Title

Structural and biochemical insight into a modular β-1,4-galactan synthase in plants

Journal

Nature Plants, 9(3)

ISSN

2055-026X

Authors

Prabhakar, Pradeep Kumar
Pereira, Jose Henrique
Taujale, Rahil
et al.

Publication Date

2023-03-01

DOI

10.1038/s41477-023-01358-4

Structural and Biochemical Insight into a Modular β-1,4-Galactan Synthase in Plants

Pradeep Kumar Prabhakar[1,2,3*], Jose Henrique Pereira[4,5*], Rahil Taujale[1], Wanchen Shao[5,6], Vivek Bharadwaj[7], Digantkumar Chapla[2], Jeong-Yeh Yang[2], Yannick J. Bomble[7], Kelley W. Moremen[1,2], Natarajan Kannan[1], Michal Hammel[4], Paul D. Adams[4,5,8], Henrik V. Scheller[5,6,9], Breeanna R. Urbanowicz[1,2,3,#]

[1]Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA 30602 USA.

[2]Complex Carbohydrate Research Center, University of Georgia, Athens, GA 30602 USA.

[3]Center for Bioenergy Innovation, Oak Ridge National Laboratory, Oakridge, TN 37830 USA.

[4]Molecular Biophysics and Integrated Bioimaging, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA.

[5]Joint BioEnergy Institute, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA.

[6]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA.

[7]National Renewable Energy Laboratory, Golden, CO 80403, USA.

[8]Department of Bioengineering, University of California, Berkeley, CA 94720, USA.

[9]Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720, USA.

*These authors contributed equally: Pradeep Kumar Prabhakar, Jose Henrique Pereira.

#e-mail: breeanna@uga.edu.

22

2

Abstract

Rhamnogalacturonan I (RGI) is a structurally complex pectic polysaccharide with a backbone of alternating rhamnose and galacturonic acid residues substituted with arabinan and galactan side chains. Galactan synthase 1 (GalS1), transfers galactose and arabinose to either extend or cap the β-1,4 galactan side chains of RGI, respectively. Here we report the structure of GalS1 from *Populus trichocarpa*, showing a modular protein consisting of an N-terminal domain that represents the founding member of a new family of carbohydrate-binding module, CBMXX (number denoted as 'XX" assigned upon publication), and a C-terminal glycosyltransferase family 92 (GT92) catalytic domain that adopts a GT-A fold. GalS1 exists as a dimer *in vitro*, with stem domains interacting across the chains in a 'handshake' orientation that is essential for maintaining stability and activity. In addition to understanding the enzymatic mechanism of GalS1, we gained insight into the donor and acceptor substrate binding sites using deep evolutionary analysis, molecular simulations, and biochemical studies. Combining all the results, a mechanism for GalS1 catalysis and a new model for pectic galactan side chain addition are proposed.

Introduction

Plants are the pre-eminent builders of complex carbohydrates, essential molecules of life that store and supply energy to nearly all organisms in the biosphere. The plant cell wall is a complex extracellular matrix composed of cellulose, hemicellulose, pectin, proteins, and polyphenolic molecules. Plants are estimated to devote at least 10% of their genomes to constructing their plant cell walls[1]. However, unlike other natural polymers, such as DNA, RNA, and proteins, far

less is known about the synthesis and essential biology of the carbohydrates that constitute plant cell walls. A significant reason why progress has been more challenging is that complex carbohydrate structures are not defined by sequence-based templates but are synthesized through the concerted actions of a diversity of carbohydrate-active enzymes (CAZymes), notably glycosyltransferases (GTs), ~~polysaccharide methyltransferases, and polysaccharide *O*-acetyltransferases,~~ **whose functions and mechanisms of action are slowly ~~being~~ revealed**[2-4]. GALACTAN SYNTHASE (GalS) enzymes that are categorized as inverting GTs from family 92 (GT92; Pfam, PF01697) in the CAZY (http://www.cazy.org/) database[5-7], that catalyze extension or capping of β-1,4-linked galactan side chains of pectic rhamnogalacturonan-I (RG-I) (Figure 1).

Rhamnogalacturonan I (RGI) is a complex pectic polysaccharide found within the primary cell walls of vascular plants[8]. RGI consists of a backbone composed of the repeating disaccharide -2)-α-**L-Rha*p*-(1-4)-α-D-Gal*p*A-(1- (Fig. 1a). The complexity of this polysaccharide is further increased by substitution with lesser amounts of other monosaccharides and non-glycosyl substituents to the backbone Rha*p* and Gal*p*A, respectively**[8, 9]. One such modification~~,~~ is RGI **galactan side chains, which are extended by ~~β-1,4-galactan galactosyltransferases, referred to as GALACTAN SYNTHASE (GalS)~~ enzymes~~, that are categorized as inverting glycosyltransferases from family 92 (GT92; Pfam, PF01697) in the CAZY (http://www.cazy.org/) database~~[5-7] ~~(Fig. 1b)~~. The GalS**1 enzyme from *Arabidopsis thaliana* (AtGalS1) is a bifunctional enzyme that elongates β-1,4 galactan side chains of RGI by adding galactose (Gal) or arabinopyranose (Ara*p*) from UDP-α-D-Gal or UDP-β-L-Ara*p* to extend or terminate the side-chains, respectively**[10]. β-1,4 galactan accounts for a

significant portion of RGI depending on the species[8, 11]. For example, β-1,4 galactan accounts for ~67% of potato RGI and nearly 10% dry weight in tension wood[12]. Galactan chains with a degree of polymerization of up to 300 galactosyl residues are thought to interact with cellulose to generate a gel-like consistency to maintain the size and shape of plant cells, hence imparting mechanical properties that bear stress[13-17], cell elongation[18], and water retention[19, 20].

In this study, we report the crystal structure of GalS1 from *Populus trichocarpa* (Potri.005G258900), which is an ortholog of AtGalS1 (Supplementary Fig. 1), and represents the primary structure of a CAZy GT92 family member. This adds to the only two other structures that have been solved for enzymes involved in plant cell wall biosynthesis, the others being xyloglucan xylosyltransferase 1(XXT1) [21] and xyloglucan fucosyltransferase 1(FUT1)[22, 23]. The **general architecture of GalS1 adopts a C-terminal domain containing a GT-A fold and an N-terminal domain that functions as an ancillary carbohydrate-binding module (CBM) that binds specifically to the backbone of RGI. This CBM is conserved across plant GT92 protein sequences present in Phytozome v12 and represents the founding member of a new CAZY family, CBMXX (number 'XX' assigned upon publication). The presence of a CBM in a glycosyltransferase such as GalS1 is unique and unexpected in a glycosyltransferase. CBMs are more commonly associated with hydrolases or lyases; its presence became more intriguing as we performed biomolecular interaction studies and showed the CBMXX module binds to the backbone of pectic RGI, while the GT92 catalytic domain interacts with β-1,4-galacto-oligosaccharides. Small-angle X-ray scattering (SAXS) experiments demonstrated that GalS1 works as a dimer in solution. Collectively, this study provides insights into the function of both domains of the**

GT92 enzymes and suggests a new model for RGI synthesis where the CBMXX is essential for enzymatic activity and stability that facilitates the ability of GalS1 to target and extend complex acceptor substrates like the repeating disaccharide backbone of RGI. Understanding the architecture and detailed mechanism of GalS1 will enable the utilization of galactan as a source for chemoenzymatic synthesis of tailored polysaccharides[24] for novel applications and the optimization of feedstocks for biomass valorization to chemicals and fuels via the alteration of the hexose to pentose ratio[24, 25].

Results

Expression, purification, and crystal structure of GalS1

GalS1 is classified in the CAZy database as a member of the GT92 family, and until now, no structural information for this family was available. Additionally, GT92 does not share significant amino acid sequence similarities with other GT families. To investigate the structure of GalS1, it was expressed as an sfGFP fusion protein (Supplementary Fig. 2) in HEK 293S GnT1- cells as a soluble secreted fusion protein (122 mg/L estimated by using GFP fluorescence) and purified[26] using affinity and size exclusion chromatography prior to crystallization (Supplementary Fig. 3, 4 & 5). ~~-~~A truncated form of PtGalS1 was generated as a fusion protein containing an NH_2-terminal signal sequence, an 8xHis tag, an AviTag, "superfolder" GFP, the TEV protease recognition site, and amino acid residues 73-495 of PtGalS1 (Supplementary Fig.~~ure~~ 2). We solved two structures, apo form GalS1 diffracting to 2.37-Å resolution and Mn^{2+} bound GalS1 diffracting to 2.56-Å resolution (Supplementary Fig. 6~~3~~). Both structures lack ~~24~~ residues 73-96 (residues 97-495 were observed) in the

electron density maps due to the highly flexible nature of the stem region. The crystal lattice contained four and two copies of GalS1 in the asymmetric unit in the Apo-state and Mn-bound GalS1, respectively. In addition to the polypeptide chain, the GalS1 structure showed 7 glycosylation sites. Each GalS1 monomer contained a stem region (residues 97 to 107) and two globular domains: a CBMXX (residues 108 to 221) connected by a linker region (residues 222-228) to a GT-A fold glycosyltransferase domain (residues 229-495) (Fig. 2a; Supplementary Fig. 2). The core GT-A domain consists of seven core β-sheets (β3, β2, β1, β4, β5, β6, β7), with β5 and β7 in an antiparallel orientation, surrounded by a helix that includes the donor and acceptor binding sites. The GT core of GalS1 displayed some distant similarity (RSMD ≥ 4.2 over ≥ 138 residues) to insect and mammalian β1,4-galactosyltransferases (β4GalTs)[27, 28] that transfer galactose from UDP-Gal to xylose or N-acetylglucosamine (GlcNAc), respectively, from CAZy family GT7 (Supplementary Fig. 4). However, the GalS1 structure is dissimilar to known β4GalTs; these changes possibly account for GalS1 activity as both extending and capping β-1,4 galactan side-chains (Supplementary Fig. 47). We also tried to obtain UDP-, UDP-Gal donor-, and acceptor-bound structures but were unsuccessful.

The oligomeric state of GalS1

SEC-MALS analysis of GalS1 indicated that it exists as a dimer *in-vitro* (the calculated molecular weight of GalS1 is 103.9kDa, and the theoretical molecular weight is 97.6 kDa). The asymmetric unit in the apo state showed 4 molecules of GalS1 and 2 possibilities of dimer formation. To identify the correct monomer-monomer interactions (Fig. 2b), we examined the solution state of GalS1 by performing experiments where small angle X-ray

scattering (SAXS) is coupled to size exclusion chromatography (SEC) and multi-angle light scattering (MALS) and quasi-elastic light scattering (QELS) detection. SEC-SAXS-MALS experiments provide accurate measurement of molecular weight and provide information on particle shape. We observed a single peak eluting from the gel filtration column that corresponded to the GalS1 homodimer, as judged by molecular weight determined from SAXS and MALS (MW_{SAXS} = 110kDa, MW_{MALS} = 119kDa). Two conformers were built based on two possible interfaces visualized in the crystal structure to determine dimer arrangement in solution: parallel (A:B) or antiparallel (A:C)(Fig. 2c). The antiparallel arrangement with the N-terminal stem region interacting across the chains matched the SAXS curve well, whereas the alternative dimer showed a poor match. (χ^2_{dimer} =7.8 and $\chi^2_{alternative\ dimer}$ =254.3, Fig. 2c). The residual discrepancy between the atomistic model and the SAXS data was due to the absence of glycans in our model and the flexibility of the disorder N-terminal region (Fig. 2c). Additionally, we reconstructed the SAXS envelope to further confirm the overall arrangement of the GalS1 homodimer in an antiparallel (A:C) orientation (Fig. 2c).

The importance of the stem region of GalS1

We identified several potential interactions between two GalS1 monomers forming the homodimer (Supplementary Fig. 8~~5~~a, & 8~~5~~b), including interactions between the N-terminal stem region of one monomer with the other. Comparative analysis of stem regions of GT92 proteins across different plant species indicated that conservation beyond residue Asp96 increases, indicating a conserved role in stability, activity, or both (Supplementary Fig. 8f~~2~~). To further investigate the role of the stem domain in dimer formation and

activity, we generated an ΔSTEM-construct lacking the stem region (Supplementary Fig. 2), and evaluated the activity of the truncated variant using galactotetraose as an acceptor. Despite the presence of the GT92 catalytic domain, the GalS1-ΔSTEM variant was inactive (Supplementary Fig. ~~5~~8c & 8~~5~~d). Comparison of thermal melting temperatures showed a decrease from 59.1°C for WT to 57.4°C in the ΔSTEM variant and indicated that the protein was correctly folded but slightly less stable (Supp Table 1). SEC-MALS analysis of the ΔSTEM variant suggested that a portion of the protein was present as a higher molecular weight aggregate (nearly 8% of the total), in addition to the expected dimer (calculated MW is 84.5, theoretical MW of the dimer is 90.1 kDa) in solution (Supplementary Fig. 8~~5~~e). These data suggested the stem region plays an essential role in the structural and functional stability of GalS1, as its presence prevents higher order aggregation of GalS1 *in vitro*, but is not entirely responsible for dimerization. Construction and analysis of additional truncation variants may shed light ~~into~~ on ~~its~~ their role in dimerization.

