Nat Chem Biol. Author manuscript; available in PMC 2022 August 01.

Published in final edited form as:
Nat Chem Biol. 2022 August ; 18(8): 841–849. doi:10.1038/s41589-022-01039-x.

Sulfated glycan recognition by carbohydrate sulfatases of the human gut microbiota

Ana S Luis[1,2,*], **Arnaud Baslé**[3], **Dominic P Byrne**[4], **Gareth SA Wright**[4], **James London**[4], **Jin Chunsheng**[2], **Niclas G Karlsson**[2,5], **Gunnar C Hansson**[2], **Patrick A Eyers**[4], **Mirjam Czjzek**[6], **Tristan Barbeyron**[6], **Edwin A Yates**[4], **Eric C. Martens**[1], **Alan Cartmell**[4,*]

[1]Department of Microbiology and Immunology, University of Michigan, Ann Arbor, MI 48109, USA

[2]Department of Medical Biochemistry and Cell Biology, University of Gothenburg, Box 440, 405 30 Gothenburg, Sweden

[3]Institute for Cell and Molecular Biosciences, Newcastle University, Newcastle upon Tyne, United Kingdom

[4]Department of Biochemistry and Systems Biology, Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool L69 3BX, United Kingdom

[5]Faculty of Health Sciences, Department of Life Sciences and Health, Pharmacy, Oslo Metropolitan University, 0130 Oslo, Norway

[6]Sorbonne Université, Univ Paris 06, CNRS, UMR 8227, Integrative Biology of Marine Models, Station Biologique de Roscoff, CS 90074, Roscoff, Bretagne, France

Abstract

Sulfated glycans are ubiquitous nutrient sources for microbial communities that have co-evolved with eukaryotic hosts. Bacteria metabolise sulfated glycans by deploying carbohydrate sulfatases that remove sulfate esters. Despite the biological importance of sulfatases, the mechanisms underlying their ability to recognise their glycan substrate remain poorly understood. Here, we utilise structural biology to determine how sulfatases from the human gut microbiota recognise sulfated glycans. We reveal 7 new carbohydrate sulfatase structures span four S1 sulfatase subfamilies. Structures of S1_16 and S1_46 represent the first structures of these subfamilies. Structures of S1_11 and S1_15 demonstrate how non-conserved regions of the protein drive specificity towards related but distinct glycan targets. Collectively, these data reveal that

*Correspondence to: Alan.Cartmell@liverpool.ac.uk, ana.luis@medkem.gu.se.

Author contributions

ASL and AC conceived and designed experiments. AC, ASL and ECM wrote the draft manuscript. ASL and AC cloned, expressed, purified sulfatases and performed the enzymatic assays. AC, DPB, and JAL carried out and analysed kinetic and binding experiments. EY and JAL performed labelling and NMR experiments. AC and AB performed structural biology experiments. MC and TB performed sulfatase phylogenetic analyses. JC and NGK performed glycan analyses. GW carried out light scattering experiments and size determination. AC, ASL, GCH, and ECM supervised and provided funding for the project. All authors read and approved the manuscript.

Competing interests statement

The authors declare no competing interests.

Europe PMC Funders Author Manuscripts

carbohydrate sulfatases are highly selective for the glycan component of their substrate. These data provide new approaches for probing sulfated glycan metabolism, whilst revealing the roles carbohydrate sulfatases play in host-glycan catabolism.

Introduction

Sulfation is a key modification found throughout nature. Sulfate esters have been identified in carbohydrates but can also decorate lipids, proteins, and steroids, drastically altering the biophysical properties of these molecules. Sulfated glycans are essential for multiple aspects of eukaryotic cell biology. The sulfated glycans of the glycosaminoglycan (GAG) class, which include heparan sulfate (HS) and chondroitin sulfate (CS) (Fig. 1a), ubiquitously coat the cells of all mammals, where they regulate extracellular cell signalling, growth, and homeostasis[1,2]. Additionally, the mucus layer coating the colon surface is composed of secreted mucins[2], highly *O*-glycosylated and heavily sulfated glycoproteins (Fig. 1a). Importantly, this mucus layer forms a protective barrier between the gut microbes and the intestinal epithelium preventing close contact and thus, inflammation[3–5]. Sulfated glycans also play critical roles in marine environments. For example, the cell walls of algae, which provide 50 % of all primary carbon fixation on Earth[6,7], contain an abundance of sulfated glycans that include the carrageenans[8], ulvans[9], porphyrans[10], fucoidans[11], and sulfated exopolysaccharides[12], which are essential for cell wall functionality and structure but have also utility as a renewable industrial resource[13]. These sulfated marine glycans also serve as critical bacterial nutrient sources, underpinning food webs and carbon recycling.

The catalytic de-sulfation of glycans by *O*- and *N*-specific carbohydrate sulfatases is a critical process for bacteria that co-inhibit mammalian host and marine environments. For instance, the loss of single sulfatases prevents the complete metabolism of sulfated host glycans such as GAGs[14] and colonic mucin *O*-glycans[3] (cMOs) (Fig. 1a), which are known to be important metabolic targets for members of the human gut microbiota (HGM), the vast microbial community inhabiting the human colon[15,16]. The HGM contributes to multiple biological activities, including protection against pathogens[17] and immunomodulation, whilst providing up to 10% of the hosts energy requirements *via* complex carbohydrate fermentation[18]. Additionally, alterations in HGM composition have been associated with multiple diseases such as cancer, inflammatory bowel disease (IBD) and diabetes. Interestingly, sulfated carbohydrates are important colonisation factors that profoundly influence the composition of the human gut microbiota[14,19,20]. Indeed, a recent study demonstrated that carbohydrate sulfatase activity is a requirement for colonic mucin metabolism and successful competitive colonisation of the gut by at least one HGM *Bacteroides* species[3]. Additionally, increased activity of carbohydrate sulfatases produced by gut microbes are associated with IBD in humans[21,22], and are directly linked to promoting colitis in a susceptible mouse model[23]. Thus, there is a fundamental need to understand the roles that carbohydrate sulfatases play in the metabolism of sulfated host glycans and how these influence biological processes.

Sulfatases are catalogued in the SulfAtlas database into four families, S1-S4, based on sequence homology, and share a conserved overall fold and catalytic mechanism[24]. The

Europe PMC Funders Author Manuscripts

S1 family, of which there are 110 subfamilies (denoted as S1_X), are part of the alkaline-phosphatase-like superfamily and comprise ~90 % of all known sulfatase sequences. This is currently the only family that has been directly implicated in carbohydrate metabolism. The conserved sulfate binding region of S1 sulfatases contains an essential calcium binding site, a His or Lys residue that functions as a catalytic acid, and a core consensus motif (**C/S**-X-P/A-S/X-R) containing a catalytic formylglycine (FGly), which is created by co-translational modification of the Cys/Ser residue[25]. Despite the invariant nature of the sulfate coordination site, wide variability is observed in the carbohydrate binding regions of S1 sulfatases, suggesting that individual subfamilies are tailored to a particular sulfated glycan substrate(s). This is likely a consequence of divergent adaptation to varied sulfated glycan niches and is in keeping with other Carbohydrate Active enZymes (CAZymes), such as glycoside hydrolases, polysaccharide lyases, and glycosyltransferases, which demonstrate exquisite specificity towards their glycan substrates. The invariant sulfate catalytic site is termed the S subsite, whilst the region that accommodates the carbohydrate is designated the 0 subsite; additional sugar binding subsites are denoted with increasing negative numbers toward the non-reducing end (–1, –2, –3, etc.), and increasingly positive numbers (+1, +2, +3, etc) toward the reducing end (free *O*1) of a released oligosaccharide[26].

Bacteroides, belonging to the Bacteroidetes phylum, encode their CAZymes into polysaccharide utilisation loci (PUL)[27], which are sets of genes that are genetically co-localised and co-regulated in response to a particular glycan. As such, prediction of CAZyme functionality and complex glycan target can sometimes be made based on PUL composition. Members of the Bacteroidetes phyla are widespread in both mammalian hosts (including the HGM) and marine environments, where sulfated glycans are common, and serve as excellent model systems to more widely understand sulfatase functionality and evolution. The model HGM organism, *Bacteroides thetaiotaomicron* (*B. theta*), possesses 28 S1 sulfatases distributed across several PULs, of both known and unknown function[3,14,28]. Of the 28 S1 sulfatases identified in the *B. theta* genome, 16 have confirmed carbohydrate sulfatase activity. Four of these enzymes, belonging to S1_9, S1_11, S1_15, and S1_27, reside in PULs for GAG metabolism, and have been extensively characterised[14,28] (Fig. 1b and Supplementary Table 1). In a recent study, we identified the activity of 12 *B. theta* S1 sulfatases in host glycan metabolism, mostly targeting colonic mucin oligosaccharides (cMOs)[3]. This work focused on the detailed structural and biochemical analyses of 3S-galacto-targeting sulfatases from the S1_20 and S1_4 subfamilies, targeting cMOs. Importantly, we also characterized the first activities for the S1_46 and S1_16 subfamilies as 3S,6S-*N*-acetylglucosamine (3,6S-GlcNAc) and 4S-Galactose/*N*-acetylgalactosamine (4S-Gal/GalNAc) sulfatases, respectively. We also demonstrated new activities for S1_15 and S1_11 subfamily members against 6S-Galactose (6S-Gal) and cMOs, respectively (Fig. 1b and Supplementary Table 1). Despite revealing new enzymatic activities amongst these four subfamilies, a detailed understanding of the molecular basis of carbohydrate recognition by HGM sulfatases remains incomplete.

Additionally, only 14 unique carbohydrate sulfatase structures have been described and, only 8 of these have been reported in complex with a ligand. This makes S1 carbohydrate sulfatases the most poorly characterised CAZymes to date. In this study, we describe

the structures of 7 *Bacteroides* S1 carbohydrate sulfatases, 6 of which are complexed with ligands (Extended Data Fig. 1). We also report the first detailed phylogenetic, biochemical, and functional analyses for these sulfatases. The new sulfatase structures span 4 S1 subfamilies and represent the first structural analysis of S1_16 and S1_46 subfamily members. Importantly, structures of sulfatases belonging to the S1_15 and S1_11 subfamilies demonstrate how non-conserved regions of the protein sequence have adapted to direct specificity for closely related sulfated glycan targets. These discoveries provide a new framework for the development of structure guided therapeutic strategies that exploit adaptive substrate specificity by targeting individual sulfatases of the host microbiome. Overall, this study improves our understanding of carbohydrate sulfatase evolution and function.

Results

Conserved features of S1 formylglycine sulfatases

To improve our mechanistic understanding of S1 sulfatase biology, we solved the structures of 7 *B. theta* sulfatases belonging to S1_11, S1_15, S1_16 and S1_46 subfamilies (Fig. 1b). Individual enzymes are subsequently referred to by their gene/locus tag number with the corresponding substrate in superscript (*e.g.*, BT1918$^{3S,6S\text{-}GlcNAc}$). As expected, all enzymes adopt a typical α/β/α fold for the core N-terminal domain, which is abutted by a smaller, C-terminal, 'sub-domain' (Fig. 2a). The sulfate binding site (S site) is invariant for S1 sulfatases and the catalytic FGly residue sits at the base of this pocket in each enzyme. Compositional analysis of the S site amino acid sequence for all available sequences of S1_11, S1_15, S1_16, and S1_46 sulfatases revealed a strong bias towards Cys residues at the critical FGly position, with Cys/Ser ratios of 83:17, 82:18, 85:15, and 84:16, respectively (Fig. 2b). However, *B. theta* sulfatases exclusively possess Ser at the FGly position, which is a common trait associated with organisms inhabiting anaerobic environments.

The S1 sulfatases are calcium-dependent enzymes, and calcium occupies the metal-binding site in all structures, with the exception of BT1918$^{3S,6S\text{-}GlcNAc}$. In the S1_11, S1_15, and S1_16 subfamilies, three Asp residues, a Gln/Asn and the FGly coordinate the calcium in an octahedral configuration completed by an appropriately oriented sulfate from the substrate (Fig. 2c). In BT1918$^{3S,6S\text{-}GlcNAc}$ Q36 and H318 replace the conserved D36 and Q346 (BT3177$^{6S\text{-}GlcNAc}$ numbering), respectively (Fig. 2c). Q292 is positioned further away than the conserved Asp, which is predicted to change the ion's coordination to trigonal bipyramidal and weaken the affinity for calcium. This was previously observed in BT1596$^{2S\text{-}\Delta4,5UA}$, an S1_9 sulfatase, where both Asp and Gln/Asn are replaced by a His residue[14] (Fig. 2c).

S1_46 targets rare host glycosaminoglycan sulfation

Of the S1_46 sequences analysed around ~50 % are derived from mammalian host environments, with the majority isolated from the HGM (Supplementary Data Set 1). BT1918$^{3S,6S\text{-}GlcNAc}$ from *B.theta* is the only S1_46 subfamily member that has been characterised to date and removes sulfate from the *O*3 position of 3S,6S-GlcNAc[3] (Supplementary Table 1). *O*3 sulfation of GlcNAc is an extremely rare modification only

currently known to occur in the GAGs Heparin (Hep) and HS[29] (Fig. 1). Interestingly, the *bt1918*$^{3S,6S\text{-}GlcNAc}$ gene does not reside in the Hep/HS PUL (spanning the genes *bt4652-bt4675*)[14]. Both the HS and CS (*bt3324-bt3334*)[28] PULs are known to utilise distally located 'orphan' genes to breakdown the respective targeted glycans. BT1918$^{3S,6S\text{-}GlcNAc}$ is a dimeric exo-acting enzyme with a pocket topography (Fig. 3a and Supplementary Fig. 1a). We predict that this enzyme is likely to be active in the latter stages of Hep/HS catabolism cleaving the *O*3-linked sulfate and generating the substrate for the 6S-GlcNAc sulfatase, BT4656$^{6S\text{-}GlcNAc/GlcNS}$, which does not act on *O*3 sulfated 6S-GlcNAc[14].

