CBM47 SpX-3 of GH98 Lewis^{Y}-specific endo-β-1,4-galactosidase	*Streptococcus pneumoniae*	Fuc	2J1U	12.1	2.7	12
CBM32 of exo-α-sialidase	*Clostridium perfringens*	Gal	2V72	12	2.6	9
CBM32 of *O*-glycopeptidases	*Clostridium perfringens*	GalNAc	7JNB	11.9	3.2	13
CBM6 of xylanase	*Clostridium thermocellum*	xylan	1GMM	5.4	3.0	13

Table S3. Sequences of oligonucleotides used in this study.

Name	Sequence
HsCBML_NheI _F	5’-TTTTGCTAGCAAAATCCATGTAAACCCACC-3’
HsCBML_XhoI_R	5’-CATATTAAAAAAGCCACCAACTGATCATCTGA-3’
BmGnT-IV_F	5’-ATTCGGATCCACCATGATTGATATTCGGCG-3’
BmGnT-IV_R	5’- AGATCGAAGTATTCAAGTTTTCTTTTCGTCCGGTG -3’
BmCBML_NdeI_F	5’- TTTTCATATGCAGTCGTATTTTCCACACCA -3’
BmCBML_HindIII_R	5’- TTTTAAGCTTTCAAGTTTTCTTTTCGTCCG-3’