GalS1 contains an N-terminal carbohydrate-binding module that is the founding member of a new CBM family (CBMXX)

Inspection of the GalS1 structure revealed an additional domain at the N-terminus of the protein (amino acids 108-221), that adopted a β-sandwich fold reminiscent of the CBM-60 present in a xylanase from *Camponotus japonicus* (2XFD[27]; RMSD for Cα of 4.4 over 64 residues; Fig. 3a) and a CBM-61 from an endo-β-1,4-galactanase from *Thermotoga maritima* (2XOM[28]; RMSD for Cα of 7.6 over 88 residues; Fig. 3b). To explore the function of this putative domain, we generated an GalS1-CBMXX construct (Supplementary Fig. 2),

and evaluated its ability to bind various cell wall oligo- and polysaccharides using MicroScale Thermophoresis (MST). We showed that the GalS1-CBMXX specifically binds unbranched pectic RGI[29, 30] isolated from non-adherent *Arabidopsis thaliana* mucilage. In contrast, GalS1-CBM-XX did not interact with galactotetraose, polygalacturonic acid, or xylohexaose based on a cutoff of a signal-to-noise ratio below five, minimally required to confirm binding. Most polysaccharide substrate-binding happens through stacking interactions with aromatic residues on the CBM surface. Therefore, we mutated various exposed tyrosine and tryptophan residues on the surface of the CBMXX domain (Fig. 3c). Additionally, basic residues such as lysine have previously been shown to act as functional residues in pectin-binding CBMs such as CBM77[31] and inspection of the GalS1 structure revealed that several were present on the surface exposed region of CBMXX and were also mutated (Fig. 3c). Recombinant CBMXX and the aforementioned mutant variants were expressed in HEK293 cells and purified using Ni-NTA. We studied the effects of mutating these residues on RGI binding. MST analysis of mutant variants using RGI as a substrate showed that K133A, W142A, Y199A, K206A, K209A displayed increase in the K_D from 3- to 6-fold, whereas K144A, W166A, and Y207A variants showed an increase in the K_D from 10 to 13-fold (Fig. 3d), indicating the latter play a more predominant role in RGI interaction. The CBMXX of GalS1 does not share any sequence homology with any other CBMs in the CAZY database[7] (confirmed by personal communication with Dr. Nicholas Terrapon, head of the CAZy database) and will be assigned as a new CBM-Family (see http://www.cazy.org/CBMXX.html for an actively updated list of sequences and source organisms).

We have established that the stem region is essential for galactan synthase activity. To investigate whether the GT-A core domain is still catalytically active in the absence of the CBM, we generated an GalS1-ΔCBM variant (Supplementary Fig. 2) that lacks the entire CBM domain and the stem region. The GalS1-ΔCBM variant of the GalS1 was successfully expressed as a soluble secreted fusion protein (75 mg/L); however, it lacked detectable galactan synthase activity (Fig. 3~~4~~d), suggesting that both the CBM and the stem domain play a crucial role in enzyme stability/folding and catalysis.

Identifying evolutionarily constrained residues in the GT92 family

Recently, a minimal structural unit for GT-A fold enzymes has been defined based on deep mining of large sequence datasets, revealing twenty residues shared throughout the common GT core[32]. Unfortunately, GT92 family proteins were not included in the study due to a lack of structural information **at the time of publication. To identify core conserved residue positions within the GT92 family, we generated an alignment of representative GT92 sequences with other GT-A fold sequences using a profile-based approach and the GalS1 structure as a template. For this, we aligned the GalS1 structure with other GT-A fold structures and used this structural alignment as a basis to then align a GT92 consensus to the GT-A profile alignment generated in a previous study**[31]. ~~The incorporation of~~Incorporating GT92 sequences into this alignment~~,~~ provided a comparative basis for mapping GT-A shared features and **residues uniquely conserved in the GT92 family (Fig. 4a & 4b**Supplementary Fig. 9a)~~)~~. Initially, based on th**e profile alignment, several GT-A fold conserved motifs were mapped: the DXD motif that is involved in coordinating the metal**

ion and the donor sugar in metal-dependent GT-A fold enzymes (D331 and D333); the G-loop involved in donor binding (R397-K400), and the conserved xED motif harboring the catalytic base (G412-H414) with H414 as the putative catalytic base (see below). Based on the alignments, H435 is predicted to function as the metal coordinating histidine at the C-terminal tail (C-His). In addition, the hydrophobic core residues that define GT-A fold enzymes are also present in GT92[32]. These include Y233, L234, Y235, M249, M253, F266, V267, F268, F328**, and I403 (**~~Fig. 4e~~Supplementary Fig. 9b). Moreover**, to identify GT92 specific residue positions, we performed a query-centric Bayesian partitioning with pattern selection (BPPS)**[33] analysis on a set of 24816 sequences that includes diverse GT-A fold sequence sets [32] and representative GT92 sequences using the GT92 consensus sequence as the query. This resulted in a foreground cluster of 153 GT92 sequences defined by multiple residue positions uniquely conserved within these sequences, suggesting family-specific functions. These residues are highlighted in **Fig. 4a along with the GT-A shared motifs. The most distinct GT92 specific feature was H414, which is invariant at this position across all GT92 sequences and is distinct from other GT-A fold enzymes, which largely conserve an Asp or a Glu that acts as a catalytic base. We also identified K400 as one of the most uniquely conserved feature of GT92. This residue is part of the G-loop, which uniquely conserves a number of charged residues in contrast to smaller amino acids with shorter side chains like Gly, Ala, or Ser in other GT-A fold enzymes. Other GT92-specific features include cysteines (C236 and C316) that form a disulfide bond and other charged residues (D315, E334) within the GT-A domain (**Supplementary **Fig. 9a~~4b~~).

Docking and Molecular Dynamics simulations reveal putative binary and ternary substrate-bound GalS1 complexes

The pursuit of crystallizing a ligand-bound structure was unsuccessful; however, we obtained an Mn^{2+} ion-bound GalS1 structure that pointed to the binding pocket at the active site. We used this structure as a starting point for docking and molecular dynamics simulations studies to identify substrate binding modes, and evaluate enzyme-substrate bound structures. The Mn-bound GalS1 structure was equilibrated under fully solvated conditions since it was crystallized without the presence of a substrate. MD simulations of the monomeric Mn bound GalS1 were performed to explore the flexibility of the various structural domains of the GT-A fold (Fig. 5a) and provide an equilibrated receptor structure for docking the donor substrate. Blind docking studies of the donor substrate revealed that most bound poses were concentrated around the Mn-binding site. Targeted binding studies suggested that the Mn-binding site could accommodate the donor molecule with favorable binding energies and showed configurational and geometric feasibility for hydrolysis based on its proximity to the putative catalytic base H414. MD simulations of the UDP-Gal-Mn-GalS1 complex revealed that the substrate remains bound throughout the 100-ns simulations (Fig. 5b), which illustrated a putative binding pose for the donor molecule at the active site. Distances between the donor sugar C1 and the putative catalytic base N during the MD simulations were observed to be consistent with hydrolysis of the sugar molecule even in the absence of the acceptor molecule and corroborated the experimental observation of the same. MD simulations of the donor-bound state also provided equilibrated structures of the binary complex for initiating docking studies of the acceptor-bound ternary complex (Fig. 5c). Docking

~~simulations with galactotetraose suggest that acceptor substrate binding is likely coordinated via key aromatic residues on the CBMXX domain of GalS1.~~

~~Furthermore, various docked conformations were observed to satisfy two critical requirements for the GalS1 reaction mechanism (*i*) the orientation of the non-reducing end of the substrate into the active site and (*ii*) proximity to the putative base. Fig. 5d illustrates a putative binding pose for the ternary complex. Although this pose in itself does not represent a catalytically competent configuration, the ability of the active site to stabilize the ternary complex over 10s of nanoseconds in the MD simulations presents promise for conducive configurations of the complex that may undergo catalysis according to the proposed reaction mechanism (Fig. 6a).~~

Biochemical characterization of active site residues in GalS1

The GT sequence comparisons combined with docking and molecular dynamics simulations analysis provided us with significant insight into the putative residues involved in substrate binding and catalysis in the GT92 domain of the enzyme. To understand the specific roles of these residues further, we performed multiple independent mutational analyses to study the effects of these non-conservative mutations on the activity of the enzyme. In the absence of a suitable acceptor substrate, most Leloir glycosyltransferases can hydrolyze suitable nucleotide sugar-donor substrates resulting in the release of a nucleotide product, such as UDP. This can be exploited to investigate donor specificity of glycosyltransferases without having any knowledge of acceptors. We used a UDP-Glo-coupled hydrolysis assay[34] to analyze the sugar-nucleotide donor specificities of **GalS1-WT using several common UDP-containing glycosyl donors present in plants: UDP-Gal, UDP-**

Ara*p*, UDP-Xyl, UDP-GlcNAc, UDP-GalNAc, UDP-Glc, UDP-GalA, and UDP-GlcA. GalS1-WT displayed specific hydrolysis activity for both UDP-Gal and UDP-Ara*p* (Supplementary Fig. 106), consistent with previous reports for *Arabidopsis thaliana* GALS1 showing utilization of UDP-Gal or UDP-Ara*p* to either extend or terminate galactan chains, respectively[10].

Key donor binding residues of the core GT-A domain were identified based on a statistical analysis of evolutionary constraints acting on primary sequences and docking simulations. Since Drosophila β4GalT7 D211N complex with manganese, UDP-Gal, and xylobiose is available in the database (PDB id 4M4K; Supplementary Fig. 4) showing key donor and acceptor binding residues[35], we merged apo PtGalS1 to Dmβ4GalT7 to point the key residues of GalS1 involved in the donor and the acceptor binding (Fig. 4b & 4c). We mutated several of these residues hypothesized to be involved in nucleotide sugar donor binding to alanine. GFP-fused mutated variants; G242A, D331A, D333A, E334A, Q309A, K400A, H414A, H435A, and H437A, were expressed and purified in HEK 293S WT cells. First, we quantified glycosyltransferase activity by UDP-Glo-assay, using UDP-Gal or UDP-Ara*p* as a donor and galactotetraose as an acceptor. The results indicate that mutating residues of the DxD motif (D331A and D333A), as well as E334A, Q309A, K400A, and H414A, reduces galactosyltransferase (GalT) and arabinopyranosyltransferase (Ara*p*T) activity to a minimum compared to the wild-type enzyme, showing that these residues are essential for catalysis/binding (Fig. 4d). In contrast, mutant variants G242A, H435A, and H437A displayed only slightly reduced GalT and Ara*p*T activity, indicating that these residues are not directly involved in binding or catalysis (Fig. 4d). A key GT92-specific feature is a

conserved disulfide bridge between C236 and C316. We generated a C236S single mutant and a C236S:C316S double mutant to evaluate the role of this disulfide bond in GalS1. Both mutations resulted in negligible protein expression (Supplementary Table S2), making purification difficult, and the resulting enzymes showed no detectable GalT or Ara*p*T activity (Fig. 4d). We postulate that this disulfide formation is essential for protein folding or stability post expression. Comparison of chain A of GalS1 with chain A of β4GalT7 (bound with Xylobiose; PDB id: 4m4k) led to the identification of three residues, W166, R396, and D398 as potential acceptor binding residues. Mutating W166A and R396A did not significantly perturb GalT or Ara*p*T activity however, the D398A variant showed loss of both GalT and ArapT activity in the presence of acceptor (Fig. 4d and 4e). The hydrolytic activity or the ability of the enzyme to transfer the sugar to water in the absence of acceptor substrate, of most variants was comparable to WT GalS1 except in W166A, where hydrolytic activity is increased by nearly 2.5-fold. To better understand the contribution of all the residues studied above in binding the nucleotide sugar donor, we estimated the equilibrium dissociation constant (K_D) of UDP-Gal in WT and mutant variants using microscale thermophoresis. The K_D of UDP-Gal in D331A, K400A, and H435A is in a similar range to WT GalS1. E334A, Q309A, and H414A hampered the binding of UDP-Gal, whereas all other mutants showed improvement in the binding of UDP-Gal (Table 1). Taken together, *in vitro* GalT or Ara*p*T assays and insights into enzyme-donor substrate interactions support that E334, Q309, and H414 directly interact with UDP-Gal/UDP-Ara*p* in GalS1 and are required for glycosyltransferase activity. The reduced activity of H435A and D333A (part of the DXD motif) is probably due to the

inability to interact with Mn^{2+}. As hypothesized, H414 acts as a catalytic base; as no other residues closer to the donor seem possible, and mutating this residue to alanine leads to a complete loss of activity. G242 and H437 are part of variable loops, allowing flexibility at the active site, explaining a partial decrease in the catalytic activity due to mutation. W166, R396, and D398 are involved in acceptor binding (Fig. 4c). We propose that this occurs through stacking interactions between the acceptor and the aromatic W166, and hydrogen bonding interactions with the planar polar side-chains of R396 and D398~~W166, R396, and D398 are essential for acceptor binding, probably by a series of aromatic interactions~~**. This is evident from these mutations that improved the K_D of UDP-Gal and yet led to a drop-in glycosyltransferase activity, particularly apparent in the D398A mutant variant. W166 is even more fascinating as it protrudes into the active site from the CBMXX domain, and the W166A variant has a 2.5-fold increase in the rate of NDP-sugar hydrolysis. Donor affinity is improved in the W166A variant by two fold but shows decreased GalT activity, suggesting W166 is directly involved in acceptor substrate binding.** This was further shown by a decrease in the dissociation constant (K_D) of the W166A mutant (nearly seven-fold), ~~showing~~confirming its role in the acceptor binding (Supplementary Fig. 11). WT and its mutant variants were tested **for galactosyltransferase activity by PACE or polysaccharide analysis using carbohydrate gel electrophoresis. The result broadly agrees with the Glo-based assays, re-confirming the role of the selected residues in the galactosyltransferase activity (Supplementary Fig. ~~7~~12).**

Docking and Molecular Dynamics simulations reveal putative binary and ternary substrate-bound GalS1 complexes

The pursuit of crystallizing a ligand-bound structure was unsuccessful; however, we obtained an Mn^{2+} ion-bound GalS1 structure that pointed to the binding pocket at the active site. We used this structure as a starting point for docking and molecular dynamics simulations studies to identify substrate binding modes, and evaluate enzyme-substrate bound structures. The Mn-bound GalS1 structure was equilibrated under fully solvated conditions since it was crystallized without the presence of a substrate. MD simulations of the monomeric Mn bound GalS1 were performed to explore the flexibility of the various structural domains of the GT-A fold (Fig. 5a) and provide an equilibrated receptor structure for docking the donor substrate. Blind docking studies of the donor substrate revealed that most bound poses were concentrated around the Mn-binding site. Targeted binding studies suggested that the Mn-binding site could accommodate the donor molecule with favorable binding energies and showed configurational and geometric feasibility for hydrolysis based on its proximity to the putative catalytic base H414. MD simulations of the UDP-Gal-Mn-GalS1 complex revealed that the substrate remains bound throughout the 100-ns simulations (Fig. 5b), which illustrated a putative binding pose for the donor molecule at the active site. Distances between the donor sugar C1 and the putative catalytic base N during the MD simulations were observed to be consistent with the hydrolysis of the sugar molecule even in the absence of the acceptor molecule and corroborated the experimental observation of the same. MD simulations of the donor-bound state also provided equilibrated structures of the binary complex for initiating docking studies of the acceptor-bound ternary complex (Fig. 5c). Docking simulations with galactotetraose suggest that acceptor substrate binding is likely coordinated via key aromatic residues on the CBMXX domain of GalS1.