Examining the structure of BT1918$^{3S,6S\text{-}GlcNAc}$, in complex with product, allowed us to undertake an informed mutagenic analysis of the 0 subsite to understand the molecular drivers behind carbohydrate recognition. Of the five residues in the carbohydrate binding region, three make direct contact with the *N*-acetyl group of GlcNAc (Y94, R327, and Y408) (Fig. 3a), and single mutation of each to Ala causes a significant loss in sulfatase activity. However, the mutation of Y94 and Y408 to Phe did not affect activity (Fig. 3b, Supplementary Table 2 and Supplementary Fig. 1b,c). Y94 and R327 are conserved in ~65% of aligned sequences with Y94 present in the motif GRVGYGDE but is replaced by Phe 10% of the time (Extended Data Fig. 2a and Supplementary Fig. 2). These findings suggests a preference for *N*-acetyl configured substrates, driven by stacking interactions, and is consistent with the observed lack of activity of BT1918$^{3S,6S\text{-}GlcNAc}$ towards either 3S- or 3S,6S-glucosamine[3] (Extended Data Fig. 2b). N174 and R143, which flank either side of the β and α faces of the GlcNAc, are conserved in 87% and 76% of S1_46 sequences (Extended Data Fig. 2a). N174 and R143 coordinate the endocyclic ring oxygen and *O*4 of the GlcNAc, respectively, helping to orientate the *O*3 sulfate for catalysis (Fig. 3a). Interestingly, N174 and the residues interacting with the *N*-acetyl group (Y94, R327, and Y408) are invariant in S1_46 sulfatases of Bacteroidetes within the HGM (Supplementary Data Set 1). This suggests that interactions with the *N*-acetyl group are a major specificity determinant in HGM S1_46 sulfatases. S1_46 sequences are also found in some pathobionts. Interestingly, only a single strain of *Clostridioides difficile*, the non-pathogenic *P28*, contains an S1_46 sulfatase, QSI_2516 (locus tag). This S1_46 sequence contains equivalent residues to Y94, N174, and R327 of BT1918$^{3S,6S\text{-}GlcNAc}$ but not Y408. This enzyme has the same substrate specificity as BT1918$^{3S,6S\text{-}GlcNAc}$ albeit with ~2000 fold lower activity, and is the only S1 sulfatase found in *C. difficile P28* (Extended Data Fig. 2b and Table S2). Moreover, Y408 is only found in 26% of S1_46 sequences (within a conserved HATCY motif), all of which were identified in Gram-negative bacteria. The reasons for this phylum distribution of Y408 are unclear.

Aromatic stacking is key for S1_16 sulfatases of the HGM

The *B.theta* S1_16 sulfatases, BT3057$^{4S\text{-}Gal/GalNAc}$ and BT3796$^{4S\text{-}Gal/GalNAc}$, display both monomeric and dimeric species *in vitro*. Although oligomerisation has been shown to be important for some sulfatases[30,31], this is likely a consequence of heterologous expression (Extended Data Fig. 3a-c and Supplementary Fig. 3). BT3057$^{4S\text{-}Gal/GalNAc}$ and BT3796$^{4S\text{-}Gal/GalNAc}$ are active against 4S-Gal/GalNAc[3]. BT3796$^{4S\text{-}Gal/GalNAc}$ is part of a small PUL that also contains a putative GH36 α-D-galactosidase, a putative GH29 α-L-fucosidase, and an S1_27 sulfatase, known from the literature as 4S-GalNAc or ulvan

sulfatases (Fig. 1b)[9,28]. In contrast, BT3057$^{4S\text{-}Gal/GalNAc}$ is an orphan gene, which can contribute additional sulfatase capacity to an existing PUL. Structural analysis revealed that both sulfatases utilise a critical aromatic residue, W109, which stacks against the α face of the sugar ring. *O*3 is coordinated via Nε2 of H423 in BT3057$^{4S\text{-}Gal/GalNAc}$, and the indole nitrogen of W431 in BT3796$^{4S\text{-}Gal/GalNAc}$ and, in BT3057$^{4S\text{-}Gal/GalNAc}$, an additional interaction is made between H182 and *O*6 (Fig. 4a,b). The importance of the W109 position for the catalytic activity of both enzymes was confirmed by mutating this position to Ala, which resulted in near complete abrogation of sulfatase activity (Fig. 4c, Supplementary Table 2 and Extended Data Fig. 3d,e). Mutation of other glycan-coordinating residues caused comparatively minor reductions in activity. The structures of both enzymes suggest they could potentially accommodate the binding of a +1 sugar, with H182 (BT3057$^{4S\text{-}Gal/GalNAc}$) and W332 (BT3796$^{4S\text{-}Gal/GalNAc}$) residues providing stacking interactions (Fig. 4a,b). Bioinformatic analysis shows that 72% of the aligned S1_16 sequences are derived from marine or aquatic environments, with sequences from a terrestrial/soil or mammalian environments making up most of the remainder, in roughly equal proportion (Supplementary Data Set 2). Interestingly, despite an abundance of S1_16 sulfatases in marine environments, and the prevalence of 4S-Gal in sulfated marine glycans (such as iota and kappa-carrageenan), only marine sulfatases from S1_19 have previously been demonstrated activity towards these substrates[8,26]. Moreover, W109 is only conserved in 37% of analysed S1_16 sequences. However, this residue is found in 84% of sequences from organisms residing in the human gut, which decreases to 17% for sequences from bacteria in marine or aquatic environments, where a hydrophobic Val is usually observed (Extended Data Fig. 4, Supplementary Fig. 4 and Supplementary Data Set 2). Furthermore, Phe and Tyr replace W109 at a frequency of ~8% and ~6%, respectively, and could theoretically perform the same stacking roles (Extended Data Fig. 4). These data suggest that only the subset of S1_16 sulfatases, possessing an equivalent residue to W109, are active as 4S-Gal/GalNAc sulfatases and that this activity is mostly restricted to human gutbacteria. S1_16 sulfatases from marine environments likely target a different, but potentially analogous, sulfated linkage.

Defining the features driving 6S-Gal/GalNAc recognition

All 4 S1_15 family members analysed here are monomeric in solution (Supplementary Fig. 5a) and target 6S-D-galacto-configured substrates, but exhibited marked substrate variability. BT3333$^{6S\text{-}GalNAc}$ and BT3109$^{6S\text{-}Gal}$ preferentially target 6S-D-*N*-acetylgalatcosamine (6S-GalNAc) and 6S-Gal respectively, whilst BT1624$^{6S\text{-}Gal/GalNAc}$ and BT4631$^{6S\text{-}Gal/GalNAc}$ cleave both 6S-GalNAc and 6S-Gal equally well (Supplementary Table 3). The genes encoding BT1624$^{6S\text{-}Gal/GalNAc}$ and BT3333$^{6S\text{-}GalNAc}$ reside in PULs associated with mucin *O*-glycan and CS metabolism respectively, whereas BT3109$^{6S\text{-}Gal}$ and BT4631$^{6S\text{-}Gal/GalNAc}$ reside in PULs with unknown glycan targets (Fig. 5a). BT3333$^{6S\text{-}GalNAc}$ was previously shown to use a galacto-recognition triad to bind GalNAc with His223 coordinating *O*3, whilst D173 and R174 hydrogen bond to *O*4[28] (Fig. 5b). The importance of these residues was confirmed through mutational analysis of BT1624$^{6S\text{-}Gal/GalNAc}$ (Extended Data Fig. 5). This configuration of amino acids was also conserved in the additional three sulfatases analysed here demonstrating similar modes of substrate recognition (Fig. 5b). However, alignments of the S1_15 subfamily shows that the complete galacto-recognition triad is

only conserved in 56% (523) of all sequences analysed. Further analyses reveal that ~65% of S1_15 sequences are derived from a marine environment and ~ 70% of these lack the galacto-recognition triad (Extended Data Fig. 6, Supplementary Fig. 6 and Supplementary Data Set 3). By contrast, 83%, 80% and 95% of sequences from human, animal, and terrestrial sources, respectively, contain the galacto-recognition triad (Supplementary Data Set 3). This suggests that S1_15 subfamily members target at least two types of sulfated substrates, a galacto-configured glycan and an unknown sulfated substrate enriched in the marine environment.

Preference for 6S-Gal or 6S-GalNAc is primarily driven by the openness of S and 0 subsites (Fig. 5c). BT4631$^{6S\text{-}Gal/GalNAc}$ has more open S and 0 subsites, which leaves the target sulfate solvent exposed and may permit the accommodation of additional carbohydrate groups. An Ile situated over the S subsite in the other 3 structures buries the sulfate group. Additional specificity is driven by residues interacting with C2 substituents of the substrate (Fig. 5c). BT1624$^{6S\text{-}Gal/GalNAc}$ lacks interactions with C2 substituents and has a more open pocket than BT3333$^{6S\text{-}GalNAc}$ (Fig. 5b, c). This results in BT1624$^{6S\text{-}Gal/GalNAc}$ showing approximately equal affinity and catalytic activity towards both 6S-Gal and 6S-GalNAc[3] (Supplementary Table 3). In contrast, BT3333$^{6S\text{-}GalNAc}$ has an aromatic residue, W464, which is orientated through a stacking interaction with Y463, to interact with the *N*-acetyl group of GalNAc (Fig. 5b), driving specificity for 6S-GalNAc (Supplementary Table 3)[3]. Mutation of W464 to Ala, or deletion of Y463/W464, significantly reduces the preference of BT3333$^{6S\text{-}GalNAc}$ for GalNAc, whilst slightly increasing its interaction with Gal (Fig. 5d). Interestingly, a residue analogous to W464 is also present in BT4631$^{6S\text{-}Gal/GalNAc}$ (W469), but the enzyme differs from BT3333$^{6S\text{-}GalNAc}$ in lacking a Y463 equivalent, which is instead replaced by T464 (Fig. 5b). In the absence of a positioning aromatic partner, W469 instead flips down and is unable to interact with the *N*-acetyl group of the 0 subsite sugar. Consistently, analysis of the sequences of S1_15 enzymes located within *Bacteroides* PULs that target CS reveals a conservation of W464. However, Y463 is only partially retained, often being substituted by Phe and His, which could also stack against W464, preserving enhanced specificity towards GalNAc substrates (Extended Data Fig. 7).

Analysis of the BT3109$^{6S\text{-}Gal}$ substrate bound structure provides a mechanistic explanation for its preference towards 6S-Gal. BT3109$^{6S\text{-}Gal}$ forms an interaction with the *O*2 of Gal via the carboxy terminus of K508 (Fig. 5b, c). This occludes binding of the *N*-acetyl group of GalNAc, resulting in a strong preference for 6S-Gal (Fig. 5e and Supplementary Table 3). To further confirm this finding, we generated a truncation mutant (BT3109$^{\Delta CT}$) by removing the 7 amino acids of the carboxy terminus (VEEEPLK), a structural element absent in the other 3 S1_15 structures analysed. As predicted, BT3109$^{\Delta CT}$ did not show a preference for Gal and bound both Gal and GalNAc with a similar affinity (Fig. 5e). Interestingly, the C-terminal VEEEPLK sequence is only found in 50 of the 920 representative S1_15 sequences analysed, and almost exclusively in Bacteroidetes inhabiting marine environments (Supplementary Table 4).

Comparison of S1_11 sulfatases targeting host glycans

BT3177$^{6S\text{-}GlcNAc}$ and BT4656$^{6S\text{-}GlcNAc/GlcNS}$ both exist in solution as monomeric proteins (Supplementary Fig. 5b) capable of de-sulfating 6S-D-*N*-acetylglucosamine (6S-GlcNAc) and 2*N*-sulfated 6S-D-glucosamine (6S-GlcNS). Their genes reside in PULs that process cMOs and Hep/HS substrates, respectively (Fig. 6a). Both enzymes utilise an invariant substrate recognition triad composed of Asp, Arg, and His, where Asp/Arg coordinate with the *O*4 of the substrate and His coordinates with the *O*3 (Fig. 6b,c). The key role of these residues in substrate recognition was confirmed by mutation to alanine of D361, R363 and H445 in BT3177$^{6S\text{-}GlcNAc}$ (Extended Data Fig 8 and Supplementary Table 3). Interestingly, these residues are found within conserved motifs present in 91 % of all 955 representative S1_11 sequences analysed [with an Asp and Arg dyad conserved in 98 and 99% of sequences, respectively (Extended Data Fig. 9, Supplementary Fig. 7 and Supplementary Data Set 4)].

The substrates targeted by these sulfatases have vastly different carbohydrate structures (Fig. 1). cMOs contain 6S-GlcNAc as one of their major glycan components[3] whilst Hep/HS can contain both 6S-GlcNAc and 6S-GlcNS[14]. Despite the invariant nature of the substrate recognition triad, the region that coordinates either the *N*-acetyl or *N*-sulfate group in both proteins displays high amino acid variance across S1_11 subfamily members and is entirely absent in around half of the S1_11 sequences analysed (Extended Data Fig. 9 and Supplementary Data Set 4). In BT3177$^{6S\text{-}GlcNAc}$, Y250 interacts with the *N*-acetyl group of 6S-GlcNAc, and sits in a hydrophobic region flanked by L263 and L266 (Fig. 6b). Mutation of Y250 to Ala reduced activity by ~10-fold, whilst substitution to Phe had no effect, demonstrating that the hydrophobic character of the aromatic phenyl ring is key for optimal interaction with the *N*-acetyl group (Supplementary Table 3). By contrast, BT4656$^{6S\text{-}GlcNAc/GlcNS}$ possesses W273 and R290 in this region, and also lacks the L263 equivalent, resulting in a more basic, open topography (Fig. 6c). R290 sits above W273 interacting through cation-π interactions, positioning R290 to form a bidentate ionic interaction with the *N*-sulfate group, whilst the Nε1 of the Trp indole ring forms a hydrogen bond with the third oxygen of the sulfate. Despite the high variability found in this region (Extended Data Fig. 9 and Supplementary Data Set 4), homologues of BT4656$^{6S\text{-}GlcNAc/GlcNS}$ harbouring W273/R290 are well conserved in PULs specifically targeting the 6S-GlcNS enriched substrates Hep/HS (26 of 28 PULs analysed with R290Q observed in 2 instances) (Extended Data Fig. 10). BT4656$^{6S\text{-}GlcNAc/GlcNS}$ is also active on 6S-GlcNAc (also found in Hep/HS) however, the presence of R290/W273 suggests an enhanced affinity towards 6S-GlcNS. Indeed, although the k_{cat}/K_M are similar for both substrates (Supplementary Table 3), thermostability analysis confirms that BT4656$^{6S\text{-}GlcNAc/GlcNS}$ has a greater affinity for GlcNS than GlcNAc. The reciprocal is true for BT3177$^{6S\text{-}GlcNAc}$, which resides in a PUL targeting cMOs[3] (Fig. 6d). These results suggest that these sulfatases evolved to specifically recognize the target substrate of the PULs in which they are encoded.