Furthermore, various docked conformations were observed to satisfy two critical requirements for the GalS1 reaction mechanism (*i*) the orientation of the non-reducing end of the substrate into the active site and (*ii*) proximity to the putative base. Fig. 5d illustrates a putative binding pose for the ternary complex. Although this pose in itself does not represent a catalytically competent configuration, the ability of the active site to stabilize the ternary complex over 10s of nanoseconds in the MD simulations presents promise for conducive configurations of the complex that may undergo catalysis according to the proposed reaction mechanism (Fig. 6a).

Discussion

The presence of a carbohydrate-rich cell wall is a ubiquitous feature of all plants. While we are beginning to understand the composition and diversity of the polysaccharide components in these walls, little is known about the molecular players involved in their synthesis. Recent studies on galactan interactions with cellulose in tension wood, and its possible implications in stress-bearing and imparting flexibility and support to plant tissues, highlight its complex role in the plant cell wall. Our ability to develop more refined synthetic biology approaches to design plant cell walls with enhanced properties for valorization of the fixed carbon locked within them requires detailed understanding of their biosynthetic processes at the molecular level. Re-engineering a biocatalyst such as GalS1 requires understanding its active site, catalytic mechanism, and interactions with other functional proteins or protein domains. Our structure of a plant β-1,4 galactan synthase revealed that GalS1 is a modular protein with an ancillary carbohydrate-binding module (CBMXX) at its N-terminus that binds specifically to the backbone of RGI. Further, we showed that the stem region plays a structural role in homodimer formation,

interacting across GalS1 monomers in a 'handshake' pose, and is essential for both GT activity and protein stability.

GalS1 belongs to the glycosyltransferase A or GT-A fold of the broader classification of glycosyltransferases. Nucleotide-sugar binding site residues within the GT-A fold are highly conserved among family members[7, 32]. Although we were unsuccessful in determining a donor or acceptor-bound X-ray structure, we were able to identify crucial residues using sequence conservation information by mining over half a million GT-A fold sequences and comparing them to those in the GT92 family. The majority of GT-A enzymes that catalyze the ~~s that~~ formation of glycosidic linkages between a donor and an acceptor substrate through a single step, inverting catalytic mechanism ~~are inverting enzymes~~ utilize Asp or Glu within a conserved, protein associated xED motif as the catalytic base[36]. In contrast, our bioinformatics and docking data led to the hypothesis, and biochemical analyses confirmed, that H414 functions as the catalytic base in GalS1. We also determined E334 is essential for UDP-Gal binding, extending the DxD motif to a DxDE motif in GT92. Docking with MD simulations provides a powerful alternative to study active site residues, their contribution to binding or catalysis, and evaluate the flexibility of loops to accommodate substrates at the active site[37-40]. Our molecular simulations ~~confirms~~support the proposed binding site and elucidate the critical active site residues in stabilizing the donor substrate (Figure 5b). They also provide insight into acceptor binding at the active site, which is shown to be mediated via a series of hydrophobic interactions (with Y241, Y307, V413, and F310 in Figure 5d), common among many glycosyltransferases[41-43]. ~~evaluate the flexibility of loops to accommodate substrates at the active site[39-42]. Our computational studies confirmed the critical interaction of the donor substrate~~

~~identified above and provided further insight into the acceptor binding site, which binds through a series of hydrophobic interactions, imparting flexibility at the active site to allow binding of galactan chain in GalS1; common among many glycosyltransferases[43-45].~~

A key finding was that GalS1 is a modular enzyme containing a CBM, a domain more commonly found in enzymes involved in carbohydrate deconstruction (GH)[44], and rarely **associated with GTs. The GT domain of GalS1 functions to extend galactan side-chains of RGI. In contrast, we demonstrated that the CBMXX binds to the RGI backbone, data that resulted in its classification as the founding member of a new family in the CAZy database. Identification and characterization of this new module led us to propose a new model in which the CBMXX functions to bring the GalS1 enzyme in proximity to the RGI backbone to enable chain elongation (Fig 6b), potentially functioning to target regions of the polymer that are sparsely substituted. In Arabidopsis's *GalS1/GalS2/GalS3 triple mutants*, the RGI backbone still has galactose substitutions even though elongated galactan chains are absent[6].** Taken together, all **available genetic and biochemical evidence supports that GalS1 catalyzes galactan chain extension but is not involved in attaching the initial galactose residue(s) to the RGI backbone. Thus, the GalT that adds the initial galactose residues to the RGI backbone is still unknown. It remains to be shown if this is a common principle of the synthesis of complex polysaccharides like RGI polysaccharide or a unique feature of galactan side-chain elongation.**

The stem region of a GT is generally defined as the stretch of amino acids after the transmembrane domain that can be truncated without changing the enzyme's activity. Biologically, stem regions are proposed to function as flexible tethers that position in the

catalytic domain away from the membrane[45]. There are also many examples that have investigating the various roles of the stem regions of GTs in flexibility, orientation to the substrate[46, 47], site of interaction to other proteins or itself, acting as chaperon[48], localization, and stability[49]. The stem region and the catalytic domain are not clearly differentiated in glycosyltransferases**; however, for this study, we have characterized the stem region as the residues not part of the globular domains and showed it plays a critical role in dimer structure. Furthermore, thermo-stability and biochemical analyses of stem deletion mutants showed that this region is essential for stability and biochemical activity of GalS1.**

GalS1 is a ~~metal~~ metal-**dependent, inverting glycosyltransferase (Fig. 1b). Combining structural information with computationally guided mutational analyses and molecular dynamics simulations, our data shows that GalS1** ~~it~~ **utilizes an S_N2 single-displacement reaction mechanism (Fig. 6a), similar to GalTs from CAZy family *GT7*** [35, 50, 51]. In other GalTs[50], it has been demonstrated that acceptor binding likely involves structural rearrangements of residues after donor binding to create an active site conducive for the reaction. ~~The process generally involves a loop movement after donor binding to form an acceptor binding site allowing completion of the reaction, followed by opening or relaxing of the loop to accept new molecules for the next reaction. In GalS1,~~ His414 of GalS1 aligns (Figure 4b and 4c) with the catalytic base in *Drosophila* beta,14galactosyltransferase 7 (PDB id: 4M4K[35]; also, mutant H414A has negligible activity indicating that in GalS1, His414 -acts as the catalytic base to deprotonate the Gal C4 nucleophilic hydroxyl group of the acceptor and the carboxylic group that stabilizes the divalent cation required to complete the GalT reaction, respectively (Fig. 6a). Investigation

of the Mn^{2+} bound GalS1structure shows that the conserved His435 is directly involved in coordination of Mn^{2+}. This structural data, combined with alignments with GalT structures from other GT families further suggests that His435 plays a direct role in coordination of Mn^{2+} with oxygen of the β-phosphate in the nucleotide sugar donor (Fig 4b).

Unlike GT7 GalTs that catalyze the addition of a single sugar, GalS1 catalyzes the extension of β-1,4-galactan side-chains composed of 100s of monosaccharides *in vivo*. To efficiently process several hundreds of reactions in the Golgi compartment, we propose that GalS1 utilizes the N-terminal carbohydrate-binding module to both target and anchor itself to the RG-I backbone (Fig. 6b) through a combination of hydrophobic and ionic interactions (Fig. 3c). This is somewhat unusual in GTs but very common in glycoside hydrolases. Plant starch synthase III has modular structure similar to that of GalS1, where the presence of an N-terminal starch-binding domain from CBM20 increases progressivity of the enzyme [52-55]. Similar to the CBMXX from GalS1, the CBM20 in starch-binding domain-containing protein 1 or STBD1 associated with glycogen metabolism and autophagy pla**ys an essential role in stability and facilitates interaction with glycogen-associated proteins**[56]. Taken together, these data suggest that the presence of CBMs in GT may play an enabling role for the efficient synthesis of long polysaccharides.

Very recently, new machine learning approaches, including RoseTTAFold[57] and Alphafold[58], have been developed for the prediction of protein secondary structures from their amino acid sequences and **have been widely released.** We compared the AlphaFold predicted structures of

Arabidopsis and Populus GalS1 with the core domains of the empirically determined Populus trichocarpa GalS1 structure and found that, in this case, they are very similar to the with root-mean-square deviation (using Cα) is 0.610 Å and 0.593 Å respectively, the slight difference in the RMSD between the two is majorly due to the orientation of the stem domain (Supplementary Fig. 13). ~~We compared the AlphaFold structures for Arabidopsis GalS1 with the core domains of the empirically determined *Populus trichocarpa* GalS1 structure, and found that in this case they are very similar to the (RSMD is only 0.641Å).~~ **The experimental insight into the 3D structure of GalS1 will accelerate the understanding of GalS1's role in tension wood and potential approaches to enzyme engineering and gene replacement.** In the future, it may be possible to use the knoledge gained from structural and functional analysis of ~~Comparing various~~diverse **bacterial and eukaryotic galactosyltransferases** ~~and applying this knowledge~~ to develop targeted engineering strategies **to** ~~create~~ modify enzymes like **GalS1** to generate variants with altered donor, acceptor or regioselectivity to enzymatically generate new saccharide structures~~to transfer other monosaccharides to an acceptor in α-1,3-, α-1,4-, or β-1,3- linkages could have many applications~~**.** However, the viability of this enzyme engineering approach will involve significant concerted efforts in understanding the detailed mechanisms of inverting and retaining glycosyl transferases ~~with~~together with elucidating the molecular basis of ~~diverse activities~~donor and acceptor substrate selectivity. ~~in terms of inverting and retaining mechanisms that act on a diverse set of substrates.~~ In particular, galactose binding to lectins has been implicated in tumor metastasis in mammals[59, 60], and unnatural substrates can pave the way to developing newer inhibitors. Also, in the future, we plan to design chimeric transferases **or**

hydrolases with CBMXX domains to modify the activity of ~~pectin~~ pectin-synthesizing and degrading enzymes.

Methods

Cloning, protein expression, and site-directed mutagenesis.

The NΔ72GalS1 coding sequence was amplified from a cDNA template prepared from terminal buds of *Populus trichocarpa* WT primers (Supplementary Table 1) and cloned into mammalian expression vector pGEn2-DEST according to our standard protocols[26, 61]; henceforward, it will be called pGEN2-DEST-GalS1WT or GalS1 WT. The resulting fusion proteins consisted of an N-terminal NH2-signal sequence, 8xHis tag (for purification), AviTag recognition site, superfolder GFP (sfGFP, for quantification), and the seven amino acid TEV protease recognition site, followed by the truncated coding region of GalS1. Mutated variants of GalS1 were generated by site-directed mutagenesis using the Q5 Site-Directed Mutagenesis Kit (New England Biolabs, USA) according to the manufacturer's instructions using pGEN2-DEST-GalS1WT as a template using primers listed in (Supplementary Table 2). Primers used to generate the GalS1-ΔSTEM- (lacks residues 73-113) and GalS1-ΔCBM truncation variant that lack the stem and the stem and CBM regions (Supplementary Fig. 2), respectively, are listed in Supp Table 1. Primers used to generate the construct for expression of Carbohydrate-Binding Module XX (CBM XX; 73-235 residues; Supplementary Fig. 2), are listed in Supplementary Table 1. Mutated variants of CBM XX were generated by site-directed mutagenesis using the Q5 Site-Directed Mutagenesis Kit (New England Biolabs, USA) according to the manufacturer's instructions using pGEN2-

DEST-CBM XX as a template with primers listed in Supplementary Table 3. All constructs were confirmed by DNA sequencing (Eurofins, USA). For transient expression, plasmid DNA was isolated using the PureLink™ HiPure Expi Plasmid Gigaprep Kit or Maxiprep Kit (Thermo Fisher Scientific) as suggested by the manufacturer. Plasmids were transfected into HEK cells (FreeStyle™ 293-F cell line, Life Technologies; HEK293S GnTI- cells, catalog number CRL-3022, ATCC[62]) as described previously[61]. Selenomethionine labeling of WT GalS1 was done by transfecting HEK293S GnTI- cells with pGEN2-DEST-GalS1WT in methionine-starved custom media for 6 hrs and then supplementing it with 60 mg/L of selenomethionine. Soluble secreted fusion proteins were harvested from the media on the sixth day. Schematics of domain organization of the full-length protein and constructs used in this study are in Supplementary Fig 2.

Purification, size-exclusion chromatography coupled with multi-angle light scattering (SEC-MALS), and thermostability of recombinant GalS1 fusion proteins

Extracellular media harvested from the culture was processed as described previously[26]. All purifications were carried out using HisPrep FF 16/10 or HisTrap FF 5 ml columns (Cytiva, USA) on an ÄKTA Go or ÄKTA Pure 25L (Cytiva, USA) protein purification system[26] . Proteins were concentrated to 5 mg/ml using Amicon Ultra 15 ml centrifugal filter devices (10 kDa MWCO, Millipore, USA) and stored at 4℃. Proteins were further purified by gel filtration using a Hi Load 16/600 Superdex 200 pg column (Cytiva, USA) in 50 mM HEPES, 400 mM NaCl at pH 7.5. Fractions were combined and dialyzed overnight in 50 mM HEPES containing 100 mM NaCl at pH 7.5, concentrated to 2 mg/ml, aliquoted (200 µl), and flash-frozen in liquid nitrogen before storing at -80℃.

For crystallization, purified GalS1-WT (94 mg) expressed in HEK293S GnTI- cells was treated with 5 mg each recombinant His-tagged GFP-TEV protease and His-tagged EndoF1[26, 61] at 4°C for 24 hrs. Tag-free protein was further purified by a second round of immobilized metal affinity chromatography (IMAC) to remove the cleaved N-terminal sfGFP tag, His-tagged GFP-TEV, and His-tagged GFP-EndoF1, concentrated to 5 mg/ml and loaded onto a **Superdex 75 Increase 10/300 GL (Cytiva, USA) column. The significant fraction was collected and dialyzed overnight into 50 mM HEPES, 100 mM NaCl, pH 7.5, and concentrated to 15 mg/ml. SEC-MALS was carried out in 50 mM HEPES, 250 mM NaCl, pH 7.5 on a Superdex 200 10/30 GL (Cytiva, USA) column using an Agilent HPLC system coupled to an Optilab T rEX Refractive Index Detector and a Mini Dawn Treos Detector (Wyatt Technology, USA). 20 µl of protein (2 mg/ml) was injected using an autosampler. Analysis was done using ASTRA 6 HPLC Software (Wyatt Technology, USA). Protein thermal shift assays were carried out using 5 µM of protein and 200X SYPRO™ Orange (Thermo Fisher Scientific) in 50 mM HEPES, 100 mM NaCl, pH 7.5 in a total volume of 50 µl in Hard-Shell® 96-well WHT/CLR (Bio-Rad, USA) plates using a CFX96™ Real-Time System (Bio-Rad, USA). Fluorescence reads using the ""FRET"" channel to measure SYPRO Orange fluorescence were taken at each 30-sec hold as temperature was increased from 25°C to 100°C. The data was analyzed using the JTSA online server (Bond, PS. JTSA., 2017, http://paulsbond.co.uk/jtsa).**

Crystallization, X-ray data collection, and structure determination

For crystallization trials, GalS1 (12 mg/ml) was screened using the following crystallization screens: Berkeley Screen[63], Crystal Screen, SaltRx, PEG/Ion, Index, and PEGRx (Hampton

Research), and MCSG-1 (Anatrace). Crystals of GalS1 were found in 0.1 M Sodium citrate tribasic dihydrate pH 5.0 and 10% w/v Polyethylene glycol 6,000. They were obtained after two days by the sitting-drop vapor-diffusion method with the drops consisting of a mixture of 0.2 µl of protein solution and 0.2 µl of reservoir solution. Crystals of **GalS1 were placed in a reservoir solution containing 20% (v/v) glycerol, then flash-cooled in liquid nitrogen. The X-ray datasets for GalS1 were collected at the Berkeley Center for Structural Biology beamline 8.2.2 at the Advanced Light Source at Lawrence Berkeley National Laboratory (LBNL). The diffraction data were recorded using an ADSC-Q315r detector and processed using the program Xia2**[64].

The GalS1 crystal structure was determined using selenomethionine (Se-Met)-labeled protein by the single-wavelength anomalous dispersion (SAD) method[65] with *phenix.autosol*[66]and *phenix.autobuild*[67] programs within the Phenix suite[68, 69]. The atomic positions obtained from the initial SAD data set were used as a search model for molecular replacement against native GalS1 data and to initiate crystallographic refinement and model rebuilding. Structure refinement was performed using the *phenix.refine* program[69]. Translation-libration-screw (TLS) refinement was used, with each protein chain assigned to a separate TLS group. Manual rebuilding using COOT[70] and the addition of water molecules allowed the construction of the final model. The final models of GalS1 and GalS1-Mn^{2+} have a R_{factor} of 0.197 / R_{free} **of 0.247 and R_{factor} of 0.235 / R_{free} of 0.267, respectively. Root-mean-square deviation differences from ideal geometries for bond lengths, angles, and dihedrals were calculated with Phenix. The stereochemical quality of the final model of GalS1 was assessed by the program MOLPROBITY.**

Summary of crystal parameters, data collection, and refinement statistics can be found in Supplementary Table S4.

Small-angle X-ray scattering (SAXS)

SAXS was performed at the SIBYLS beamline at the Advanced Light Source[71, 72]. For SAXS coupled with a multi-angle light scattering in line with size-exclusion chromatography (SEC-SAXS-MALS) experiments, 60 µL containing 10 mg/ml GalS1 in 25 mM Hepes pH 7.5, and 100 mM NaCl was used during the experiments. SEC-SAXS-MALS data were collected at the ALS beamline 12.3.1 LBNL Berkeley**, California. The X-ray wavelength was set at λ=1.127 Å, and the sample-to-detector distance was 2100 mm, resulting in scattering vectors, q, ranging from 0.01 Å-1 to 0.4 Å-1. The scattering vector is defined as q = 4πsinθ/λ, where 2θ is the scattering angle. The SAXS flow cell was directly coupled with an online Agilent 1260 Infinity HPLC system using a Shodex KW803 SEC column equilibrated with a running buffer as indicated above with a flow rate of 0.5 mL/min. Each sample was run through the SEC, and three second X-ray exposures were collected continuously during a 30-minute elution. The SAXS frames recorded prior to the protein elution peak were used to subtract all other frames. The subtracted frames were investigated by the radius of gyration (Rg) derived by the Guinier approximation $I(q) = I(0)\ \exp(-q^2R_g^2/3)$ with the limits qRg<1.5**[73]. The elution peak was mapped by comparing the integral of ratios to background and Rg relative to the recorded frame using the program SCÅTTER. Uniform Rg values across an elution peak represent a homogeneous sample. Final merged SAXS profiles, derived by integrating multiple frames at the peak of the elution peak, were used for further analysis, including Guinier plot, which determined aggregation**-free state. Eluent was subsequently split 3 to 1 between the**

SAXS line and a series of detectors, including UV at 280 and 260 nm, multi-angle light scattering (MALS), quasi-elastic light scattering (QELS), and refractometer detector. MALS experiments were performed using an 18-angle DAWN HELEOS II light scattering detector connected in tandem to an Optilab refractive index concentration detector (Wyatt Technology). System normalization and calibration was performed with a BSA monomer using a 45 µL sample at 10 mg/mL in the same SEC running buffer and a dn/dc value of 0.19. The light scattering experiments were used to perform analytical scale chromatographic separations for Mw determination of the principal peaks in the SEC analysis. UV, MALS, and differential refractive index data were analyzed using Wyatt Astra 7 software to monitor the homogeneity of the sample across the elution peak complementary to the above-mentioned SEC-SAXS signal validation.

Two atomistic models of the GalS1 dimer were built based on close interfaces found in the crystal structure. The missing N-terminal region was modelled as a random coil using the program MODELLER[74]. The experimental SAXS profiles were then compared to theoretical scattering curves generated from atomistic models using FOXS[75, 76]. The SAXS envelope was restored in the P2 symmetry from the experimental data using the program GASBOR[77]. The average SAXS envelope was determined from 10 reconstructions using the DAMAVER program[78]. Structures and the SAXS envelopes were superimposed and visualized in CHIMERA[79].

Bayesian pattern-based evolutionary analysis of GT92 sequences

We first collected 259 GT92 sequences curated at the CAZy database[7]. Using the alignment of the GalS1 structure and other GT-A fold structures as a template, we then used a profile-based

alignment strategy, mapgaps[80], to align them to the core GT-A fold profile generated in our previous study[32]. This alignment allowed the mapping of GT-A features into the GalS1 structure. A representative set of 24816 sequences[32] was generated, including diverse GT-A fold families and GT92 sequences purged using an 80% sequence identity cutoff. This set was used to perform a query-centric Bayesian partition based on pattern selection (BPPS)[33] analysis with the GT92 consensus sequence as the query. This procedure clusters GT92 sequences into a distinct foreground group based on alignment positions that are most conserved within the GT92 family and distinguishes them from other GT-A fold enzymes grouped into the background.

Identification of critical residues in at the active site of GalS1

Docking and molecular dynamics (MD) simulations were employed to deduce the donor (UDP-Gal) and acceptor (Gal_4) binding sites and poses in GalS1. The Mn-bound monomeric structure of GalS1 was considered for both MD and docking studies. A sequential combination of molecular dynamics simulations, conducted using the CHARMM MD engine, and docking studies conducted using Autodock Vina, were used for modeling the enzyme-substrate complexes [81-83]. Considering that the Mn-bound crystal structure was elucidated in the absence of substrate molecules, the first set of simulations conducted were of the apo state (Mn-bound) of GalS1 under fully solvated conditions. The CHARMM 36 forcefield was used for proteins [84] and ions, including Mn^{2+}, and the TIP3P [85] forcefield for water molecules. The protonation states of the titratable amino acids in the proteins were estimated based on the H++ package[86] and disulfide linkages between residues 145-179, 236-316 and 369-447 were considered. In a 100 ns unbiased simulation of the solvated Mn-bound state of GALS1 was conducted and snapshots from this run were considered for the donor

molecule docking studies. An initial blind docking study was conducted wherein the whole GALS1 structure was considered for docking the donor molecule, followed by a more targeted docking study centered around the bound Mn^{2+} **ion. The targeted docking calculations involved a 40x40x40Å box with a grid spacing of 0.375Å, an exhaustiveness value of 128 and a total of 40 binding modes were explored. The best binding pose was selected for conducting 100 ns production MD simulations of the GalS1-UDP-Gal complex to evaluate the validity of the docked binding pose under fully solvated unbiased dynamical conditions. Prior to these MD simulations, a series of short, restrained simulations (totaling 2.24ns) were conducted to ensure proper equilibration of the active site residues around the bound donor molecule. Snapshots chosen from the donor bound simulations were then considered for docking studies of the acceptor molecule to obtain ternary complexes of Mn-bound GalS1 with UDP-Gal and Gal4 bound at the active site. Suitable docked poses of a putative ternary complex were then subjected to fully solvated unbiased MD simulations. The CHARMM36 forcefield was also employed for the MD simulations that involved the UDP-Gal and galactotetraose substrates** [87].

Generation and purification of galactotetraose $(Gal)_4$ acceptor substrate

The plasmid for heterologous expression of the β-1,4-galactanase GanA from *Geobacillus stearothermophilus* in pET9d was a kind gift provided by Dr. Yuval Shoham[88]. His-tagged GanA was heterologously expressed in *Escherichia coli,* purified using Ni-NTA chromatography, and concentrated stocks (2.5 mg/ml) were stored at -80℃ in 50 mM MES (pH 6.5), 100 mM NaCl, and 10% (v/v) glycerol. Potato galactan (500 mg) **(Megazyme, Product code: P-GALPOT) was dissolved in 50 mM MES, pH 6.5 to a final concentration of 10 mg/**

ml. 50 µg of GanA β-1,4-galactanase was added, and the digestion was allowed to proceed for 3 hr at 30°C shaking at 1000 rpm. Galacto-oligosaccharides were separated from the reaction mixture via diafiltration (10 kDa MWCO, Millipore, USA). An additional 50 µg of galactanase was added to undigested potato galactan retained in the filter device. The digest was repeated five times in total, with intermittent addition of enzyme and product removal. The galacto-oligosaccharides collected in the filtrates were pooled and lyophilized before loading onto a Bio-Gel P-2 (Bio-Rad, USA) column (120 ml, self-packed column) attached to an HPLC with water as a running buffer. The fractions were collected and analyzed using matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF MS) using a Microflex LT spectrometer (Bruker) as described below. Fractions containing galactotetraose were pooled and lyophilized.

Galactan synthase activity assays

All activated nucleotide sugars were purchased from CarboSource (USA), Promega (USA), or Sigma (USA). Screening of sugar-nucleotide donor specificities in the absence of acceptor substrate was done with the UDP-Glo™ Glycosyltransferase Assay (Promega, USA) kit[34]. Reactions (20 µl) consisted of 100 uM individual UDP-sugars (UDP-Gal, UDP-Ara*p*, UDP-Xyl, UDP-Glc, UDP-GalA, UDP-GlcA, UDP-GlcNAc, and UDP-GalNAc) and 4 µg of purified GalS1 in 50 mM HEPES, 100 mM NaCl, pH7 at 30°C for 18 hrs. 5 µl of the reaction mixture was mixed with an equal amount UDP-Glo™ reagent in a 384 well assay plate (Corning 4513) and incubated for 1 hr at room temperature before measuring luminescence using a Synergy LX Multi-mode Microplate Reader (BioTek, USA). A standard curve was used for the quantification of UDP produced.

The quantity of UDP formed as a by-product of the galactosyltransferase reaction was determined using the UDP-Glo™ Glycosyltransferase Assay (Promega) according to the manufacturer's instructions using either UDP-Gal (Promega, USA) or UDP-Ara*p* (CarboSource Services, USA) as donor substrates. Standard galactosyltransferase assays (20 µl) consisted of either UDP-Gal (250 µM) or UDP-Ara*p* (400 µM) as activated nucleotide sugar donors, galactotetraose (400 µM) as an acceptor and 5 mM manganese(II) chloride in 50 mM HEPES pH 7.0. Reactions were allowed to proceed at 30°C for 2 hr, and the amount of UDP produced was determined as described above.

Polysaccharide analysis using carbohydrate gel electrophoresis (PACE)

Reactions (25 µl) consisted of 2 µg galactotetraose as galacto-oligosaccharide substrate, 200 µM UDP-Gal and 20 µg purified protein and contained 10 mM $MnCl_2$ and 1% (v/v) Triton X-100 in 50 mM HEPES, pH 7.0. Reactions were incubated at 30°C for 2 h and then terminated by heating at 100°C for 5 min, followed by centrifugation at 10,000 x g for 10 min. Supernatants (15 µl) were mixed with 15 µl 3 M urea, and 5-µl-samples were loaded on large format Tris-borate acrylamide gel prepared as described previously[89], and electrophoresed at 200 V for 30 min followed by 1000 V for 1.5 h. The PACE gels were visualized with a G-Box gel doc system (Syngene, USA) at Tumi-wavelength with a UV detection filter and long-wave UV tubes (365 nm emission).