Despite 6S-GlcNAc being a common component of host glycans, only ~15% of S1_11 sulfatases are found in mammalian host environments. The majority of S1_11 sequences isolated (~50%) are of a marine origin (Supplementary Data Set 4). Substrates of a

6S-gluco configuration are rare in marine environments. But it has recently been shown that two marine S1_11 sulfatases utilise the conserved recognition triad to recognise 6S-L-galactose[32], a common component of the algal galactan porphyran, rather than 6S-D-GlcNAc. Rotation of the 6S-L-galactose by ~90°, relative to 6S-D-GlcNAc, in BT3177$^{6S\text{-}GlcNAc}$ and BT4656$^{6S\text{-}GlcNAc/GlcNS}$, allows coordination of *O*3 and *O*4 with the critical recognition triad. The region interacting with *C*2 substituents, found in S1_11 HGM sulfatases analysed here, is absent in these enzymes. Interestingly, ~27% of S1_11 sequences are derived from terrestrial/soil based environments, representing a 2-3 fold enrichment when compared to S1_15, S1_16, and S1_46 (Supplementary Data Set 4).

Discussion

CAZymes are exquisitely specific enzymes, capable of distinguishing between single epimeric features thus driving specificity towards individual carbohydrates. S1 carbohydrate sulfatases are no exception to this. For example, S1_11 and S1_15 subfamilies utilise an identical recognition triad but specificity for GlcNAc or GalNAc (epimeric at *O*4) is imparted by residues coming from the C-terminus in S1_11 and the N-terminus in S1_15, meaning no interactions are spatially conserved and thus recognition is specific. Inactive forms of BT4656$^{6S\text{-}GlcNAc/GlcNS}$ (S1_11 subfamily) and BT3333$^{6S\text{-}GalNAc}$ (S1_15 subfamily) have similar binding energies, for their respective substrates. Yet, BT3333$^{6S\text{-}GalNAc}$ had a specific activity >100 lower than BT4656$^{6S\text{-}GlcNAc/GlcNS}$. In addition, there was no significant difference in the reactivity of the 6S-GlcNAc and 6S-GalNAc substrates under acid hydrolysis conditions. This suggests that the difference in rate may be due to the ability of the *E.coli* expression host to correctly insert the catalytic FGly residue into S1_15 subfamily members at levels comparable to other subfamilies characterised. (Supplementary Fig. 7a-c, and Supplementary Table 5) for further comparisons). The finer details of sulfatase substrate recognition occur through features in the highly variable regions of the proteins. For example, S1 sulfatases such as BT3177$^{6S\text{-}GlcNAc}$ and BT3333$^{6S\text{-}GalNAc}$ have evolved aromatic residues capable of interacting with the *N*-acetyl group of their respective target substrates. In contrast, the same region of BT1624$^{6S\text{-}Gal/GalNAc}$ shows no tailored adaptation to *C*2 substituents, exhibiting similar activity towards both 6S-Gal and 6S-GalNAc. Finally, BT4656$^{6S\text{-}GlcNAc/GlcNS}$ possesses an equivalent area comprising more positively charged residues, which provides a stronger interaction with the doubly sulfated 6S-GlcNS, a component found exclusively in the GAGs Hep and HS, the substrates specifically targeted by this sulfatases PUL. Interestingly, GAGs have been shown to be high priority substrates for several *Bacteroides* species[16] and the enhanced activities of sulfatases targeting host glycans may bestow a critical advantage on these substrates within the competitive gut environment. Interestingly, the S1_11 and S1_15 subfamilies are more closely related to the mainly eukaryotic subfamilies, S1_6 (6S-GlcNAc activity) and S1_5 (6S-GalNAc activity), respectively, than to each other. Despite this similarity in substrate specificity, and evolutionary relatedness, analysis of select mammalian sequences from S1_6 and S1_5 suggests only lysosomal sulfatases from the S1_6 subfamily contain any residues from the critical recognition triads identified in the S1_11 and S1_15 subfamilies (Supplementary Fig. 7d,e). Exo-acting lysosomal S1_11 enzymes from *Homo sapiens*, *Pan troglodytes* (Chimpanzee), *Mus*

musculus (House mouse), and *Rattus norvegicus* (Brown rat) retain the Asp and Arg dyad, which coordinates *O*4. In the endo-acting S1_6 enzymes SULF1 and SULF2 the Asp and Arg of the dyad are replaced by Gly and Ser, respectively. These observations are in line with the endo-acting nature of the SULF enzymes. They remove *O*6 sulfation from GlcNAc in the linear polysaccharide heparan sulfate, which is linked by β1,4 and α1,4 glycosidic bonds. This would make the *O*4 hydroxyl unavailable for coordination by an Asp/Arg dyad as it is involved in a glycosidic linkage and room must be left to accommodate additional sugar residues.

The S1_15 sulfatase BT3109$^{6S\text{-}Gal}$ is found in a PUL of unknown function that also contains GH2 and GH43_31 enzymes, both of which target Gal in the pyranose[33] and furanose[34] form, respectively. Since BT3109$^{6S\text{-}Gal}$ orthologues are most commonly found in marine environments, we propose that this PUL may target marine polysaccharides containing 6S-Gal, such as λ-carrageenan[35]. Indeed, the ability of the HGM to metabolise marine glycans was recently revealed to be more extensive than previously thought[36], suggesting they may be important resources for some bacterial species within the colonic ecosystem.

The S1_16 family is most common in marine environments but these sequences mostly lack the critical aromatic stacking interactions required for 4S-Gal/GalNAc activity within HGM sulfatases. This suggests that there is significant divergence in S1_16 function between the two environmental niches. Furthermore, only marine sulfatases of the S1_19 subfamily have previously demonstrated to have 4S-Gal activity[8,26], which is independent of aromatic stacking at the 0 subsite or recognition of the *O*3 hydroxyl group. This indicates that S1_16 and S1_19 subfamilies have evolved different strategies to deal with 4S de-sulfation of Gal in the structurally distinct glycans found in mammalian and marine environments.

Sequences of S1_46 subfamily sulfatases from both Firmicutes and Bacteroidetes inhabiting the human gastrointestinal tract largely retain the *N*-acetyl group recognition features. All of the key residues are strongly conserved with the exception of Y408 which is preserved in Bacteroidetes but is mostly absent in gut Firmicutes. These conserved recognition features are also found in several pathobionts, as well as a single strain of non-pathogenic *C. difficile*, which are considered to prevent infection by pathogenic strains[37]. S1_46 sulfatases could serve to allow access to host glycan breakdown products by these species thereby increasing their fitness. S1_46 sequences from marine environments exhibit greater variance in the residues that recognise *N*-acetyl groups. This could suggest that carbohydrates modified by this moiety are rarely encountered in aquatic microbiomes. Interestingly, some sulfatases found in the gastrointestinal tract also lack the *N*-acetyl group recognition features in this region, and may target a similar, as of yet unidentified, substrate.

The two S1_11 *B. theta* sulfatases studied here demonstrated it was the non-conserved region of the sulfatases that drove specificity for differing host glycans. Interestingly, this subfamily was found enriched in terrestrial/soil based environments. We propose that this intriguing finding may be due to the presence of 6S-GlcNAc in nodulation (NOD) factors. These bacterially produced lipooligosaccharides are essential for bacterial invasion of plant roots as a precursor to establishment of nitrogen fixing nodules[38]. Indeed, several bacterial S1_11 sequences were isolated from rhizosphere communities and root nodules suggesting

that S1_11 sulfatases could be involved in NOD factor metabolism. Additionally, S1_11 sequences found within the fungal phylum Ascomyota include notable plant pathogens. This alludes to a potential dichotomy, where plant associated bacterial species may utilise S1_11 sulfatases to de-sulfate NOD factors during nodule formation, whilst S1_11 sequences in fungal Ascoymota may act as virulence factors, potentially by hijacking the NOD factor system.

S1 carbohydrate sulfatases are exquisitely specific enzymes deriving binding energy and specificity from the nature of the glycan to which the target sulfate is appended. HGM S1_46 subfamily members utilise the *N*-acetyl group as an absolute specificity determinant, with these features conserved in HGM commensals and pathobionts. In contrast, S1_11 and S1_15 subfamilies utilise the *N*-acetyl group as a specificity modifier with absolute specificity being targeted toward the carbohydrate's unique epimeric features. S1_16 members of the HGM utilise a critical aromatic stacking interaction, and recognition of O3, to target 4S-Gal/GalNAc, but these features are not well conserved in the marine environment suggesting a divergence in 4S-Gal desulfation activity. Significant divergence of the key recognition features is also observed for S1_46 and S1_15 sequences from the marine environment, but not for S1_11. The specificity of S1 carbohydrate sulfatases makes these enzymes excellent targets for small molecule intervention to modify their function, develop tools to probe sulfated glycan metabolism, and for disease intervention where sulfated glycan metabolism is perturbed.

Methods

Recombinant Protein Production

Genes were amplified by PCR using the appropriate primers and the amplified DNA cloned into pET28b using NheI/XhoI restriction sites generating constructs with N-terminal His_6 tags (Supplementary Table 6). Recombinant genes were expressed in *Escherichia coli* (*E.coli*) strains BL21 (DE3) or TUNER (Novagen), containing the appropriate recombinant plasmid, and cultured to mid-exponential phase in Luria Broth (LB) supplemented with 50 μg/mL kanamycin at 37 °C and 180 rpm. Cells were then cooled to 16°C, and recombinant gene expression was induced by the addition of 0.1 mM isopropyl β-D-1-thiogalactopyranoside; cells were cultured for another 16 h at 16°C and 180 rpm. The cells were then centrifuged at 5,000 × g and resuspended in 20 mM HEPES ((4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid), pH 7.4, with 500 mM NaCl before being sonicated on ice. Recombinant protein was then purified by immobilized metal ion affinity chromatography using a cobalt-based matrix (Talon, Clontech) and eluted with 100 mM imidazole. For the proteins selected for structural studies, another step of size exclusion chromatography was performed using a Superdex 16/60 S75 or S200 column (GE Healthcare), with 10 mM HEPES, pH 7.5, and 150 mM NaCl as the eluent, and they were judged to be ≥95% pure by SDS-PAGE. Protein concentrations were determined by measuring absorbance at 280 nm using the molar extinction coefficient calculated by ProtParam on the ExPasy server (web.expasy.org/protparam/).

Site-Directed Mutagenesis

Site-directed mutagenesis was conducted using the PCR-based QuikChange kit (Stratagene) according to the manufacturer's instructions using the appropriate plasmid as the template and appropriate primer pairs (Supplementary Table 6).

Glycan labelling

Sulfated saccharide samples were labelled according to a modification of the method by Das *et al.*, reporting the formation of N-glycosyl amines for 4,6-O-benzilidene protected D-gluopyranose monosaccharides with aromatic amines[39]. Briefly, the lyophilised sugar (1 mg) was dissolved in anhydrous methanol (0.50 mL, Sigma-Aldrich) in a 1.5 mL screw-top PTFE microcentrifuge tube and BODIPY-FL hydrazide (4,4-difluoro-5,7-dimethyl-4-bora-3a,4a-diaza-*s*-indacene-3-propionic acid, hydrazide, 0.1 mg, ThermoFischer, $\lambda_{ex./em.}$ 493/503, ε 80,000 $M^{-1}cm^{-1}$) was added and the mixture vortexed (1 min), then reacted (65°C, 24 h) in darkness. The products were then cooled and a portion purified by TLC on silica coated aluminium plates (silica gel 60, Sigma-Aldrich, Millipore) developed with methanol or 1:1 v/v ethyl acetate/methanol. The unreacted BODIPY-FL label (orange on the TLC plate) was identified by reference to a lane containing the starting material (BODIPY-FL hydrazide), allowing differentiation from the putative labelled product (also orange). This latter band (which can migrate ahead or behind the label depending on the sugar; e.g. labelled 4S and 6S GlcNAc both migrate with R_f 0.84 compared to label, R_f 0.70; others, such as labelled 4S or 6S GalNAc require, 1:1 v/v ethyl acetate/methanol and the product runs behind the label on TLC) was scraped from the plates and extracted in fresh methanol (2 x 0.5 mL), spun (benchtop centrifuge, 3 minutes), the supernatant recovered and dried (rotary evaporator) to afford the fluorescent, coloured product (bright green in aqueous solution), which was then employed in subsequent experiments.

Microfluidics de-sulfation assays

Sulfated carbohydrates were labelled at their reducing end with BODIPY-FL which has a maximal emission absorbance of ~503nm, which can be detected by the EZ Reader via LED-induced fluorescence[40]. Non-radioactive microfluidic mobility shift carbohydrate sulfation assays were optimised in solution with a 12-sipper chip coated with CR8 reagent and a PerkinElmer EZ Reader II system using EDTA-based separation buffer and real-time kinetic evaluation of substrate de-sulfation. Pressure and voltage settings were adjusted manually (1.8 psi, upstream voltage: 2250 V, downstream voltage: 500 V) to afford optimal separation of the sulfated and unsulfated product with a sample (sip) time of 0.2 s, and total assay times appropriate for the experiment. Individual de-sulfation assays were carried out at 28°C and assembled in a 384-well plate in a volume of 80 μl in the presence of substrate concentrations between 0.5 and 20 μM with 100 mM Bis-Tris-Propane or Tris, depending on the pH required, 150 mM NaCl, 0.02% (v/v) Brij-35 and 5 mM $CaCl_2$. The degree of de-sulfation was directly calculated using the EZ Reader software by measuring the sulfated:unsulfated carbohydrate ratio at each time-point. The activity of sulfatase enzymes was quantified in 'kinetic mode' by monitoring the amount of unsulfated glycan generated over the assay time, relative to control assay with no enzyme; with sulfate loss limited to ~20% to prevent of substrate depletion and to ensure assay linearity. k_{cat}/K_M values,

using the equation $V_0 = (k_{cat}/K_m)[E][S]$, were determined by linear regression analysis with GraphPad Prism software. Substrate concentrations were halved and doubled to assess linearity of the reaction rates to ensure substrate concentrations were significantly $<K_M$.