Microscale Thermophoresis (MST)

MST experiments to investigate the ability of the full-length protein and variants to bind UDP-Gal were performed on a NanoTemper ® Monolith NT.115 (NanoTemper Technologies,

Germany) with blue/red filters, as previously described[90]. His-GFP-GalS1 (or variants) were diluted 200X in MST Buffer 1 (1% Triton X-100, 10 mM $MnCl_2$, 50 mM HEPES, pH 7.0), and the final concentration yielded detectable fluorescent signals, between 200 and 1600 units of fluorescence (FU units). 10 µL of 5 mM UDP-Gal solution was diluted 1:1 in 10 µL MST buffer 1 to make a 16-sample serial dilution f**rom 2.5 mM to 76.3 nM. 10 µL of 5 µg/ml purified protein was added to 10 µL of each ligand solution and incubated at room temperature for 10 min. Prepared samples were loaded into standard treated capillaries for measurements using 20% MST power with laser off/on times of 0 s and 10 s, respectively, at 22 °C. All experiments were repeated three times for each measurement.**

MST experiments to evaluate the CBMXX were performed on a Monolith NT.115Pico (NanoTemper Technologies, Germany) equipped with blue/red filters. Non-adherent *Arabidopsis thaliana* seed coat mucilage, composed of almost pure RGI, was prepared according to the previously described method[91]. A protein solution of His-GFP-CBMXX (80 nM) or variants was prepared in MST Buffer 2 (0.02% of Tween 20, 10 mM of $MnCl_2$, 600 **mM NaCl, and 100 mM HEPES pH 7.0), mixed and centrifuged at 15,000 × g for 10 min to remove any potential aggregates. Substrate solutions (50 µl of 0.1 mg/ml) of non-adherent mucilage, galactotetraose, polygalacturonic acid (Sigma), and xylohexaose (Megazyme) were mixed 1:1 with 50 µl of the CBMXX protein solution. Samples were incubated for 5 min in the dark before MST analysis. Four aliquots of Standard capillaries were loaded with prepared samples, and the binding was checked. Binding affinity was measured using a 16-sample serial dilution from 8.3 µM to 0.25 nM. 10 µL of 160 nM purified protein was added to 10 µL of each ligand solution and incubated at room temperature for 10 min.**

Prepared samples were loaded into standard treated capillaries for measurements using 40% MST power with laser off/on times of 0 s and 10 s, respectively, at 22 °C. All experiments were repeated two times for each measurement.

MST experiments to investigate the ability of the full-length protein and variants to bind galactotetraose were also performed on a NanoTemper ® Monolith NT.115 (NanoTemper Technologies, Germany) with blue/red filters, similar to CBMs above, except that binding affinity was measured using a 16-sample serial dilution from 5 mM to 153 nM of the acceptor.

MALDI

MALDI spectra were acquired by using Microflex LT™ (Bruker Daltonics, Germany). 4 µg of the WT PtGalS1 with 10mM of UDP-Gal or UDP-Ara*p*, 0.5mM of galactotetrose, 1mM manganese(II) chloride in 50mM HEPES pH 7.0 in a total of 20µl reaction, the mixture was incubated overnight at 25°C. 5µl aliquots of each reaction were mixed with 1µl of Dowex-50 cation exchange resin (Bio-rad) and incubated for 1h on a microplate mixer. The tubes were centrifuged at 1250xg for 5 min. 1 µl sample of each sample is mixed with 1 µl matrix (2, 5-dihydroxybenzoic acid (DHB;100mg/ml in 50% methanol) on the plate and dried using blow-drying the spots to crystalize. Positive-ion spectra from 200 laser shots were added to generate the MALDI spectrum for each sample.

Sequence analysis of the stem domain

To prepare the sequence alignments of the stem domain, the sequence region spanning residues 1 -113 of GalS1 was blasted against the NCBI database using PSI blast (National

Centre for Biotechnology Institute). The top 100 sequences were taken for analysis, and hypothetical, predicted, and protein sequences with low-quality sequences were removed before. The sequences were aligned using the T-Coffee web server[92]; the web logo was created using (https://weblogo.berkeley.edu/logo.cgi).

Acknowledgments

Funding was provided in part by the Center for Bioenergy Innovation (CBI), from the U.S. Department of Energy Bioenergy Research Centers supported by the Office of Biological and Environmental Research in the DOE Office of Science. The work conducted by the Joint BioEnergy Institute is supported by the US Department of Energy, Office of Science, Office of Biological and Environmental Research under contract no. DE-AC02-05CH11231 between LBNL and the US Department of Energy. The Advanced Light Source is a Department of Energy Office of Science User Facility under Contract No. DE-AC02-05CH11231. The Berkeley Center for Structural Biology is supported in part by the Howard Hughes Medical Institute. The ALS-ENABLE beamlines are supported in part by the National Institutes of Health, National Institute of General Medical Sciences, grant P30 GM124169. Funding for the SIBYLS beamline at the Advanced Light Source was provided in part by the Offices of Science and Biological and Environmental Research, US Department of Energy, under Contract DE-AC02-05BH11231 and NIGMS grant P30 GM124169-01, ALS-ENABLE. Funding for NK from R35 GM139656 and R01 GM130915 is acknowledged.

Author Contributions

37

PKP, JHP, RT, WS, VB NK, MH, PDA, HVS, and BRU, designed experiments, performed experiments, and analyzed data. RT, VSB, NK performed computational simulations and machine learning studies. KWM and YJB and designed experiments, analyzed and interpreted data and edited the manuscript. PKP, JHP, RT, VB, NK, MH, PDA, HVS, BRU wrote the manuscript. PDA, HVS, and BRU conceived the project and BRU led the project.

Competing Interests statement

The authors declare no competing interests.

Data Availability

The data that supports the findings of this study and any computer code used herein are available from the corresponding authors upon request.

References

1. Carpita, N.C. Update on mechanisms of plant cell wall biosynthesis: how plants make cellulose and other (1->4)-β-D-glycans. *Plant physiology* **155**, 171-184 (2011).
2. Zabotina, O.A., Zhang, N. & Weerts, R. Polysaccharide Biosynthesis: Glycosyltransferases and Their Complexes. ***Frontiers in plant science*** **12, 625307-625307 (2021).**
3. Rini JM, E.J. in Essentials of Glycobiology [Internet]. Edn. 3rd. (ed. C.R. Varki A, Esko JD, et al., editors.) (Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press., 2017).
4. Na, L., Li, R. & Chen, X. Recent progress in synthesis of carbohydrates with sugar nucleotide-dependent glycosyltransferases. ***Curr Opin Chem Biol*** **61, 81-95 (2021).**
5. Liwanag, A.J.M. et al. Pectin Biosynthesis: GALS1 in Arabidopsis thaliana Is a β-1,4-Galactan β-1,4-Galactosyltransferase ***The Plant cell*** **24, 5024-5036 (2012).**
6. Ebert, B. et al. The Three Members of the Arabidopsis Glycosyltransferase Family 92 are Functional β-1,4-Galactan Synthases. ***Plant & cell physiology*** **59, 2624-2636 (2018).**

7. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P.M. & Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic acids research* 42, D490-495 (2014).
8. Atmodjo, M.A., Hao, Z. & Mohnen, D. Evolving views of pectin biosynthesis. *Annual review of plant biology* 64, 747-779 (2013).
9. Luis, A.S. & Martens, E.C. Interrogating gut bacterial genomes for discovery of novel carbohydrate degrading enzymes. *Current Opinion in Chemical Biology* 47, 126-133 (2018).
10. Laursen, T. et al. Bifunctional glycosyltransferases catalyze both extension and termination of pectic galactan oligosaccharides. *The Plant Journal* 94, 340-351 (2018).
11. Harholt, J., Suttangkakul, A. & Vibe Scheller, H. Biosynthesis of pectin. *Plant physiology* 153, 384-395 (2010).
12. Gorshkova, T. et al. Aspen Tension Wood Fibers Contain β-(1---> 4)-Galactans and Acidic Arabinogalactans Retained by Cellulose Microfibrils in Gelatinous Walls. *Plant physiology* 169, 2048-2063 (2015).
13. Ulvskov, P. et al. Biophysical consequences of remodeling the neutral side chains of rhamnogalacturonan I in tubers of transgenic potatoes. *Planta* 220, 609-620 (2005).
14. Obro, J. et al. Simultaneous in vivo truncation of pectic side chains. *Transgenic Res* 18, 961-969 (2009).
15. Mellerowicz, E.J. & Gorshkova, T.A. Tensional stress generation in gelatinous fibres: a review and possible mechanism based on cell-wall structure and composition. *Journal of experimental botany* 63, 551-565 (2012).
16. Zykwinska, A., Thibault, J.-F. & Ralet, M.-C. Organization of pectic arabinan and galactan side chains in association with cellulose microfibrils in primary cell walls and related models envisaged. *Journal of experimental botany* 58, 1795-1802 (2007).
17. Lin, D., Lopez-Sanchez, P. & Gidley, M.J. Binding of arabinan or galactan during cellulose synthesis is extensive and reversible. *Carbohydrate polymers* 126, 108-121 (2015).
18. McCartney, L., Steele-King, C.G., Jordan, E. & Knox, J.P. Cell wall pectic (1-->4)-beta-d-galactan marks the acceleration of cell elongation in the Arabidopsis seedling root meristem. *The Plant journal : for cell and molecular biology* 33, 447-454 (2003).
19. McCartney, L., Ormerod, A.P., Gidley, M.J. & Knox, J.P. Temporal and spatial regulation of pectic (1-->4)-beta-D-galactan in cell walls of developing pea cotyledons: implications for mechanical properties. *The Plant journal : for cell and molecular biology* 22, 105-113 (2000).
20. Klaassen, M.T. & Trindade, L.M. RG-I galactan side-chains are involved in the regulation of the water-binding capacity of potato cell walls. *Carbohydrate polymers* 227, 115353 (2020).
21. Culbertson, A.T., Ehrlich, J.J., Choe, J.Y., Honzatko, R.B. & Zabotina, O.A. Structure of xyloglucan xylosyltransferase 1 reveals simple steric rules that define biological patterns of xyloglucan polymers. *Proceedings of the National Academy of Sciences of the United States of America* 115, 6064-6069 (2018).

39

22. Urbanowicz, B.R. et al. Structural, mutagenic and in silico studies of xyloglucan fucosylation in Arabidopsis thaliana suggest a water-mediated mechanism. *The Plant journal : for cell and molecular biology* 91, 931-949 (2017).
23. Rocha, J. et al. Structure of Arabidopsis thaliana FUT1 Reveals a Variant of the GT-B Class Fold and Provides Insight into Xyloglucan Fucosylation. *The Plant cell* 28, 2352-2364 (2016).
24. Smith, P.J. et al. Enzymatic Synthesis of Artificial Polysaccharides. *ACS Sustainable Chemistry & Engineering* 8, 11853-11871 (2020).
25. Loqué, D., Scheller, H.V. & Pauly, M. Engineering of plant cell walls for enhanced biofuel production. *Curr Opin Plant Biol* 25, 151-161 (2015).
26. Prabhakar, P.K. et al. in Methods in Cell Biology, Vol. 160. (eds. C.T. Anderson, E.S. Haswell & R. Dixit) 145-165 (Academic Press, 2020).
27. Montanier, C. et al. Circular permutation provides an evolutionary link between two families of calcium-dependent carbohydrate binding modules. *The Journal of biological chemistry* 285, 31742-31754 (2010).
28. Cid, M. et al. Recognition of the helical structure of beta-1,4-galactan by a new family of carbohydrate-binding modules. *The Journal of biological chemistry* 285, 35999-36009 (2010).
29. Macquet, A., Ralet, M.C., Kronenberger, J., Marion-Poll, A. & North, H.M. In situ, chemical and macromolecular study of the composition of Arabidopsis thaliana seed coat mucilage. *Plant & cell physiology* 48, 984-999 (2007).
30. Haughn, G. & Western, T. Arabidopsis Seed Coat Mucilage is a Specialized Cell Wall that Can be Used as a Model for Genetic Analysis of Plant Cell Wall Structure and Function. *Frontiers in plant science* 3 (2012).
31. Venditto, I. et al. Complexity of the Ruminococcus flavefaciens cellulosome reflects an expansion in glycan recognition. *Proceedings of the National Academy of Sciences* 113, 7136-7141 (2016).
32. Taujale, R. et al. Deep evolutionary analysis reveals the design principles of fold A glycosyltransferases. *Elife* 9 (2020).
33. Neuwald, A.F. A Bayesian sampler for optimization of protein domain hierarchies. *J Comput Biol* 21, 269-286 (2014).
34. Sheikh, M.O. et al. Rapid screening of sugar-nucleotide donor specificities of putative glycosyltransferases. *Glycobiology* 27, 206-212 (2017).
35. Tsutsui, Y., Ramakrishnan, B. & Qasba, P.K. Crystal structures of β-1,4-galactosyltransferase 7 enzyme reveal conformational changes and substrate binding. *The Journal of biological chemistry* 288, 31963-31970 (2013).
36. Venkat, A. et al. Modularity of the hydrophobic core and evolution of functional diversity in fold A glycosyltransferases. *Journal of Biological Chemistry*, 102212 (2022).
37. Prabhakar, P.K., Rao, K.K. & Balaji, P.V. The Cys78–Asn88 loop region of the Campylobacter jejuni CstII is essential for α2,3-

sialyltransferase activity: analysis of the His85 mutants. *The Journal of Biochemistry* 156, 229-238 (2014).
38. Prabhakar, P.K., Srivastava, A., Rao, K.K. & Balaji, P.V. Monomerization alters the dynamics of the lid region in Campylobacter jejuni CstII: an MD simulation study. *Journal of Biomolecular Structure and Dynamics* 34, 778-791 (2016).
39. Lunin, V.V. et al. Molecular Mechanism of Polysaccharide Acetylation by the Arabidopsis Xylan O-acetyltransferase XOAT1. *The Plant cell* 32, 2367-2382 (2020).
40. Wang, H.-T. et al. Rational enzyme design for controlled functionalization of acetylated xylan for cell-free polymer biosynthesis. *Carbohydrate polymers* 273, 118564 (2021).
41. Zhang, Y. et al. Roles of active site tryptophans in substrate binding and catalysis by α-1,3 galactosyltransferase. *Glycobiology* 14, 1295-1302 (2004).
42. van der Veen, B.A. et al. Hydrophobic Amino Acid Residues in the Acceptor Binding Site Are Main Determinants for Reaction Mechanism and Specificity of Cyclodextrin-glycosyltransferase*. *Journal of Biological Chemistry* 276, 44557-44562 (2001).
43. Yang, T. et al. Hydrophobic recognition allows the glycosyltransferase UGT76G1 to catalyze its substrate in two orientations. *Nature Communications* 10, 3214 (2019).
44. Abbott, D.W. & van Bueren, A.L. Using structure to inform carbohydrate binding module function. *Current opinion in structural biology* 28, 32-40 (2014).
45. Breton, C., Mucha, J. & Jeanneau, C. Structural and functional features of glycosyltransferases. *Biochimie* 83, 713-718 (2001).
46. Lowe J.B., V.A. in Essentials of Glycobiology. (ed. C.R. Varki A, Esko J, et al., editors.) (Cold Spring Harbor Laboratory Press; Cold Spring Harbor (NY). 1999).
47. Kellokumpu, S., Hassinen, A. & Glumoff, T. Glycosyltransferase complexes in eukaryotes: long-known, prevalent but still unrecognized. *Cellular and Molecular Life Sciences* 73, 305-325 (2016).
48. Boeggeman, E.E., Ramakrishnan, B. & Qasba, P.K. The N-terminal stem region of bovine and human β1,4-galactosyltransferase I increases the in vitro folding efficiency of their catalytic domain from inclusion bodies. *Protein expression and purification* 30, 219-229 (2003).
49. Grabenhorst, E. & Conradt, H.S. The cytoplasmic, transmembrane, and stem regions of glycosyltransferases specify their in vivo functional sublocalization and stability in the Golgi. *J Biol Chem* 274, 36107-36116 (1999).
50. Ramakrishnan, B., Balaji, P.V. & Qasba, P.K. Crystal structure of beta1,4-galactosyltransferase complex with UDP-Gal reveals an oligosaccharide acceptor binding site. *Journal of molecular biology* 318, 491-502 (2002).
51. Ramakrishnan, B., Boeggeman, E. & Qasba, P.K. Binding of N-Acetylglucosamine (GlcNAc) β1–6-branched Oligosaccharide Acceptors to β4-Galactosyltransferase I Reveals a New Ligand

Binding Mode*. *Journal of Biological Chemistry* 287, 28666-28674 (2012).
52. Valdez, H.A. et al. Role of the N-terminal starch-binding domains in the kinetic properties of starch synthase III from Arabidopsis thaliana. *Biochemistry* 47, 3026-3032 (2008).
53. Gámez-Arjona, F.M. & Mérida, Á. Interplay Between the N-Terminal Domains of Arabidopsis Starch Synthase 3 Determines the Interaction of the Enzyme With the Starch Granule. *Frontiers in plant science* 12, 704161 (2021).
54. Christiansen, C. et al. The carbohydrate-binding module family 20--diversity, structure, and function. *The FEBS journal* 276, 5006-5029 (2009).
55. Hedin, N., Velazquez, M.B., Barchiesi, J., Gomez-Casati, D.F. & Busi, M.V. CBM20CP, a novel functional protein of starch metabolism in green algae. *Plant Mol Biol* 108, 363-378 (2022).
56. Zhu, Y., Zhang, M., Kelly, A.R. & Cheng, A. The carbohydrate-binding domain of overexpressed STBD1 is important for its stability and protein-protein interactions. *Biosci Rep* 34, e00117 (2014).
57. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373, 871-876 (2021).
58. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583-589 (2021).
59. Cornil, I., Kerbel, R.S. & Dennis, J.W. Tumor cell surface beta 1-4-linked galactose binds to lectin(s) on microvascular endothelial cells and contributes to organ colonization. *J Cell Biol* 111, 773-781 (1990).
60. Raz, A. & Lotan, R. Endogenous galactoside-binding lectins: a new class of functional tumor cell surface molecules related to metastasis. *Cancer Metastasis Rev* 6, 433-452 (1987).
61. Moremen, K.W. et al. Expression system for structural and functional studies of human glycosylation enzymes. *Nat Chem Biol* 14, 156-162 (2018).
62. Reeves, P.J., Callewaert, N., Contreras, R. & Khorana, H.G. Structure and function in rhodopsin: high-level expression of rhodopsin with restricted and homogeneous N-glycosylation by a tetracycline-inducible N-acetylglucosaminyltransferase I-negative HEK293S stable mammalian cell line. *Proceedings of the National Academy of Sciences of the United States of America* 99, 13419-13424 (2002).
63. Pereira, J.H., McAndrew, R.P., Tomaleri, G.P. & Adams, P.D. Berkeley Screen: a set of 96 solutions for general macromolecular crystallization. *J Appl Crystallogr* 50, 1352-1358 (2017).
64. Winter, G., Lobley, C.M. & Prince, S.M. Decision making in xia2. *Acta Crystallogr D Biol Crystallogr* 69, 1260-1273 (2013).
65. Hendrickson, W.A. Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. *Science* 254, 51-58 (1991).
66. Terwilliger, T.C. et al. Decision-making in structure solution using Bayesian estimates of map quality: the PHENIX AutoSol wizard. *Acta Crystallographica Section D %@ 0907-4449* 65, 582-601 (2009).

67. Terwilliger, T.C. et al. Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. *Acta Crystallogr D Biol Crystallogr* 64, 61-69 (2008).
68. Adams, P.D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* 66, 213-221 (2010).
69. Afonine, P.V. et al. Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr D Biol Crystallogr* 68, 352-367 (2012).
70. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* 60, 2126-2132 (2004).
71. Classen, S. et al. Implementation and performance of SIBYLS: a dual endstation small-angle X-ray scattering and macromolecular crystallography beamline at the Advanced Light Source. *J Appl Crystallogr* 46, 1-13 (2013).
72. Hura, G.L. et al. Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS). *Nat Methods* 6, 606-612 (2009).
73. Guinier A, F.F. Small angle scattering of X-rays. (Wiley Interscience, New York, 1955).
74. Sali, A. & Blundell, T.L. Comparative protein modelling by satisfaction of spatial restraints. *Journal of molecular biology* 234, 779-815 (1993).
75. Schneidman-Duhovny, D., Hammel, M., Tainer, John A. & Sali, A. Accurate SAXS Profile Computation and its Assessment by Contrast Variation Experiments. *Biophysical Journal* 105, 962-974 (2013).
76. Schneidman-Duhovny, D., Hammel, M. & Sali, A. FoXS: a web server for rapid computation and fitting of SAXS profiles. *Nucleic acids research* 38, W540-W544 (2010).
77. Svergun, D.I., Petoukhov, M.V. & Koch, M.H.J. Determination of Domain Structure of Proteins from X-Ray Solution Scattering. *Biophysical Journal* 80, 2946-2953 (2001).
78. Volkov, V.V. & Svergun, D.I. Uniqueness of ab initio shape determination in small-angle scattering. *Journal of Applied Crystallography %@ 0021-8898* 36, 860-864 (2003).
79. Pettersen, E.F. et al. UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* 25, 1605-1612 (2004).
80. Neuwald, A.F. Rapid detection, classification and accurate alignment of up to a million or more related protein sequences. *Bioinformatics* 25, 1869-1875 (2009).
81. Brooks, B.R. et al. CHARMM: the biomolecular simulation program. *J Comput Chem* 30, 1545-1614 (2009).
82. Eberhardt, J., Santos-Martins, D., Tillack, A.F. & Forli, S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *Journal of Chemical Information and Modeling* 61, 3891-3898 (2021).
83. Trott, O. & Olson, A.J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient

optimization, and multithreading. *J Comput Chem* 31, 455-461 (2010).
84. Jorgensen, W., Chandrasekhar, J., Madura, J., Impey, R. & Klein, M. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* 79, 926-935 (1983).
85. Huang, J. & MacKerell, A.D., Jr. CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. *J Comput Chem* 34, 2135-2145 (2013).
86. Anandakrishnan, R., Aguilar, B. & Onufriev, A.V. H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic acids research* 40, W537-W541 (2012).
87. Guvench, O., Hatcher, E., Venable, R.M., Pastor, R.W. & MacKerell, A.D. CHARMM Additive All-Atom Force Field for Glycosidic Linkages between Hexopyranoses. *Journal of Chemical Theory and Computation* 5, 2353-2370 (2009).
88. Tabachnikov, O. & Shoham, Y. Functional characterization of the galactan utilization system of Geobacillus stearothermophilus. *The FEBS journal* 280, 950-964 (2013).
89. Goubet, F., Jackson, P., Deery, M.J. & Dupree, P. Polysaccharide analysis using carbohydrate gel electrophoresis: a method to study plant cell wall polysaccharides and polysaccharide hydrolases. *Anal Biochem* 300, 53-68 (2002).
90. Shao, W., Sharma, R., Clausen, M.H. & Scheller, H.V. Microscale thermophoresis as a powerful tool for screening glycosyltransferases involved in cell wall biosynthesis. *Plant Methods* 16, 99 (2020).
91. Zhao, J. et al. Isolation of a lectin binding rhamnogalacturonan-I containing pectic polysaccharide from pumpkin. *Carbohydrate polymers* 163, 330-336 (2017).
92. Notredame, C., Desmond, H.G. & Heringa, J. T-Coffee: A novel method for multiple sequence alignments" C. Notredame, D. Higgins, J. Heringa. *J of Mol Bio* 302, 205-217 (2000).

Figure Legends

Fig.~~ure~~ 1: ~~_~~–The role of GalS1 in rhamnogalacturonan I (RG-I) synthesis. a~~A.~~) Schematic model of RGI highlighting common, known side chains arabinan, galactan, and arabinogalactan[9]. ~~b)~~b. Schematic of an *~~i~~In vitro* galactosyltransferase reaction scheme showing

the transfer of a ~~galactosyl~~galactopyranosyl or arabinopyranosyl **residue from UDP-α-D-Gal*p*** or UDP-β-L-Ara*p* to a β-1,4-galactooligosaccharide acceptor ~~by GalS1~~illustrating the extension or capping activity of GalS1, respectively. **c-e**, MALDI-TOF MS of the GalS1 saccharide products after a 16-h reaction using galactotetraose as an an acceptor.~~MALDI showing~~ **c~~C~~.** Control (no-enzyme). **~~d,~~** PtGalS1 activity using (**d**) ~~galactotetraose and~~ UDP-α-D-Gal*p* ~~UDP-Gal as substrates~~or (**e**) UDP-β-L-Ara*p* as a donor to demonstrate extension versus capping, respectively. **d**, PtGalS1 activity using UDP-α-D-Galp as a donor The series of annotated [M + H]+ ions are the result of structures with a mass difference of 162 Da consistent with the sequential addition of galactosyl residues to the galactotetraose acceptor to generate oligosaccharides products with degrees of polymerization (DP) ranging from 5 to 30. **e**, PtGalS1 activity using UDP-β-L-Arap as a donor. The annotated [M + H]+ ion is a capped structure with a mass difference of 264 Da consistent with the sequential addition of two arabinosyl (132 Da, pentosyl) residues to the galactotetraose acceptor. ~~e, PtGalS1 activity using galactotetraose and UDP-Ara*p* as substrates.~~

Fig~~ure~~ 2: The structure of GalS1 obtained by X-ray crystallography. a~~)~~ Monomer of GalS1 showing highlighting the stem domain (blue), CBMXX (magenta), and core GT-A domain **(grey). The secondary structures are displayed as cartoon models with transparent surfaces. b~~)~~** Cartoon representation of a GalS1 homodimer emphasizing the dimer interface. Interacting residues (up to 4Å) between each monomer in a dimer **are shown in yellow and green, respectively. c~~)~~** Experimental (black) and theoretical (colored as indicated) SAXS profiles for the solution state models fitting of GalS1. Parallel (*P) and antiparallel (*AP) configurations of the homodimer are indicated. **Fit are shown together with the fit residuals**

and goodness of fit values (□2). Guinier plots that determined the aggregation free state for the experimental SAXS curve are shown in the inset. Average SAXS envelop (gray transparent) is superimposed onto the atomic model of the solution state–dimer (red and blue). Glycosylation, as seen in the crystal structure, is highlighted (green).

Fig.~~ure~~ 3: The CBM XX domain is the founding member of a new CAZy family. a~~A,~~) and **b~~),~~** GalS1-CBMXX represented in gray superimposed with ***Thermotoga maritima*** **CBM61 (PDB ID: 2XOM, red) and *Cellvibrio japonicus* vCBM60 (PDB ID: 2XFD, green), respectively. c~~),~~** Putative binding residues on the surface of the CBMXX domain. **d~~),~~** Comparison of dissociation **constants (K_D) of CBMXX WT and its variants obtained by MST.** The values shown are **K_D** obtained after using **K_D** fit model in the MO.Affinity Analysis software (NanoTemper Technologies) ~~with K_D confidence (in brackets) of a representative experiment performed in duplicates.~~ ± KD confidence (SD) is indicated next to KD values. Confidence (SD) values define the range where the KD falls with 68% of certainty. Error Bar: SD, n = 2.