HPAEC and TLC sulfatase enzymatic assays

For reactions analysed by thin layer chromatography (TLC) 2 μL of each sample was spotted onto silica plates and resolved in butanol:acetic acid:water (2:1:1) running buffer. The TLC plates were dried, and the sugars were visualized using diphenylamine stain (1 ml of 37.5% HCl, 2 ml of aniline, 10 ml of 85% H_3PO_3, 100 ml of ethyl acetate and 2 g diphenylamine) and heated at 450°C for 2-5 min with a heat gun. Where possible, the enzymatic activity was confirmed by high-performance anionic exchange chromatography (HPAEC) with pulsed amperometric detection using standard methodology[41]. The sugars (reaction products) were bound to a Dionex CarboPac PA200 column and eluted with an isocratic flow of 80 mM NaOH for 15 min, the column was then cleaned with 500 mM NaOH for 10 min before being ran back into 80 mM NaOH at a flow rate of 0.25 ml min^{-1} before injection of the next sample. The reaction products were identified using appropriate standards.

Differential scanning fluorimetry

Thermal shift/stability assays (TSAs) were performed using a StepOnePlus Real-Time PCR machine (LifeTechnologies) and SYPRO-Orange dye (emission maximum 570 nm, Invitrogen) as previously described[42] with thermal ramping between 20 and 95°C in 0.3°C step intervals per data point to induce denaturation in the presence or absence of various carbohydrates as appropriate for the sulfatase being analysed. The melting temperature (Tm) corresponding to the midpoint for the protein unfolding transition was calculated by fitting the sigmoidal melt curve to the Boltzmann equation using GraphPad Prism, with R^2 values of >0.99. Data points after the fluorescence intensity maximum were excluded from the fitting. Changes in the unfolding transition temperature compared with the control curve (ΔT_m) were calculated for each ligand. A positive ΔT_m value indicates that the ligand stabilises the protein from thermal denaturation, and confirms binding to the protein. All TSA experiments were conducted using a final protein concentration of 5μM in 100 mM Bis-Tris-Propane (BTP), pH 7.0, and 150 mM NaCl supplemented with the appropriate ligand. Three independent assays were performed for each protein and protein ligand combination (Supplementary Table 7 and 8).

Isothermal titration calorimetry (ITC)

The affinity of $BT3333^{6S\text{-}GalNAc}$ and $BT4656^{6S\text{-}GlcNAc/GlcNS}$ against 6S-GalNAc and 6S-GlcNAc, respectively, was quantified by ITC using a Microcal ITC^{200} calorimeter. The protein samples (70 μM for $BT3333^{6S\text{-}GalNAc}$ and 60 μM for $BT4656^{6S\text{-}GlcNAc/GlcNS}$), stirred at 400 rpm in a 0.2-mL reaction cell, was injected 18 times with 2 μL aliquots of ligand, preceded by 1 injection of 0.2 μL with a delay of 180 seconds between injections (0.8 mM 6S-GalNAc was used for $BT3333^{6S\text{-}GalNAc}$ and 0.4 mM 6S-GlcNAc for $BT4656^{6S\text{-}GlcNAc/GlcNS}$). Titrations were carried out in 50 mM Tris-HCl buffer, pH 8.0, at 25 °C. Integrated binding heats minus dilution heat controls were fit to a single set of sites binding model to derive K_A, ΔH, and n (number of binding sites on each molecule of protein) using Microcal Origin v7.0.

Light scattering and determination of molecular weight

Molecular masses were determined using an Agilent Multi-Detector System calibrated with bovine serum albumin. Proteins were separated by size exclusion chromatography using an Agilent BioSEC Advance 300 Å, 4.6 x 300 mm or GE Superdex 200 10 300 columns equilibrated with 20 mM tris(hydroxymethyl)aminomethane-HCl pH 7.4, 150 mM NaCl buffer. Light scattering data was collected at 90° and refractive index used to calculate absolute molecular mass.

Crystallisation of carbohydrate sulfatases

After purification, all proteins were concentrated in centrifugal concentrator with a molecular weight cutoff of 30 KDa in the size exclusion chromatography buffer. Sparse matrix screens were set up in 96-well sitting drop SPT Labtech plates plates (400-nL drops). Initial hits crystals for all proteins were obtained between 20 and 35 mg/mL supplemented with between 10 and 30 mM ligand unless otherwise stated. For all sulfatase the wildtype *B. theta* variants were used, having a Ser at the catalytic formylglycine position. BT1624$^{6S\text{-}Gal/GalNAc}$ with 6S-GalNAc crystallised in 20% Polyethyleneglycol (PEG) 6000, 0.2 M ammonium chloride and 0.1 M sodium acetate pH 5.5. BT1918$^{3S\text{-}GlcNAc}$ with 6S-GlcNAc crystallised in 45% Methylpentanediol (MPD) 0.2 M $CaCl_2$ and 0.1 M Bis-Tris pH 5.5. BT3057$^{4S\text{-}Gal/GalNAc}$ with 4S-GalNAc crystallised in 20% PEG 3350 and 0.2 M sodium nitrate. BT3109$^{6S\text{-}Gal}$ with 6S-Gal crystallised in 30% PEG 4000, 0.2 M ammonium acetate and sodium citrate pH 5.6. BT3177$^{6S\text{-}GlcNAc}$ with 6S-GlcNAc crystallised in 50 % precipitant mix 1 (40% v/v PEG 500 MME; 20% w/v PEG 20000), 0.1 M carboxcylic acids (0.2 M Sodium formate; 0.2 M Ammonium acetate; 0.2 M Sodium citrate tribasic dihydrate; 0.2 M Sodium potassium tartrate tetrahydrate; 0.2 M Sodium oxamate) and 0.1 M buffer system 3 pH 8.5 (Tris (base); BICINE). BT3796$^{4S\text{-}Gal/GalNAc}$ with 4S-GalNAc crystallised in 20% PEG 6000, 0.2 M magnesium chloride and 0.1 MES pH 6.0. BT4631$^{6S\text{-}Gal/GalNAc}$ was crystallised at 80 mg/ml with 10 mM 6S-Gal in 20% PEG 10000 with 0.1 M Bicine pH 8.5. All crystals were cryo-cooled with the addition of the ligand they were crystallised with. 20% PEG 400 was used as the cryoprotectant for BT1624$^{6S\text{-}Gal/GalNAc}$, BT3057$^{4S\text{-}Gal/GalNAc}$, and BT3109$^{6S\text{-}Gal}$ and 100% paratone-N oil for BT3796$^{4S\text{-}Gal/GalNAc}$. PEG 200 was used as cryoprotectant for BT4631$^{6S\text{-}Gal/GalNAc}$ crystals. No cryoprotectant was added to BT1918$^{3S\text{-}GlcNAc}$ or BT3177$^{6S\text{-}GlcNAc}$ crystals as the crystallisation condition afforded sufficient cryoprotection. Data were collected at Diamond Light Source (Oxford) on beamlines I03, I04, I04-1 and I24, and SOLEIL on the PROXIMA_1 beamline at 100 K. The data were integrated with XDS[43], or Xia2 3di or 3dii, and scaled with Aimless[44,45]. Five percent of observations were randomly selected for the R_{free} set. The phase problem was solved by molecular replacement using the automated molecular replacement server Balbes[46] for all proteins except BT3109$^{6S\text{-}Gal}$ and BT4631$^{6S\text{-}Gal/GalNAc}$, which were solved using Phaser[47] and BT1624$^{6S\text{-}Gal/GalNAc}$ as the search model after preparation with sculptor. Models underwent recursive cycles of model building in Coot[48] and refinement cycles in Refmac5[49]. Where necessary ligand restraint and coordinates were generated with Jligand[50]. The models were validated using Coot and MolProbity[51]. Structural figures were made using Pymol (The PyMOL Molecular graphics system, Version 2.0 Schrodinger, LLC.) and all other programs used were from the CCP4 suite[52,53]. The data processing and refinement statistics are reported in Supplementary Tables 9 and 10.

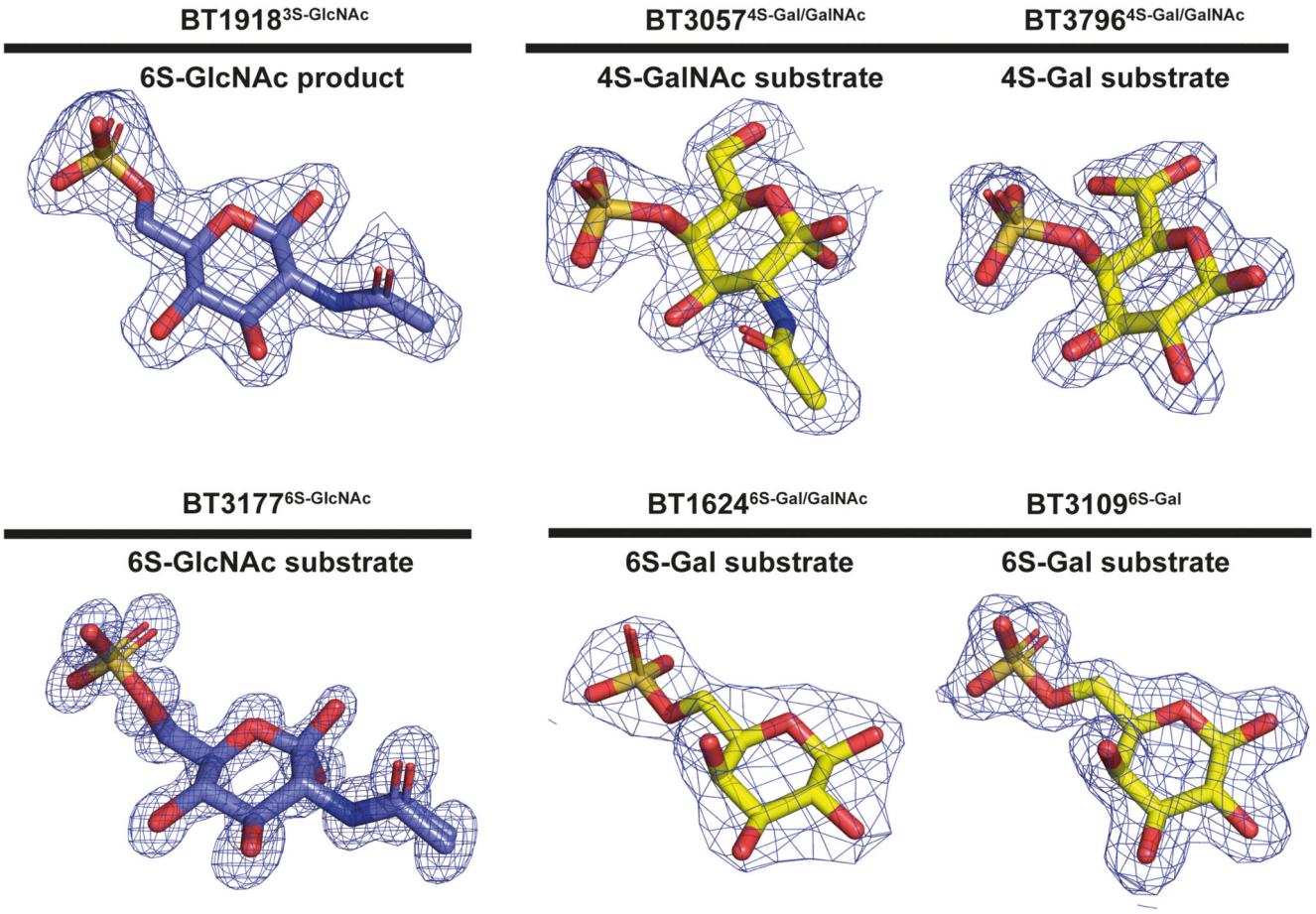
Europe PMC Funders Author Manuscripts

Global phylogenetic trees of S1_11, S1_15, S1_16, S1_46 sequences

On the basis of the taxonomic diversity, to avoid identical sequences we selected a representative number of sequences within each subfamily: a) for S1_11, 955 sequences were selected among the 2177 sequences present in this subfamily in the SulfAtlas database and 411 positions were used for phylogeny; b) for S1_15, 920 sequences were selected among the 1906 sequences present in SulfAtlas and 365 positions were used for phylogeny; c) for S1_16, 800 sequences were selected among the 1361 sequences present in SulfAtlas and 342 positions were used for phylogeny; and d) for S1_46, 349 out of the total 574 sequences present in SulfAtlas were selected, 401 positions were used for phylogeny. In each case, the sequences were aligned by MAFFT v.7[54] using L-INS-i algorithm. The multiple sequence alignments were visualized by Jalview software v.11.0[55], non-aligned regions were removed, and the above listed respective numbers of positions were used for the phylogeny. Phylogeny was made using RAxML v. 8.2.4[56]. The phylogenetic tree was build with the Maximum Likelihood method[57] and the LG matrix as evolutive model[58] using a discrete Gamma distribution to model evolutionary rate differences among sites (4 categories). The rate variation model allowed for some sites to be evolutionarily invariable. The reliability of the trees was tested by bootstrap analysis using 1000 resamplings of the dataset[59]. In all cases, fifteen S1_0 sequences from the SulfAtlas database were used as outgroup.