Fig.~~ure~~ 4: Conserved pattern positions that distinguish the GT92 family. a,~~a)~~ Web logos depict the extent of conservation of residues at any given position in 153 GT92 sequences (top) versus other GT-A fold sequences (bottom). Residue numbers based on GalS1 positions are indicated above the red bars. Based on the alignment, the four core GT-A motifs (DXD, G-loop, xED, and C-His) are labeled at the bottom. Structural alignment of ~~*b)*~~ *Drosophila* β4GalT7 D211N complex with manganese, UDP-Gal, and xylobiose (in red) with GalS1 (current study, in blue) showing critical residues identified in the GalS1 for **b**, donor binding **c**, acceptor binding based on proximity or conservation in GT92 sequences. ~~**Highlighted**~~

~~residues studied in the current work. c) Hypervariable regions (HV; predicted to impart acceptor specificity to GalS1) and core-hydrophobic regions are shown in yellow.~~ **d~~)~~, Galactosyltransferase (GalT) and e~~)~~, Arabinosyltransferase (AraT) activity of GalS1 WT and its variants** using UDP-α-D-Gal*p* or UDP-β-L-Ara*p* as donors, respectively, **in the presence and absence of galactotetraose acceptor (Gal_4).** The values shown are average ± standard deviation of a representative experiment performed in triplicate.

Fig~~ure~~ 5: Insights from docking and MD simulations. a~~)~~, RMSF difference plot showing regions that are more flexible during the apo state (blue) and regions that are more flexible in the donor bound state (red). Most structural regions don't show significant differences between the two states. **b~~)~~,** Snapshot of the active site from the MD simulation of the donor bound GalS1 complex. **c~~)~~,** Docking results show primary binding grooves on GalS1 capable of binding the acceptor substrate. **d~~)~~,** Snapshot of the active site from the MD simulation of the acceptor and donor bound GalS1 ternary complex.

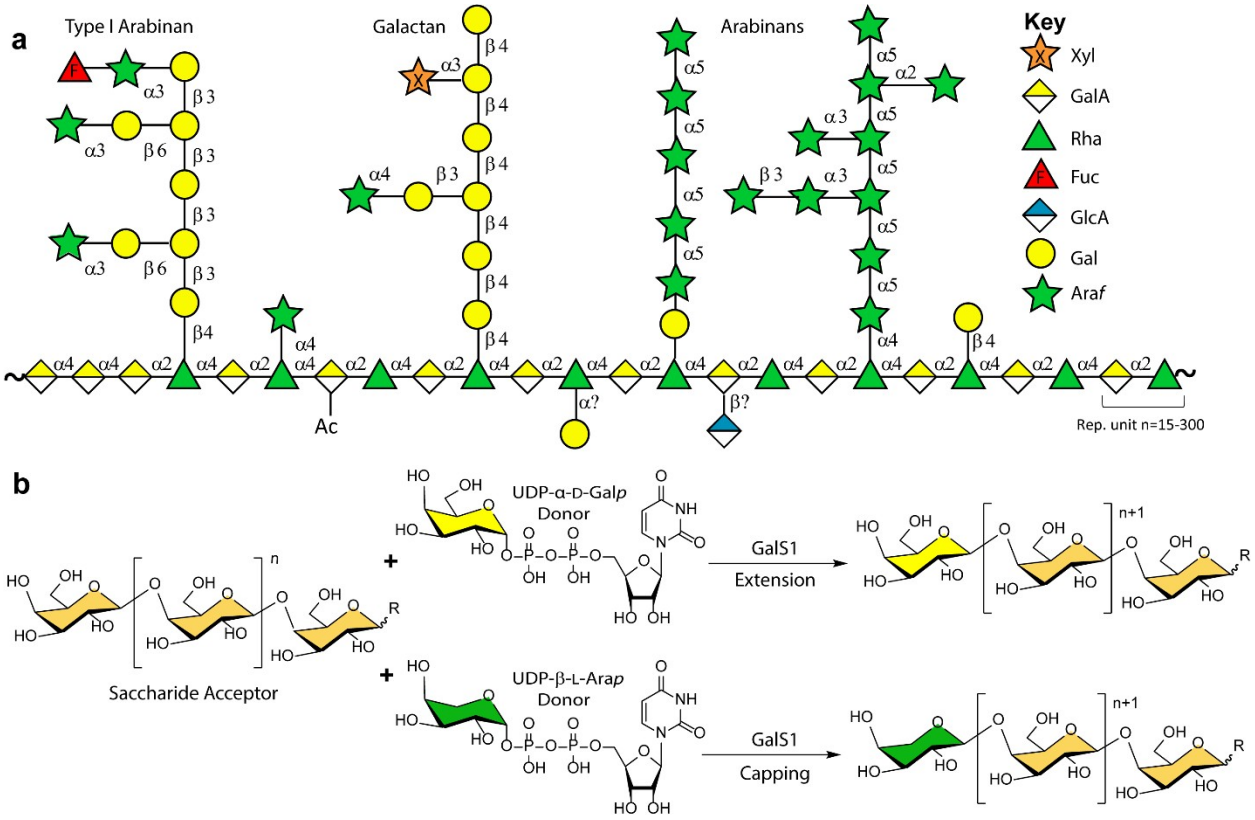
Fig~~ure~~ 6: Proposed mechanism of galactan synthesis by GalS1. a~~A~~~~)~~, GalS1 is proposed to utilize an inverting, S_N2 single displacement reaction mechanism. The catalytic base (His 414, blue outline) deprotonates the acceptor nucleophilic hydroxyl, which in turn attacks the anomeric carbon of the UDP-α-D-Gal donor nucleophile and displaces the phosphate leaving group. The Red O highlights the nucleophilic oxygen at O-4 of the galactooligosaccharide acceptor and the Red C indicates the anomeric carbon of the UDP-α-D-Gal donor. b~~)~~, Schematic representation of RGI binding and galactan chain elongation by homodimeric GalS1. Chain A (CBM in pink and core GT-A domain in light pink**) with Chain B (CBM in blue and core GT-A domain in light blue cyan with some part of stem region),**

transmembrane and cytoplasmic domains are shown by dashed lines. The RGI backbone with a single galactan chain is shown for simplicity (see Fig.1a for glycan symbols). The redline outlines the contour of the GalS1 active site, and black lines indicate the CBMXX binding site proposed to interact with the RGI backbone as the GT domain catalyzes galactan synthesis.

Table 1: Comparison of UDP-Gal (donor substrate) equilibrium dissociation constant (K_D) ± standard deviation of GalS1 WT and its variants.

GalS1 variants	K_D* for UDP-Gal (µM)
GalS1 WT	198±64
D331A	253±131
D333A	18.2±11.6
E334A	No binding
K400A	265±116
Q309A	1000±300
H435A	163±57
H414A	508±739
G242A	116±29
H437A	55±23
D398A	46.6±18.7
R396A	23.0±10.4
W166A	115 ±44
ΔStem	No binding

Fig. 1 (option 1)



1186

1187

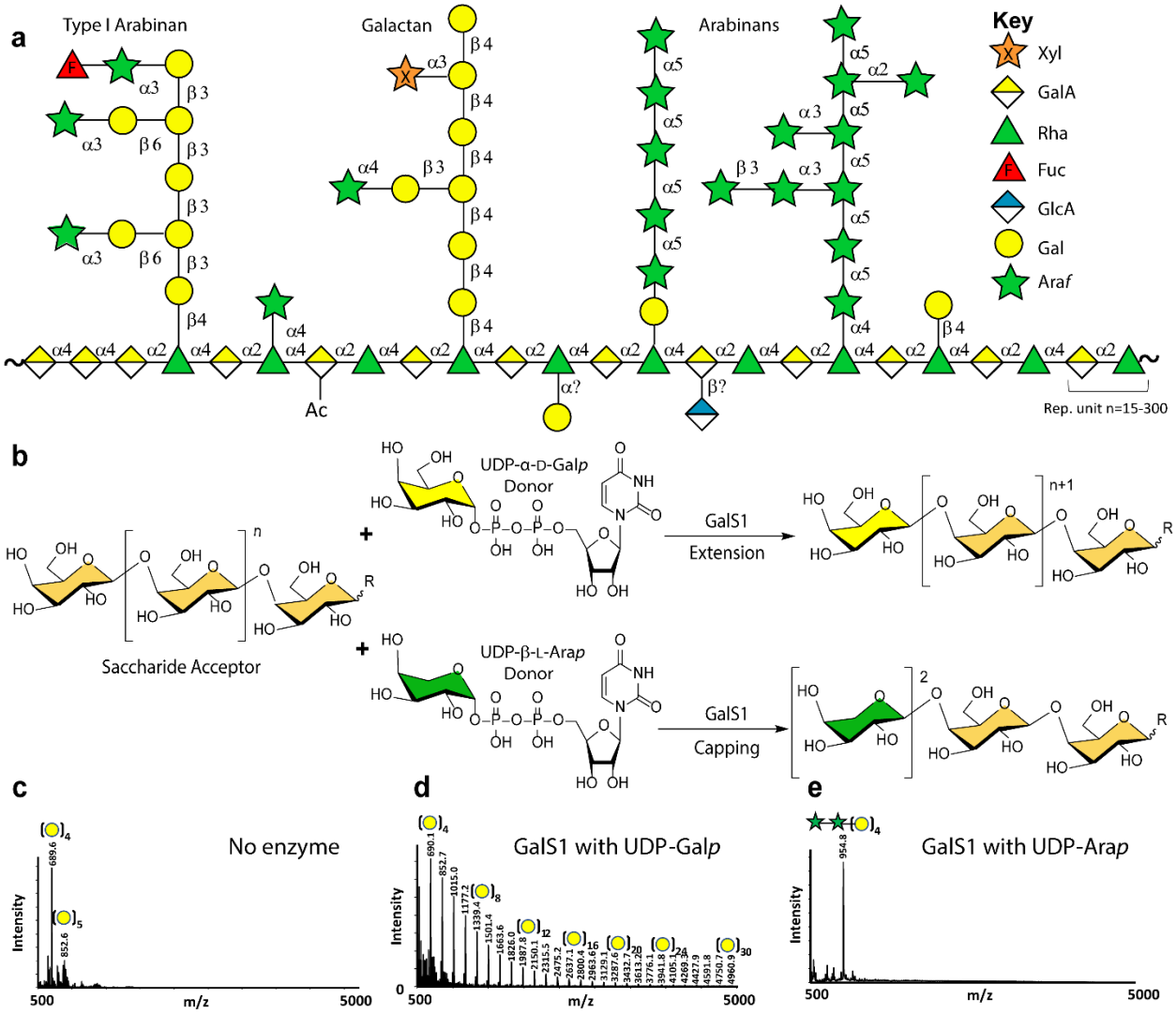
1188

1189

50

1190 | **Fig. 1 (2nd option)**

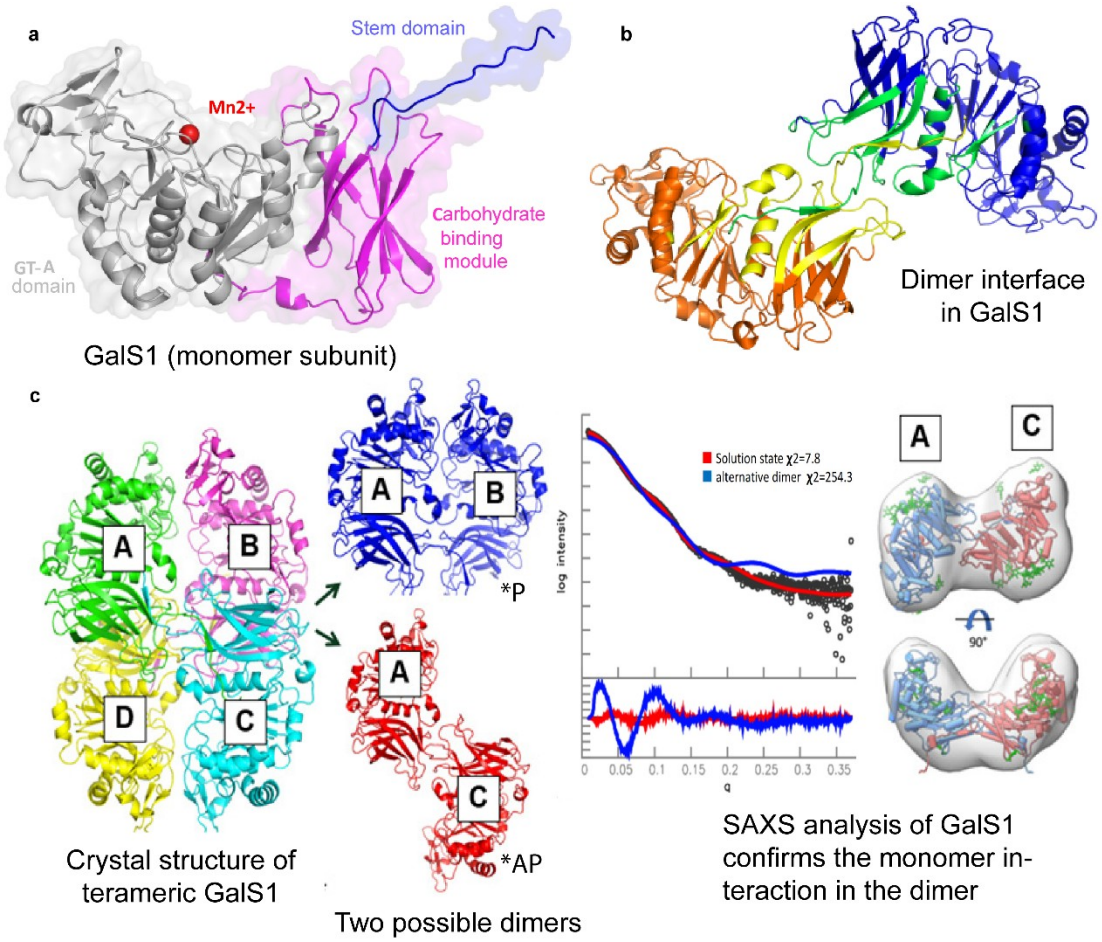
1191



1192

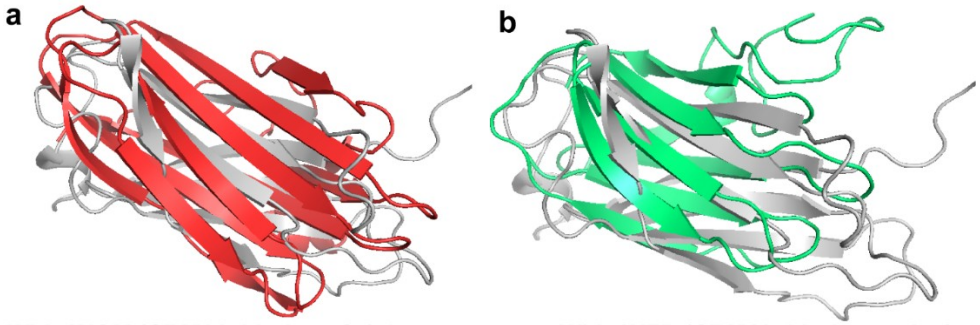
51

Fig. 2


a
Stem domain
Mn2+
Carbohydrate binding module
GT-A domain
GalS1 (monomer subunit)
b
Dimer interface in GalS1
c
A
B
D
C
Crystal structure of terameric GalS1
*P
*AP
Two possible dimers
Solution state χ2=7.8
alternative dimer χ2=254.3
log intensity
q
SAXS analysis of GalS1 confirms the monomer interaction in the dimer

52

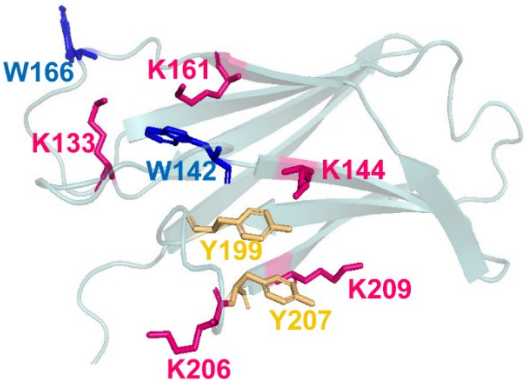
Fig. 3



With 2XOM (CBM61; binds to β-1,4-galactan) RMSD is 7.6Å over 88 residues.

With 2XFD (CBM60; binds to xylan) RMSD is 4.4Å over 64 residues.

c

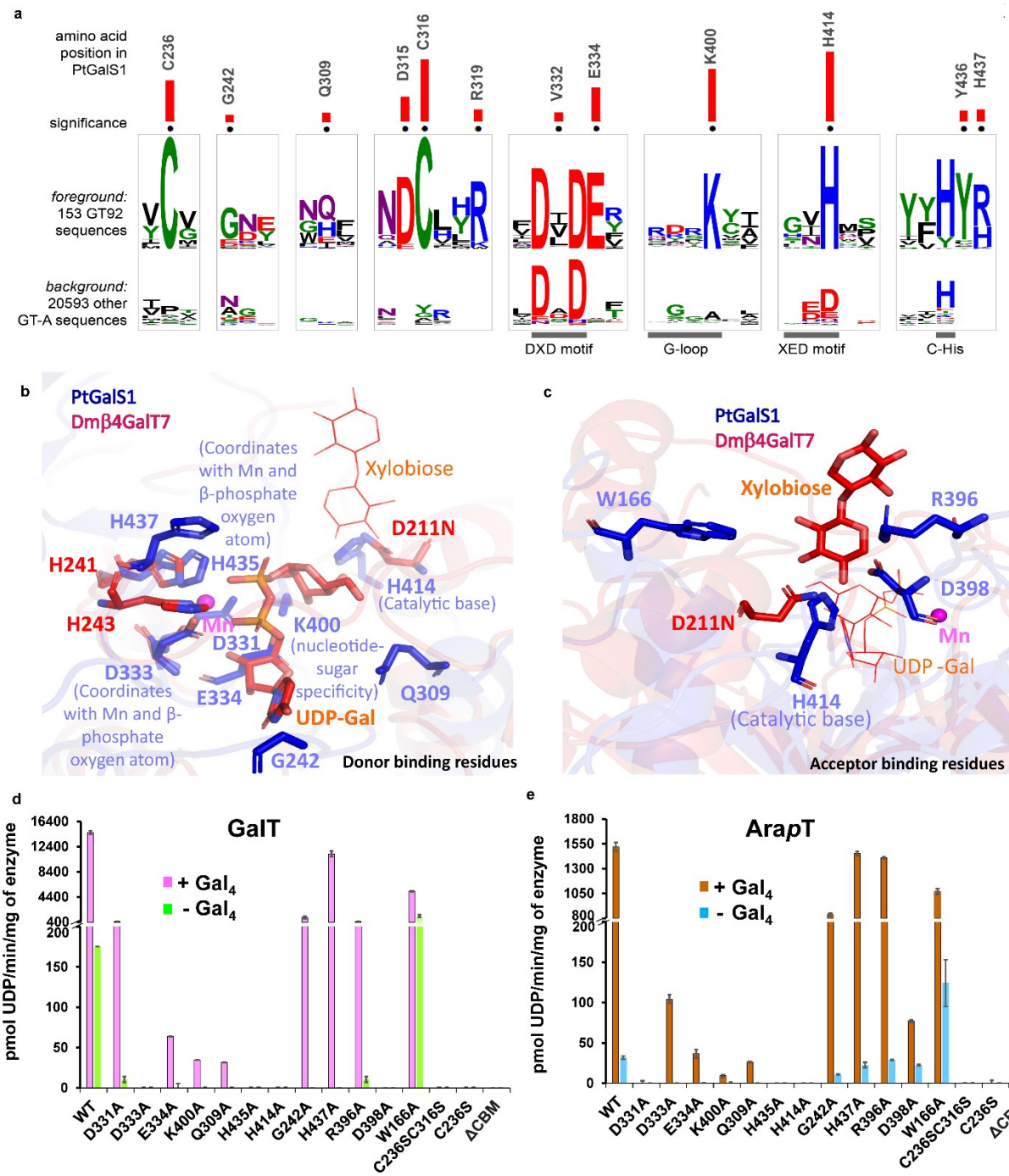


d

CBM domain variants	Dissociation constant (μM)
K133A	0.44 ± 0.1
K144A	1.7 ± 2.3
W142A	0.63 ± 0.7
W166A	1.12 ± 0.4
Y199A	0.73 ± 0.3
K206A	0.79 ± 0.8
Y207A	1.18± 2.1
K209A	0.42 ± 0.5
WT	0.13 ± 0.1

1208 **Fig. 4**

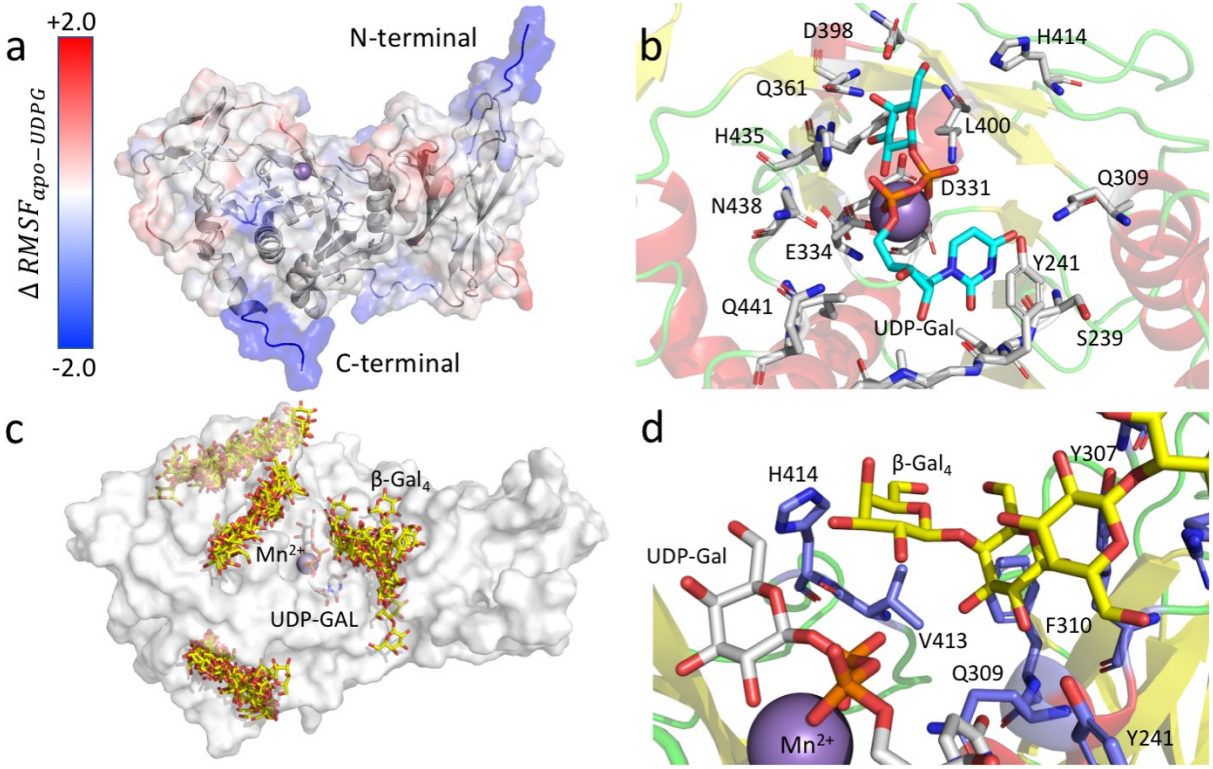
54


a
amino acid position in PtGalS1
significance
foreground: 153 GT92 sequences
background: 20593 other GT-A sequences
C236
G242
Q309
D315
C316
R319
V332
E334
K400
H414
Y436
H437
DXD motif
G-loop
XED motif
C-His
b
PtGalS1
Dmβ4GalT7
(Coordinates with Mn and β-phosphate oxygen atom)
Xylobiose
D211N
H437
H241
H435
H414
(Catalytic base)
H243
Mn
D331
K400
(nucleotide-sugar specificity)
D333
(Coordinates with Mn and β-phosphate oxygen atom)
E334
Q309
UDP-Gal
G242
Donor binding residues
c
PtGalS1
Dmβ4GalT7
W166
Xylobiose
R396
D398
D211N
Mn
UDP-Gal
H414
(Catalytic base)
Acceptor binding residues
d
GalT
pmol UDP/min/mg of enzyme
+ Gal4
- Gal4
WT
D331A
D333A
E334A
K400A
Q309A
H435A
H414A
G242A
H437A
R396A
D398A
W166A
C236SC316S
C236S
ΔCBM
e
ArapT
pmol UDP/min/mg of enzyme
+ Gal4
- Gal4
WT
D331A
D333A
E334A
K400A
Q309A
H435A
H414A
G242A
H437A
R396A
D398A
W166A
C236SC316S
C236S
ΔCBM

1209

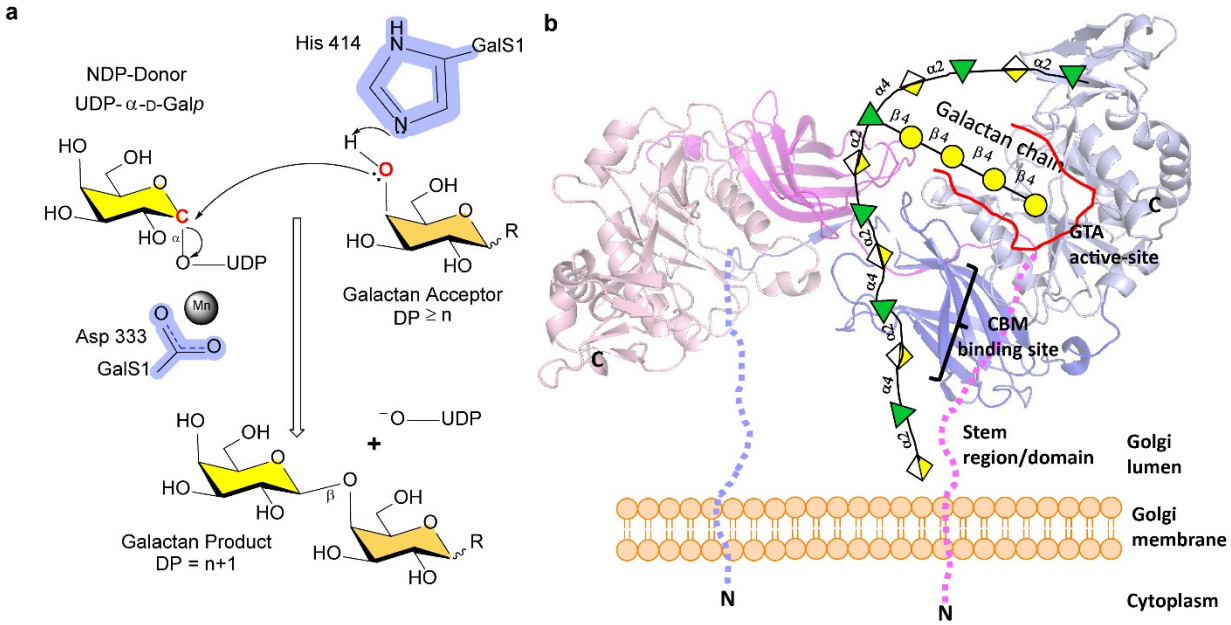
55

Fig. 5


a
+2.0
-2.0
$\Delta RMSF_{apo-UDPG}$
N-terminal
C-terminal
b
D398
H414
Q361
L400
H435
D331
Q309
N438
E334
Y241
Q441
UDP-Gal
S239
c
β-Gal4
Mn2+
UDP-GAL
d
H414
β-Gal4
Y307
UDP-Gal
V413
F310
Q309
Mn2+
Y241

56

Fig. 6


a
His 414
GalS1
NDP-Donor
UDP-α-D-Galp
Galactan Acceptor
DP ≥ n
Asp 333
GalS1
Mn
Galactan Product
DP = n+1
b
Galactan chain
GTA
active-site
CBM
binding site
Stem
region/domain
Golgi
lumen
Golgi
membrane
Cytoplasm
C
N

57