Europe PMC Funders Author Manuscripts

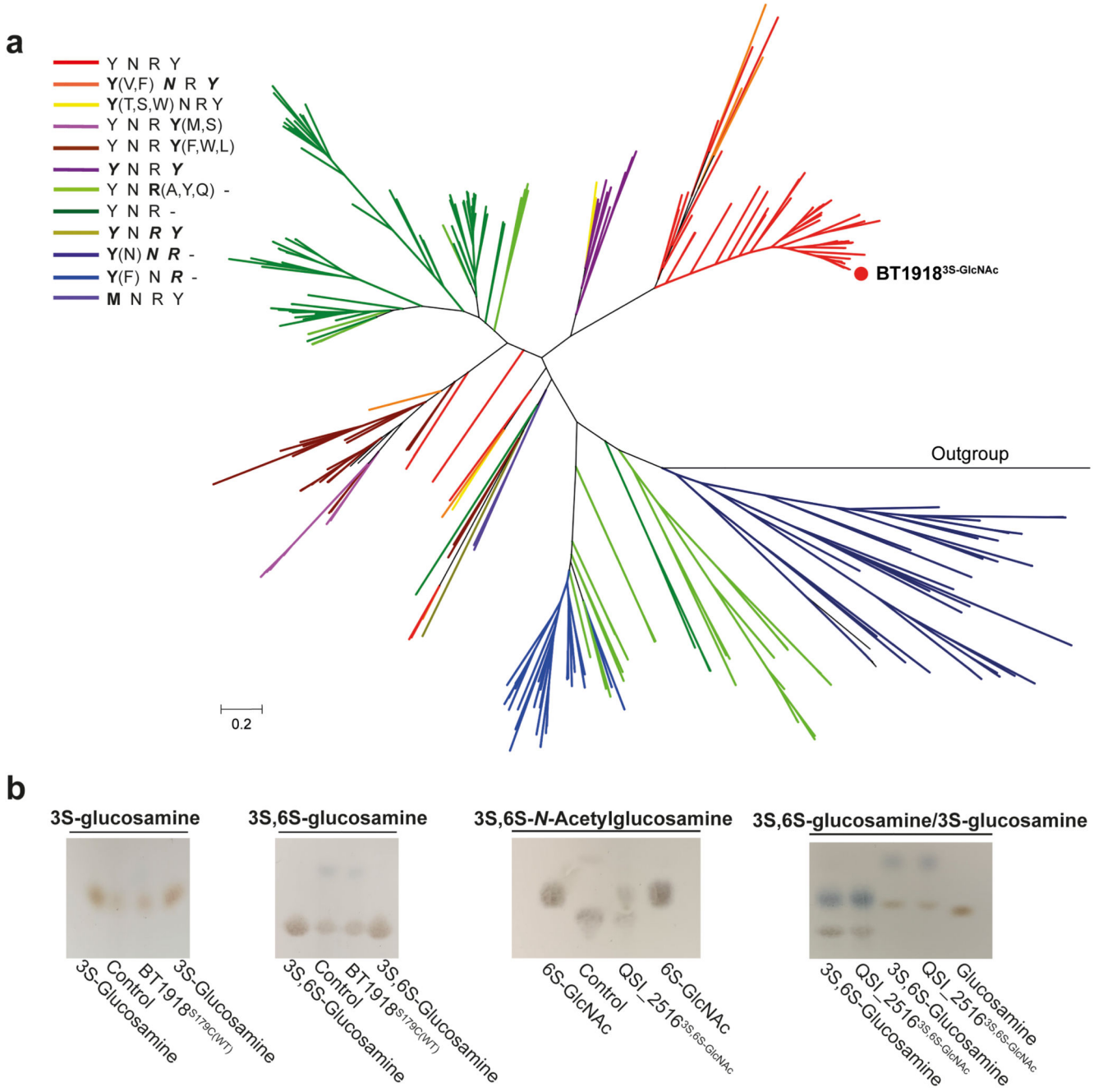
Extended Data



Extended Data Fig. 1. Electron density maps of extracted ligands.

The $2mF_{obs}$-F_c maps are shown contoured at 1σ for all substrates and products co-crystallised with their respective sulfatase.

Europe PMC Funders Author Manuscripts

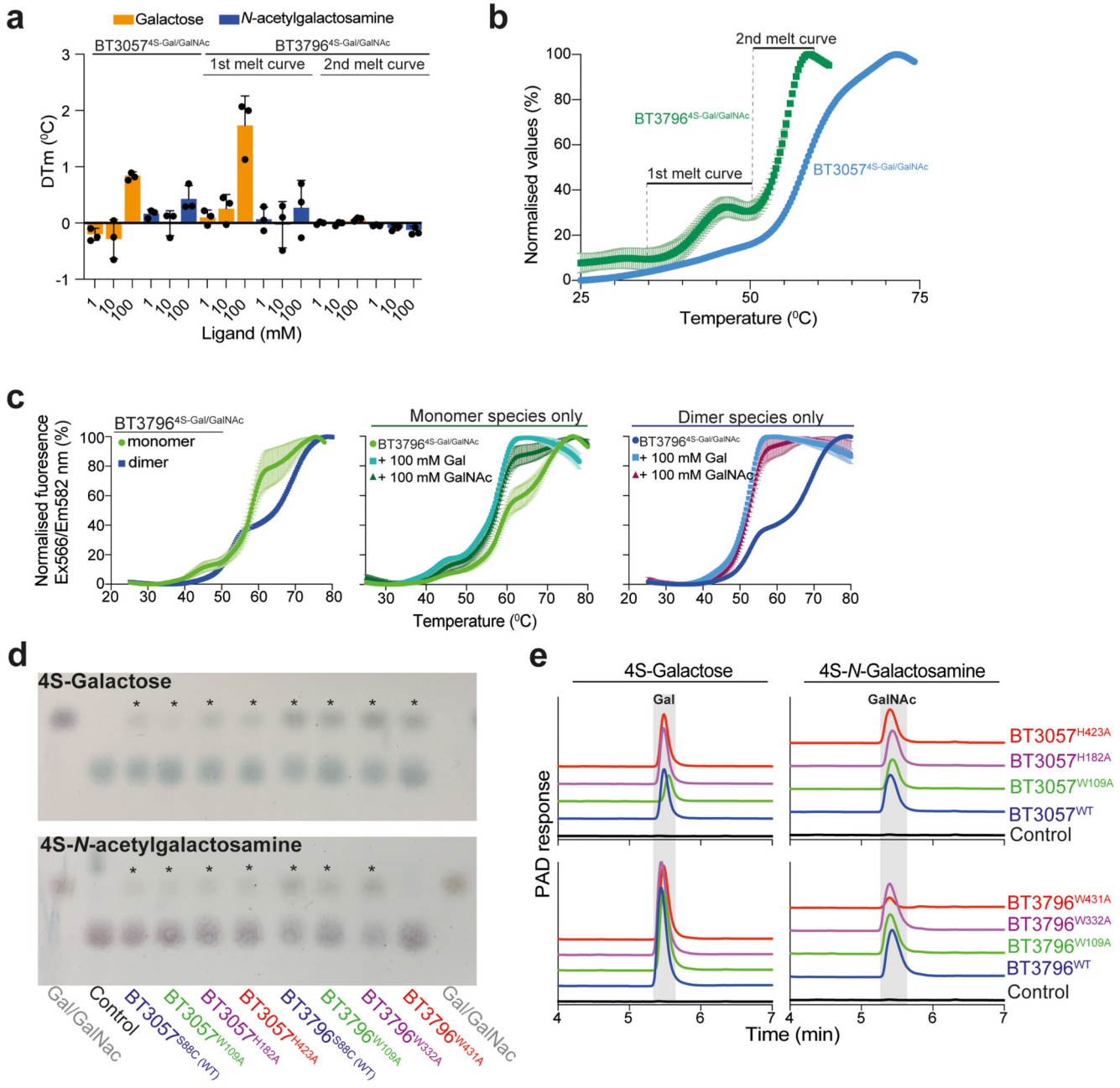


Extended Data Fig. 2. Biophysical, Specificity and phylogenetic analysis of BT1918$^{3S\text{-}GlcNAc}$ and S1_46.

a, Radial version of the phylogenetic tree of representative sulfatases from subfamily S1_46. The tree comprise a total of 564 sequences with 250 being Firmicutes; 156 are Bacteroidetes; 54 are Actinobacteria; 25 are Proteobacteria; 20 are Lentisphaerae. For clarity all labels and sequence accession codes have been omitted. The annotations next to the colour code reveal the presence or absence of conserved residues crucial for substrate recognition by BT1918$^{3S\text{-}GlcNAc}$ (acc-code Q8A6G6) in the following order: Y94, N174, R327 and Y408. These residues are invariant in HGM Bacteroidetes, whilst in Firmicutes

from the HGM, the Y408 equivalent is not conserved. The residues are coloured as following: black means an equivalent residue is present; a grey and bold letter at any position means that the corresponding residue is replaced by that amino acid; a grey, bold and italic letter at any position means that the equivalent position is replaced by any type of amino acid; a bold grey letter followed by one-letter codes in parentheses indicates that the equivalent position can be substituted by any of those amino acids; the dash at the Y408-equivalent position indicates that no equivalent amino acid can be deduced from the multiple alignment. Branches of the same colour have the corresponding pattern in common. The red filled circle designates the sequence of the S1_46 sulfatase from *B. thetaiotaomicron* (See Supplementary Fig. 2 for full tree). **b,** Thin layer chromatography analysis of BT1918$^{3S\text{-}GlcNAc}$ versus 3S-glucosamine and 3S,6S-glucosamine and QSI_2516$^{3S,6S\text{-}GlcNAc}$ versus 3S-glucosamine and 3S,6S-glucosamine. All assays described were performed for 48 h at 37°C, containing 6 mM substrate and 5 μM (BT1918$^{3S\text{-}GlcNAc}$) or 100 μM enzyme (QSI_2516$^{3S,6S\text{-}GlcNAc}$) and 3 mM HEPES pH 7.0, 45 mM NaCl and 5 mM $CaCl_2$.

Europe PMC Funders Author Manuscripts

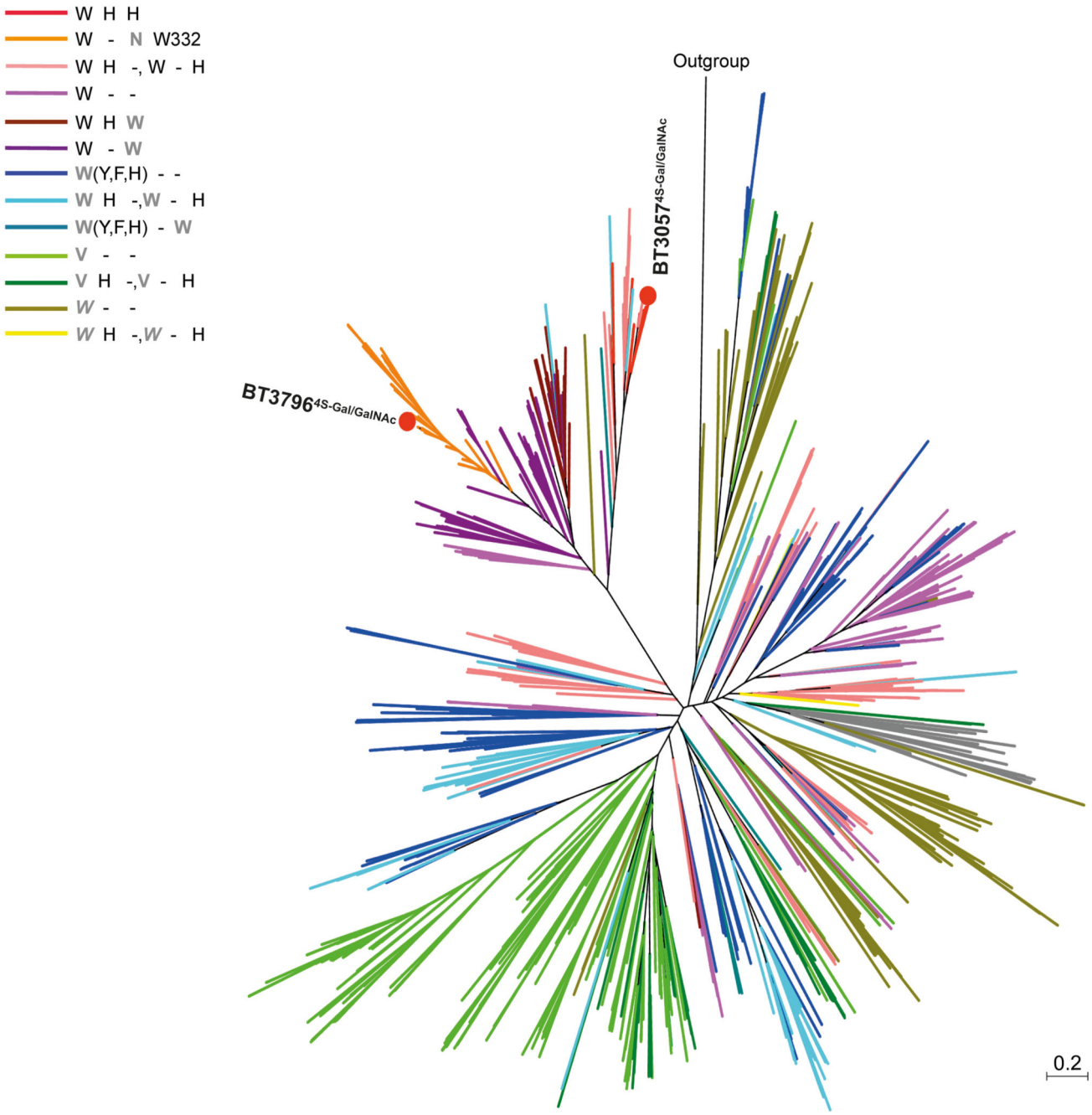


Extended Data Fig. 3. Activity and stability analysis of S1_16 sulfatases and their mutant variants.

a, DSF analysis of the effects of galactose and *N*-acetylgalactosamine on thermostability, with a positive-shift indicative of substrate binding. **b,** Normalised DSF melt curves of BT3057$^{4S\text{-}Gal/GalNAc}$ and BT3796$^{4S\text{-}Gal/GalNAc}$. **c,** DSF melt curves of purified monomer and dimer species (left), monomer species in the presence of galactose and *N*-acetylgalactosamine (middle), or dimer species in the presence of galactose and *N*-acetylgalactosamine (right). **d,** Thin layer chromatography (TLC) analysis of wild-type (WT) and mutant S1_16 sulfatases. Asterisks are placed above lanes where sulfatase activity

Europe PMC Funders Author Manuscripts

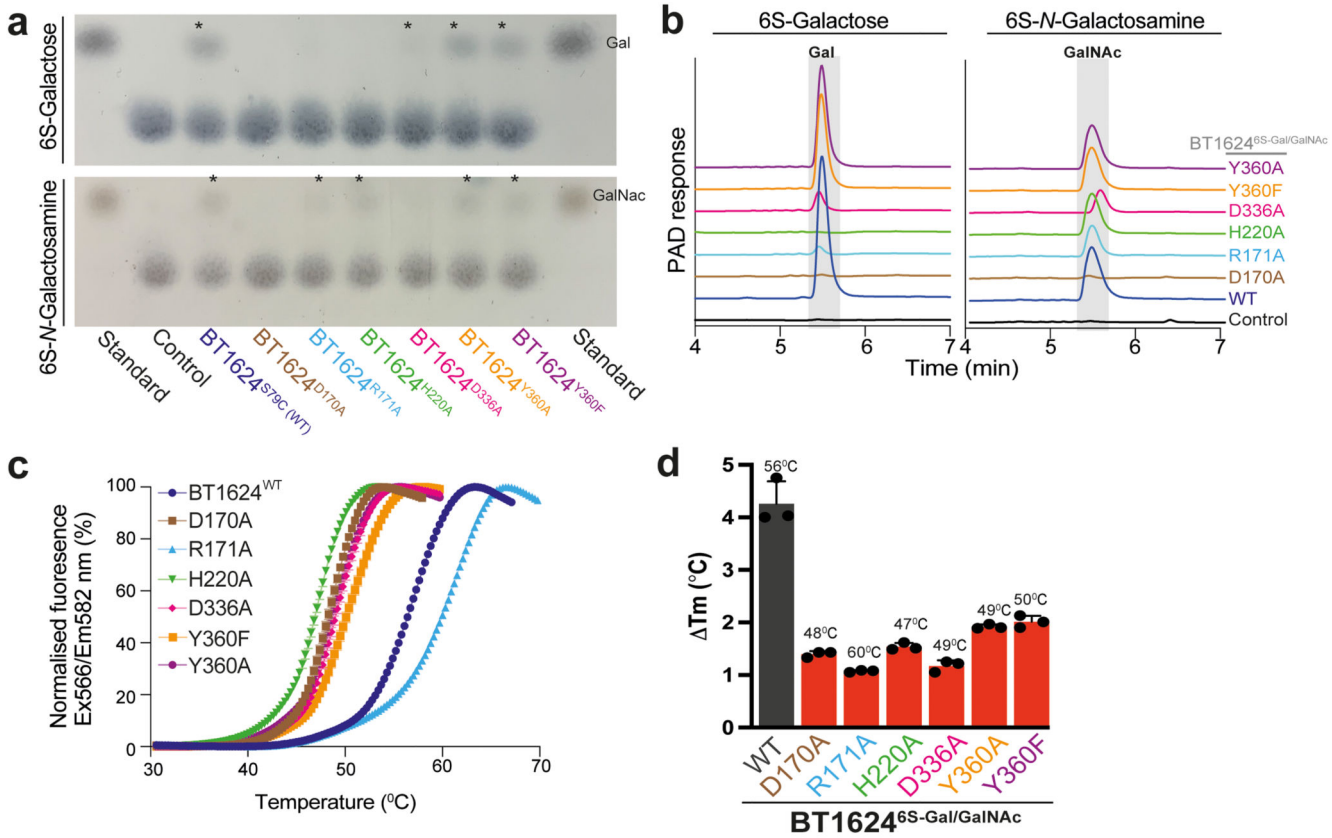
is observed. **e,** High pressure anion exchange chromatography (HPAEC) of WT and mutants. A grey block highlights the desulfated product. Both TLC and HPAEC reactions utilised 6 mM substrate and 1 μM enzyme, except for W109A variants where 10 μM was used, with 3 mM HEPES, 45 mM NaCl, and 5 mM $CaCl_2$. Reactions incubated at 37°C for 48 h. Control represents the substrate incubated in same conditions without adding enzyme. Experiments are technical triplicates and error bars represent SEM.



Extended Data Fig. 4. Radial phylogenetic tree of representative sulfatases from subfamily S1_16.

Europe PMC Funders Author Manuscripts

The tree comprises a total of 1368 sequences of which 854 are Bacteroidetes; 211 are Planctomycetes; 107 are Kiritimatiellaeota; 64 are Verrucomicrobia; 53 are Lentisphaerae For clarity all labels and sequence accession codes have been omitted. The annotations next to the colour code reveal the presence or absence of conservation of the critical residues in substrate recognition by BT3057$^{4S\text{-}Gal/GalNAc}$ (acc-code Q8A397) in the following order: W109, H182 and H423. Sequences coded by orange branches contain an additional W332 present in BT3796$^{4S\text{-}Gal/GalNAc}$ (acc-code Q8A171) but absent in other sequences. For simplification the residue numbers have been omitted, except for W332. The residues are coloured as following: black means an equivalent amino acid is present; a grey and bold letter at any position means that the corresponding residue is replaced by that amino acid; a grey and italic letter at any position means that the equivalent position is replaced by any type of amino acid; a bold grey letter followed by one-letter codes in parentheses indicates that the equivalent position can be substituted by any of those amino acids; the dash at the H-equivalent position indicates that no equivalent amino acid can be deduced from the multiple alignment. When two patterns are indicated separated by a comma (i.e. W - H, W H -) both have been given the same colour code. Branches having the same colour have the corresponding pattern in common. Red filled circles designate sequences of S1_16 sulfatases from *B. thetaiotaomicron* (See Supplementary Fig. 4 for full tree).

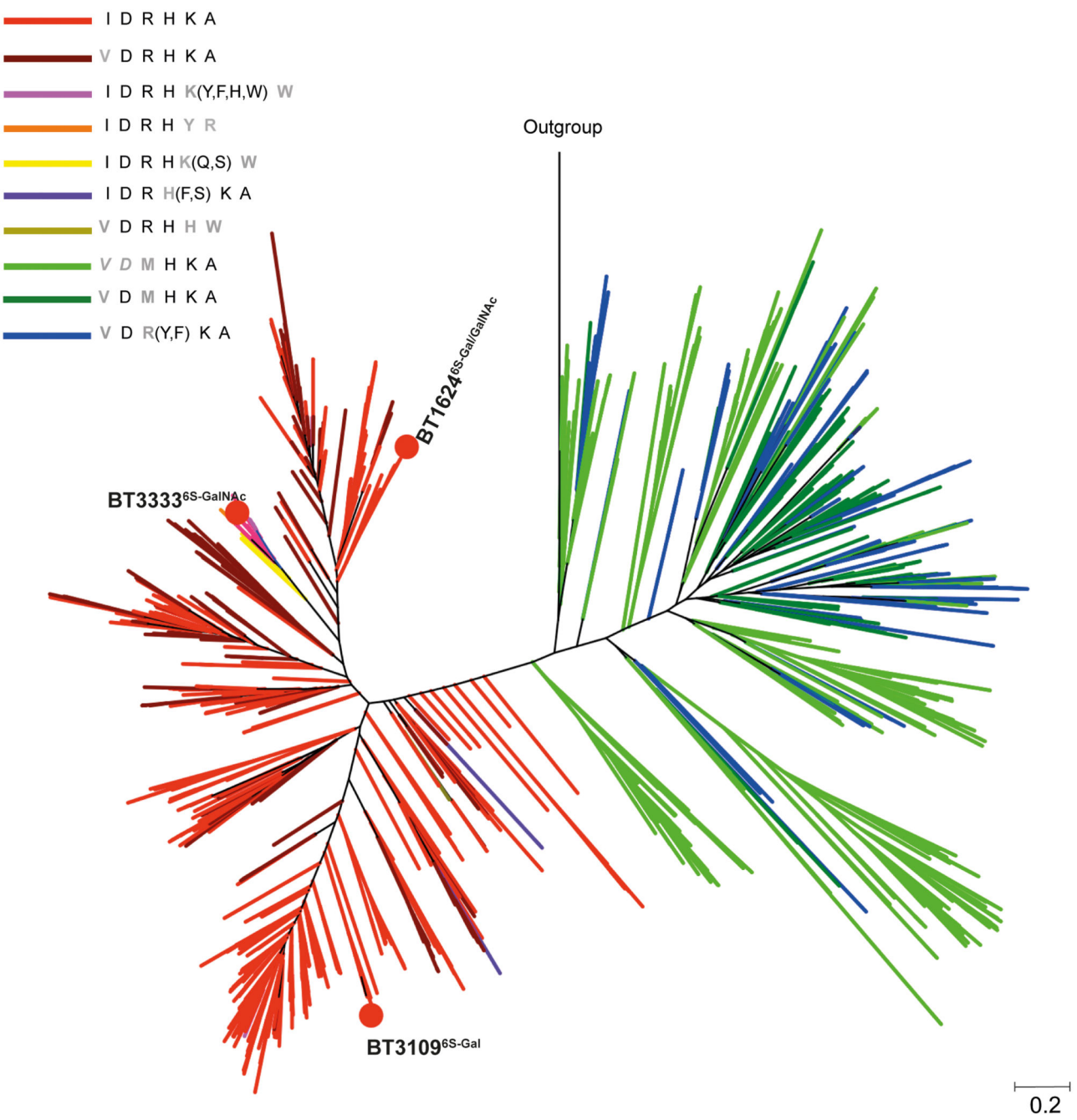


Extended Data Fig. 5. Analysis of the activity and stability of BT1624$^{6S\text{-}Gal/GalNAc}$ and its mutant variants.

Europe PMC Funders Author Manuscripts

a, Thin layer chromatography (TLC) analysis of wild-type (WT) BT1624$^{6S\text{-}Gal/GalNAc}$ and its mutants. Asterisks are placed above lanes where activity is observed. **b,** High pressure anion exchange chromatography (HPAEC) of wild-type BT1624$^{6S\text{-}Gal/GalNAc}$ WT and its mutants. The desulfated product is highlighted by a grey box. All TLC (a) HPAEC (b) reactions utilised 6 mM substrate and 5 μM enzyme, with 3 mM HEPES, 45 mM NaCl, and 5 mM $CaCl_2$. Reactions incubated for 48 h at 37°C. Control represents the substrate incubated in same conditions without adding enzyme. **c,** DSF analysis showing relative thermostability of BT1624$^{6S\text{-}Gal/GalNAc}$ mutant proteins with respect to the WT enzyme. **d,** DSF analysis of the effects of alanine scanning on the ability of BT1624$^{6S\text{-}Gal/GalNAc}$ to bind galactose, with the Tm of the protein shown above the bar. The experiments were performed using 5 μM of protein and 324 mM of galactose in 100 mM BTP and 150 mM NaCl. Experiments are technical triplicates and error bars represent SEM.

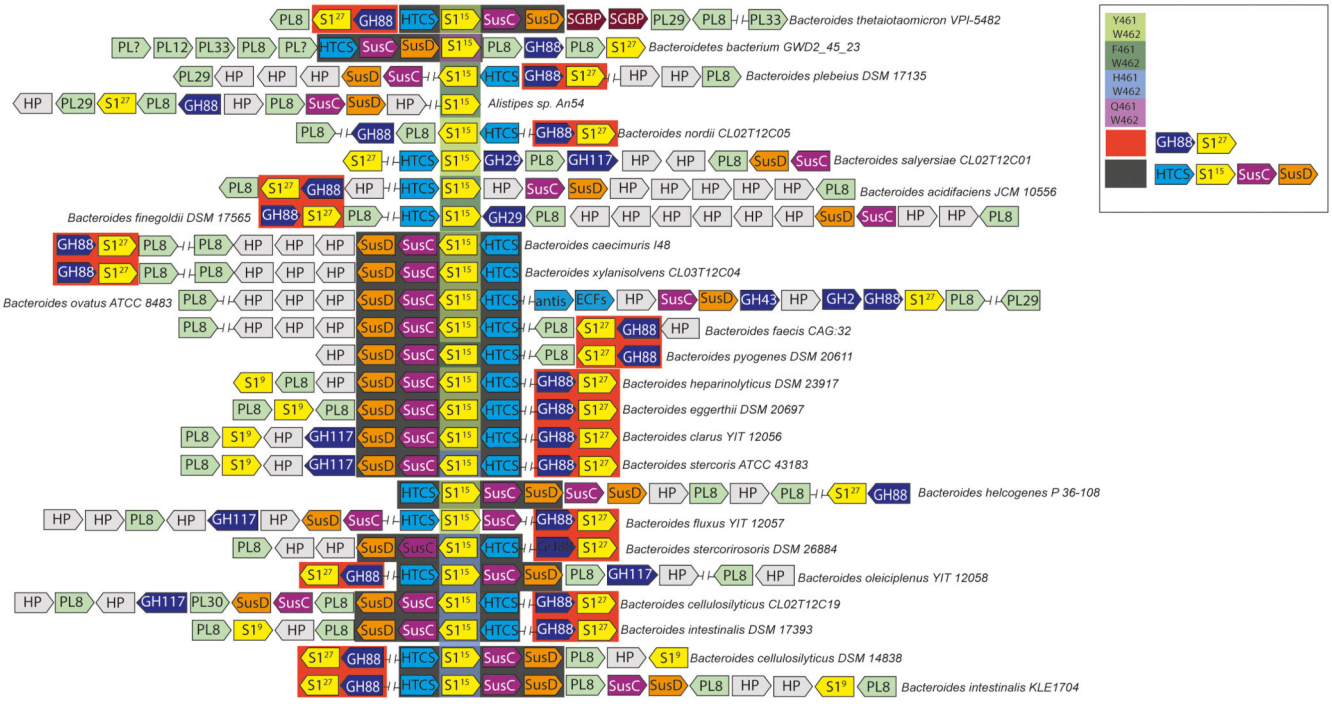
Europe PMC Funders Author Manuscripts



Extended Data Fig. 6. Radial phylogenetic tree of S1_15 showing the conservation of the galacto-recognition triad and *N*-acetyl-D-galactosamine specificity features.

The tree comprises a total of 1906 sequences of which 1424 are Bacteroidetes; 172 are Planctomycetes ;119 are Kiritimatiellaeota ;57 are Proteobacteria ;53 are Verrucomicrobia. The annotations next to the colour code concern the presence or absence of conservation of the BT1624$^{6S\text{-}Gal/GalNAc}$ (acc-code Q8A7A1) indicated residues and in this order: I100, D170, R171, H220, K461 and A462. These residues are crucial in substrate recognition and D170, R171, and H220 represent the galacto-recognition triad within S1_15 subfamily. For simplification the residue numbers have been omitted. For example, an I in black

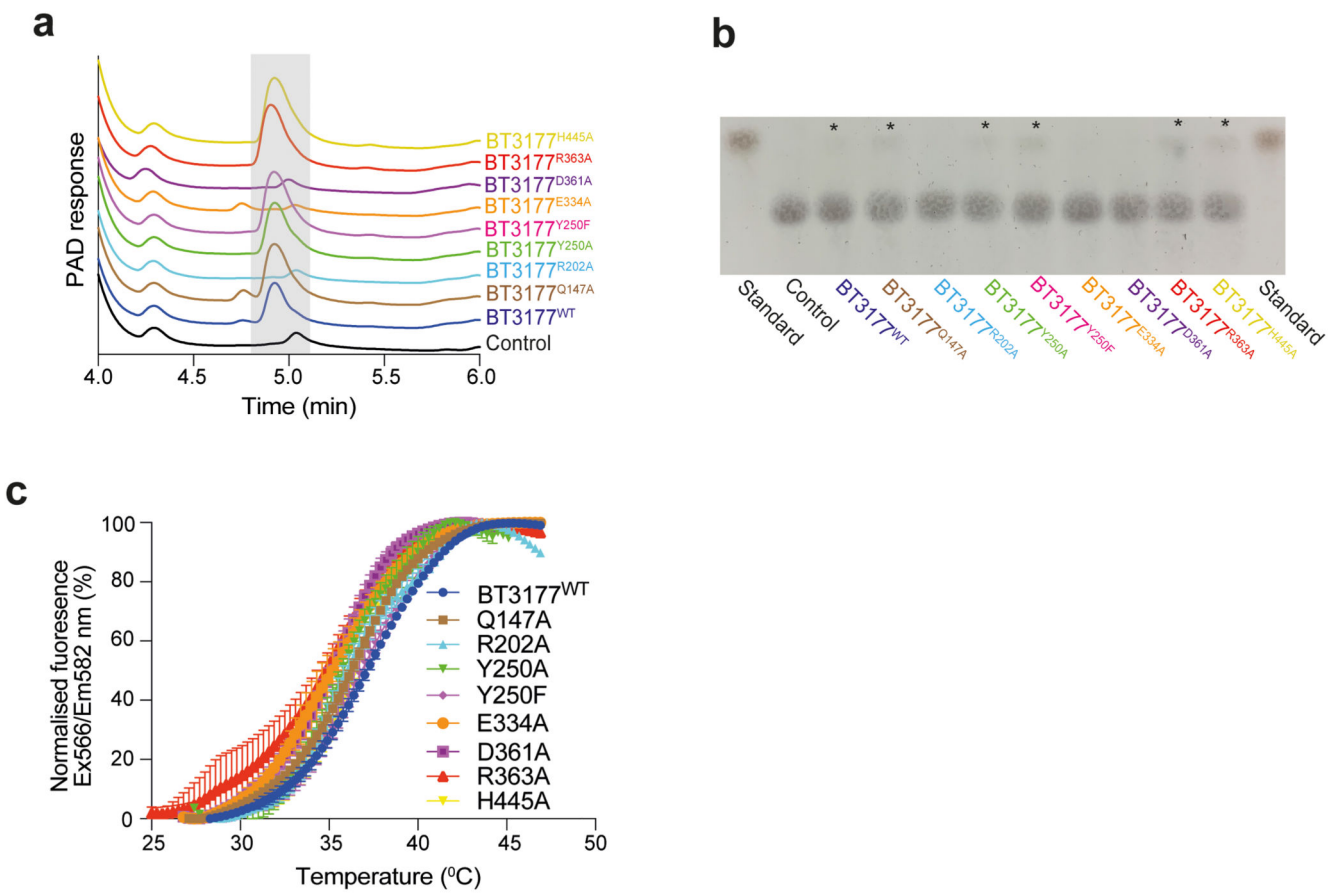
means an equivalent isoleucine is present; a grey and bold letter at any position means that the corresponding residue is replaced by that amino acid; a grey and italic letter at any position means that the equivalent position can be replaced by any type of amino acid; a bold grey letter followed by one-letter codes in parentheses indicates that the equivalent position is substituted by any of those amino acids. Branches having the same colour have the corresponding pattern in common. For clarity all labels and sequence accession codes have been omitted. Red filled circles designate sequences of S1_15 sulfatases from *B. thetaiotaomicron* (See Supplementary Fig. 6 for full tree).



Extended Data Fig. 7. Conservation of the *N*-acetyl-D-galactosamine specificity features (Y463/W464) in S1_15 enzymes within PULs targeting chondroitin sulfate.

Schematic representation of PULs targeting chondroitin sulfate aligned by orthologues of BT3333$^{6S\text{-}GalNAc}$. Light green background shows orthologues with Y463/W464, a dark green background highlights orthologues with F463/W464, a light blue background highlights orthologues with H463/W464, and a purple background highlights orthologues with Q463/W464. The numbering used corresponds to the sequence of BT3333$^{6S\text{-}GalNAc}$. A red background highlights the presence of GH88 and S1_27 (an endo 4S-chrondroitin sulfatase), which is encoded by a discrete genetic region not always physically localised next to the core PUL. A black background highlights a core block observed in CS PULs containing BT3333$^{6S\text{-}GalNAc}$ orthologues. HP (protein of unknown function), S1 (sulfatase S1 with the respective subfamily number superscript), GHXX (glycoside hydrolase with X representing the family number), PL (polysaccharide lyase), DUF (domain of unknown function).

Europe PMC Funders Author Manuscripts

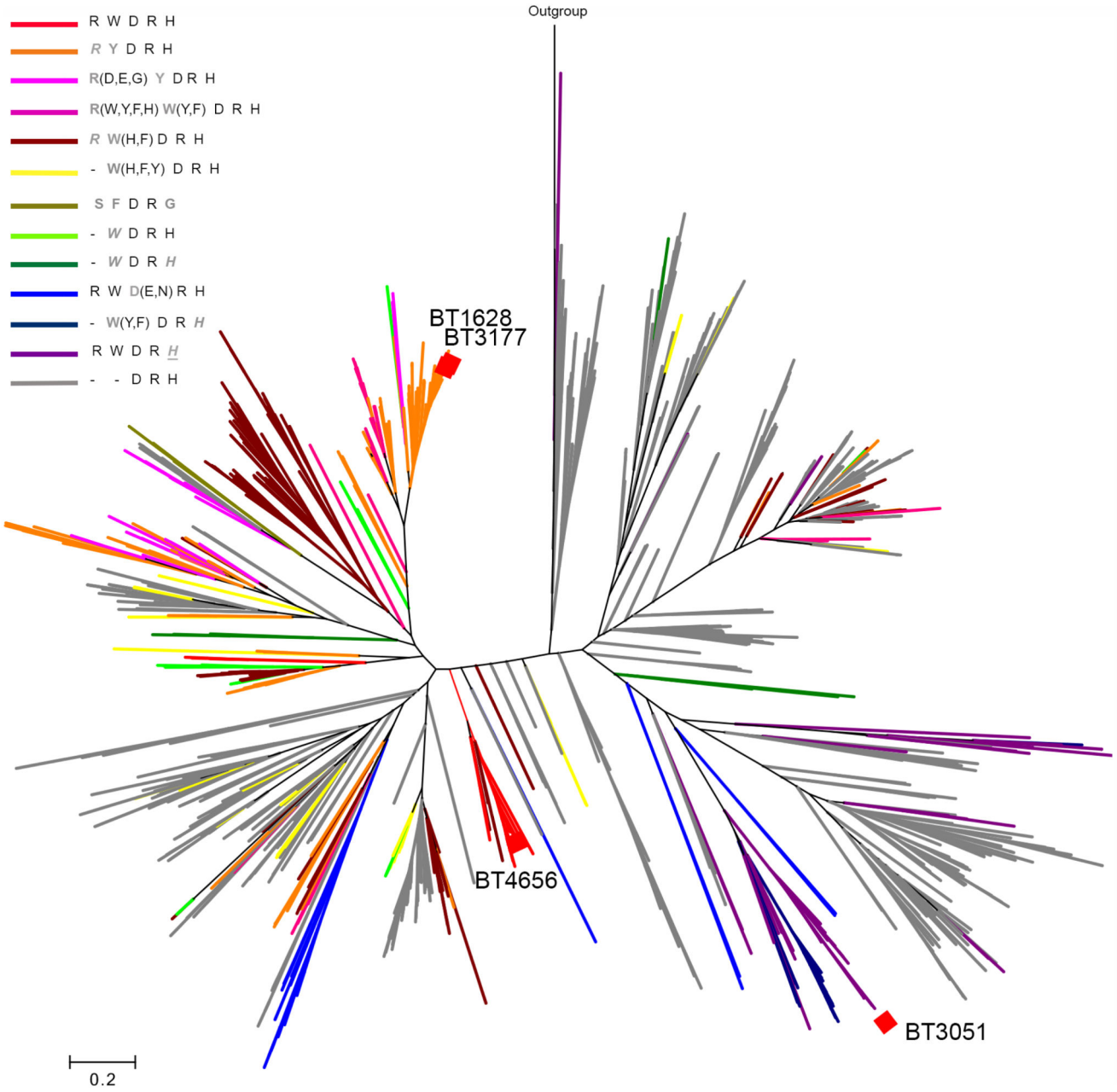


Extended Data Fig. 8. Analysis of the activity and stability of BT3177$^{6S\text{-}GlcNAc}$ and mutant variants.

a, High pressure anion exchange chromatography (HPAEC) of wild-type BT3177$^{6S\text{-}GlcNAc}$ wild-type (WT) and substituted variants. The produced product is highlighted by a grey box. **b,** Thin layer chromatography (TLC) analysis of WT BT3177$^{6S\text{-}GlcNAc}$ and its mutants. Asterisks are placed above lanes where activity is observed. Both HPAEC (a) and TLC (b) reactions utilised 6 mM substrate and 5 μM enzyme, with 3 mM HEPES, 45 mM NaCl, and 5 mM $CaCl_2$ over a 48 h period at 37°C. Control represents the substrate incubated in same conditions without enzyme. **c,** DSF analysis showing relative thermostability of mutant proteins of BT3177$^{6S\text{-}GlcNAc}$ in comparison to the WT enzyme. Experiments are technical triplicates and error bars represent SEM.

Europe PMC Funders Author Manuscripts

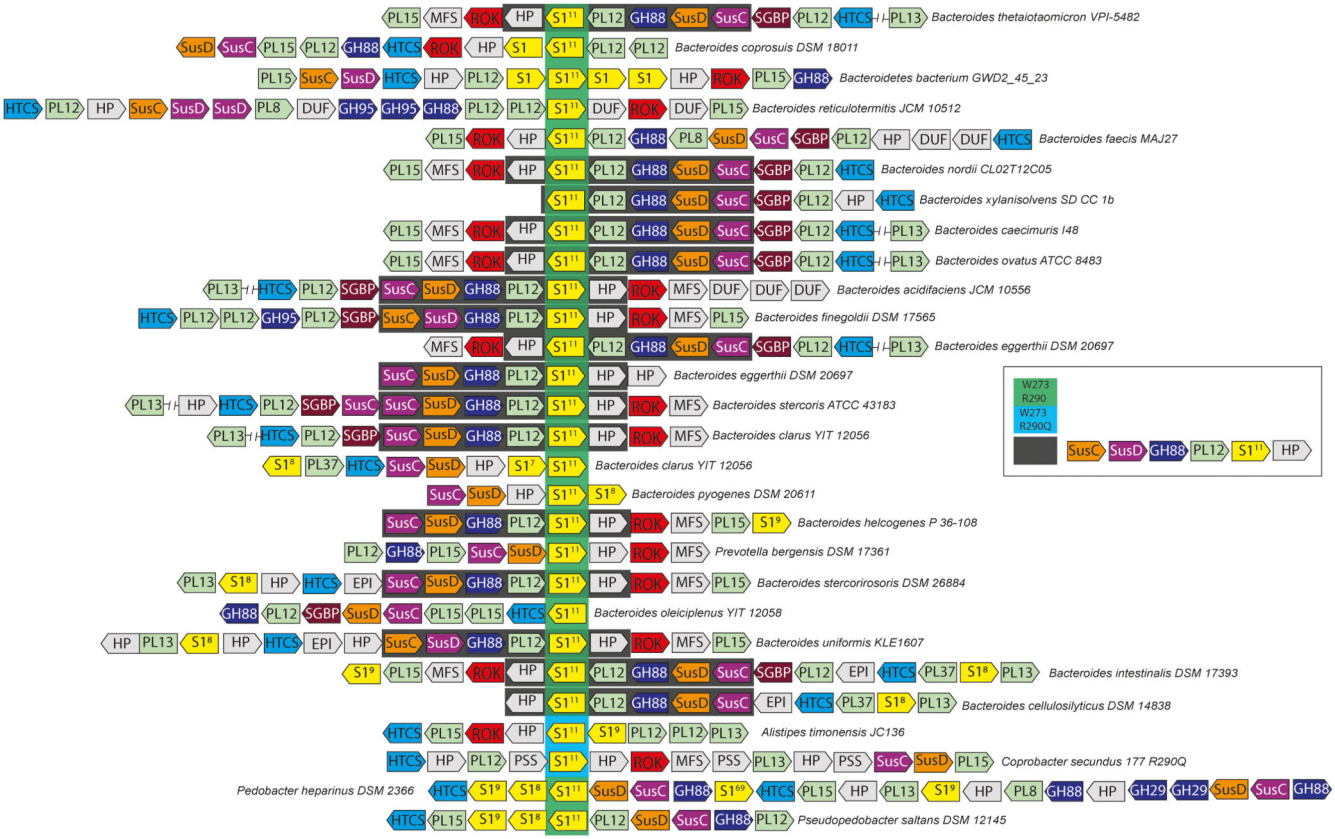
Europe PMC Funders Author Manuscripts



Extended Data Fig. 9. Radial phylogenetic tree of S1_11 showing the conservation of the gluco-recognition triad and *N*-sulfate specificity features.

The tree comprises a total of 2178 sequences of which 1190 are Bacteroidetes; 233 are Verrucomicrobia;184 are Planctomycetes;143 are Ascomycota (fungi); 100 are Actinobacteria.The annotations next to the colour code concern the presence or absence of conservation of the indicated residues and in this order: R290, W273, D385, R387 and H471. These residues are required for substrate recognition by BT4656$^{6S\text{-}GlcNAc/GlcNS}$ (acc-code Q89YS5). D385, R387, and H471 represent the gluco-recognition triad, whilst the presence of W or R at positions 273 and 290, respectively, represent *N*-sulfate specificity features. Residue numbers have been omitted for simplicity. For example, an R in black

means an equivalent arginine is present; a grey and bold letter at this position means that the corresponding residue is replaced by that amino acid; the grey and italic R at this position means that the R-equivalent position is replaced by any type of amino acid; a bold grey R followed by one-letter codes in parentheses indicates that the R-equivalent position can be substituted by any of those amino acids; the dash at the R-equivalent position indicates that no equivalent amino acid can be deduced from the multiple alignment. Branches having the same colour have the corresponding pattern in common. Red filled diamonds designate sequences of S1_11 sulfatases from *B. thetaiotaomicron*. All sequences in the specific branch that contains BT4656$^{6S\text{-}GlcNAc/GlcNS}$ are found within a conserved heparan sulfate PUL. For clarity, all labels and sequence accession codes have been omitted (See Supplementary Fig 7 for full tree).



Extended Data Fig. 10. Conservation of the *N*-sulfate targeting features, W273/R290, in S1_11 enzymes within PULs targeting heparan sulfate.

PULs targeting heparan sulfate (HS) aligned by orthologues of BT4656$^{6S\text{-}GlcNAc/GlcNS}$. Orthologues of BT4656$^{6S\text{-}GlcNAc/GlcNS}$ with W273/R290 and W273/Q290 are highlighted with a green and blue background, respectively. A black background highlights a core block observed in HS PULs containing BT4656$^{6S\text{-}GlcNAc/GlcNS}$ orthologues. HP (protein of unknown function), S1 (sulfatase S1 with the respective subfamily number superscript), GHXX (glycoside hydrolase with X representing the family number), PL (polysaccharide lyase), DUF (domain of unknown function).

Europe PMC Funders Author Manuscripts

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement N° 748336 and Wilhelm och Martina Lundgrens Vetenskapsfond (2020.3597) awarded to ASL. The European Research Council ERC (694181), the Knut and Alice Wallenberg Foundation (2017.0028), Swedish Research Council (2017-00958) awarded to GCH, the National Institute of Health (R01 DK118024 and DK125445 and U01AI095473) award to ECM and GCH, and the Academy of Medical Sciences/Wellcome Trust through the Springboard Grant (SBF005\1065 163470) awarded to AC. The authors acknowledge access to the SOLEIL and Diamond Light sources via both the University of Liverpool and Newcastle University BAGs (proposals mx21970 and mx18598, respectively). We thank the staff of DIAMOND, SOLEIL, and members of the Liverpool's Molecular biophysics group for assistance with data collection. We are also grateful for Dr. Erwan Corre's help regarding bioinformatics analyses (ABIMS platform, Station Biologique de Roscoff, France).

Data availability statement

Source Data for all experiments, along with corresponding statistical test values, where appropriate, are provided within the paper and in Supplementary information. The crystal structure dataset generates have been deposit in the in the Protein Data Bank (PDB) under the following accession numbers: 7OZ8, 7OZ9, 7OZA, 7OZE, 7OZC, 7P26, and 7P24.

Code availability statement

No new codes were developed or compiled in this study

References

1. Sarrazin S, Lamanna WC, Esko JD. Heparan sulfate proteoglycans. Cold Spring Harbor perspectives in biology. 2011; 3 doi: 10.1101/cshperspect.a004952
2. Soares da Costa D, Reis RL, Pashkuleva I. Sulfation of Glycosaminoglycans and Its Implications in Human Health and Disorders. Annu Rev Biomed Eng. 2017; 19: 1–26. DOI: 10.1146/annurev-bioeng-071516-044610 [PubMed: 28226217]
3. Luis AS, et al. A single sulfatase is required to access colonic mucin by a gut bacterium. Nature. 2021; 598: 332–337. DOI: 10.1038/s41586-021-03967-5 [PubMed: 34616040]
4. Bloom SM, et al. Commensal Bacteroides species induce colitis in host-genotype-specific fashion in a mouse model of inflammatory bowel disease. Cell host & microbe. 2011; 9: 390–403. DOI: 10.1016/j.chom.2011.04.009 [PubMed: 21575910]
5. Johansson ME, Larsson JM, Hansson GC. The two mucus layers of colon are organized by the MUC2 mucin, whereas the outer layer is a legislator of host-microbial interactions. Proceedings of the National Academy of Sciences of the United States of America. 2011; 108 (Suppl 1) 4659–4665. DOI: 10.1073/pnas.1006451107 [PubMed: 20615996]
6. Field CB, Behrenfeld MJ, Randerson JT, Falkowski P. Primary production of the biosphere: integrating terrestrial and oceanic components. Science. 1988; 281: 237–240. DOI: 10.1126/science.281.5374.237
7. Chen J, et al. Laminarin, a Major Polysaccharide in Stramenopiles. Mar Drugs. 2021; 19 doi: 10.3390/md19100576
8. Hettle AG, et al. Insights into the κ/ι-carrageenan metabolism pathway of some marine Pseudoalteromonas species. Communications Biology. 2019; 2: 474. doi: 10.1038/s42003-019-0721-y [PubMed: 31886414]

9. Reisky L, et al. A marine bacterial enzymatic cascade degrades the algal polysaccharide ulvan. Nature chemical biology. 2019; 15: 803–812. DOI: 10.1038/s41589-019-0311-9 [PubMed: 31285597]
10. Zhang Q, et al. Chemical characteristics of a polysaccharide from Porphyra capensis (Rhodophyta). Carbohydr Res. 2005; 340: 2447–2450. DOI: 10.1016/j.carres.2005.08.009 [PubMed: 16150429]
11. Ponce NMA, Stortz CAA. Comprehensive and Comparative Analysis of the Fucoidan Compositional Data Across the Phaeophyceae. Front Plant Sci. 2020; 11 556312 doi: 10.3389/fpls.2020.556312 [PubMed: 33324429]
12. Panggabean JA, et al. Antiviral Activities of Algal-Based Sulfated Polysaccharides. Molecules. 2022; 27 doi: 10.3390/molecules27041178
13. Pereira, L. Carrageenans: Sources and Extraction Methods, Molecular Structure, Bioactive Properties and Health Effects. Nova Science Publishers, Incorporated; 2016.
14. Cartmell A, et al. How members of the human gut microbiota overcome the sulfation problem posed by glycosaminoglycans. Proceedings of the National Academy of Sciences of the United States of America. 2017; 114: 7037–7042. DOI: 10.1073/pnas.1704367114 [PubMed: 28630303]
15. Tuncil YE, et al. Reciprocal Prioritization to Dietary Glycans by Gut Bacteria in a Competitive Environment Promotes Stable Coexistence. mBio. 2017; 8 doi: 10.1128/mBio.01068-17
16. Raghavan V, Groisman EA. Species-specific dynamic responses of gut bacteria to a mammalian glycan. J Bacteriol. 2015; 197: 1538–1548. DOI: 10.1128/JB.00010-15 [PubMed: 25691527]
17. Cheng HY, Ning MX, Chen DK, Ma WT. Interactions Between the Gut Microbiota and the Host Innate Immune Response Against Pathogens. Front Immunol. 2019; 10: 607. doi: 10.3389/fimmu.2019.00607 [PubMed: 30984184]
18. McNeil NI. The contribution of the large intestine to energy supplies in man. The American journal of clinical nutrition. 1984; 39: 338–342. DOI: 10.1093/ajcn/39.2.338 [PubMed: 6320630]
19. Goodman AL, et al. Identifying genetic determinants needed to establish a human gut symbiont in its habitat. Cell host & microbe. 2009; 6: 279–289. DOI: 10.1016/j.chom.2009.08.003 [PubMed: 19748469]
20. Li H, et al. The outer mucus layer hosts a distinct intestinal microbial niche. Nature communications. 2015; 6 8292 doi: 10.1038/ncomms9292
21. Tsai HH, Dwarakanath AD, Hart CA, Milton JD, Rhodes JM. Increased faecal mucin sulphatase activity in ulcerative colitis: a potential target for treatment. Gut. 1995; 36: 570–576. DOI: 10.1136/gut.36.4.570 [PubMed: 7737566]
22. Alipour M, et al. Mucosal Barrier Depletion and Loss of Bacterial Diversity are Primary Abnormalities in Paediatric Ulcerative Colitis. J Crohns Colitis. 2016; 10: 462–471. DOI: 10.1093/ecco-jcc/jjv223 [PubMed: 26660940]
23. Hickey CA, et al. Colitogenic Bacteroides thetaiotaomicron Antigens Access Host Immune Cells in a Sulfatase-Dependent Manner via Outer Membrane Vesicles. Cell host & microbe. 2015; 17: 672–680. DOI: 10.1016/j.chom.2015.04.002 [PubMed: 25974305]
24. Barbeyron T, et al. Matching the Diversity of Sulfated Biomolecules: Creation of a Classification Database for Sulfatases Reflecting Their Substrate Specificity. PloS one. 2016; 11 e0164846 doi: 10.1371/journal.pone.0164846 [PubMed: 27749924]
25. Hanson SR, Best MD, Wong CH. Sulfatases: structure, mechanism, biological activity, inhibition, and synthetic utility. Angewandte Chemie. 2004; 43: 5736–5763. DOI: 10.1002/anie.200300632 [PubMed: 15493058]
26. Hettle AG, et al. The Molecular Basis of Polysaccharide Sulfatase Activity and a Nomenclature for Catalytic Subsites in this Class of Enzyme. Structure. 2018; 26: 747–758. e744 doi: 10.1016/j.str.2018.03.012 [PubMed: 29681469]
27. Terrapon N, Lombard V, Gilbert HJ, Henrissat B. Automatic prediction of polysaccharide utilization loci in Bacteroidetes species. Bioinformatics. 2015; 31: 647–655. DOI: 10.1093/bioinformatics/btu716 [PubMed: 25355788]
28. Ndeh D, et al. Metabolism of multiple glycosaminoglycans by Bacteroides thetaiotaomicron is orchestrated by a versatile core genetic locus. Nature communications. 2020; 11: 646. doi: 10.1038/s41467-020-14509-4

29. Wei W, Ninonuevo MR, Sharma A, Danan-Leon LM, Leary JA. A comprehensive compositional analysis of heparin/heparan sulfate-derived disaccharides from human serum. Anal Chem. 2011; 83: 3703–3708. DOI: 10.1021/ac2001077 [PubMed: 21473642]
30. Sidhu NS, et al. Structure of sulfamidase provides insight into the molecular pathology of mucopolysaccharidosis IIIA. Acta crystallographica. Section D, Biological crystallography. 2014; 70: 1321–1335. DOI: 10.1107/S1399004714002739 [PubMed: 24816101]
31. von Bulow R, et al. Defective oligomerization of arylsulfatase a as a cause of its instability in lysosomes and metachromatic leukodystrophy. The Journal of biological chemistry. 2002; 277: 9455–9461. DOI: 10.1074/jbc.M111993200 [PubMed: 11777924]
32. Robb CS, et al. Metabolism of a hybrid algal galactan by members of the human gut microbiome. Nature chemical biology. 2022; doi: 10.1038/s41589-022-00983-y
33. Juers DH, et al. A structural view of the action of Escherichia coli (lacZ) beta-galactosidase. Biochemistry. 2001; 40: 14781–14794. DOI: 10.1021/bi011727i [PubMed: 11732897]
34. Helbert W, et al. Discovery of novel carbohydrate-active enzymes through the rational exploration of the protein sequences space. Proceedings of the National Academy of Sciences of the United States of America. 2019; 116: 6063–6068. DOI: 10.1073/pnas.1815791116 [PubMed: 30850540]
35. Lapebie P, Lombard V, Drula E, Terrapon N, Henrissat B. Bacteroidetes use thousands of enzyme combinations to break down glycans. Nature communications. 2019; 10: 2043. doi: 10.1038/s41467-019-10068-5
36. Pudlo NA, et al. Extensive transfer of genes for edible seaweed digestion from marine to human gut bacteria. bioRxiv. 2020; 2020.2006.2009.142968 doi: 10.1101/2020.06.09.142968
37. Verma S, et al. Identification and engraftment of new bacterial strains by shotgun metagenomic sequence analysis in patients with recurrent Clostridioides difficile infection before and after fecal microbiota transplantation and in healthy human subjects. PloS one. 2021; 16 e0251590 doi: 10.1371/journal.pone.0251590 [PubMed: 34252073]
38. Roche P, et al. Molecular basis of symbiotic host specificity in Rhizobium meliloti: nodH and nodPQ genes encode the sulfation of lipo-oligosaccharide signals. Cell. 1991; 67: 1131–1143. DOI: 10.1016/0092-8674(91)90290-f [PubMed: 1760841]
39. Das TM, Rao CP, Kolehmainen E. Synthesis and characterisation of N-glycosyl amines from the reaction between 4,6-O-benzylidene-D-glucopyranose and substituted aromatic amines and also between 2-(o-aminophenyl)benzimidazole and pentoses or hexoses. Carbohydr Res. 2001; 334: 261–269. DOI: 10.1016/s0008-6215(01)00202-6 [PubMed: 11527527]
40. Byrne DP, London JA, Eyers PA, Yates EA, Cartmell A. Mobility shift-based electrophoresis coupled with fluorescent detection enables real-time enzyme analysis of carbohydrate sulfatase activity. The Biochemical journal. 2021; 478: 735–748. DOI: 10.1042/BCJ20200952 [PubMed: 33480417]
41. Labourel A, et al. Structural and functional analysis of glycoside hydrolase 138 enzymes targeting chain A galacturonic acid in the complex pectin rhamnogalacturonan II. The Journal of biological chemistry. 2019; doi: 10.1074/jbc.RA118.006626
42. Byrne DP, et al. cAMP-dependent protein kinase (PKA) complexes probed by complementary differential scanning fluorimetry and ion mobility-mass spectrometry. The Biochemical journal. 2016; 473: 3159–3175. DOI: 10.1042/BCJ20160648 [PubMed: 27444646]
43. Kabsch W. Xds. Acta crystallographica. Section D, Biological crystallography. 2010; 66: 125–132. DOI: 10.1107/S0907444909047337 [PubMed: 20124692]
44. Evans P. Scaling and assessment of data quality. Acta crystallographica. Section D, Biological crystallography. 2006; 62: 72–82. DOI: 10.1107/S0907444905036693 [PubMed: 16369096]
45. Evans P. R. An introduction to data reduction: space-group determination, scaling and intensity statistics. Acta crystallographica. Section D, Biological crystallography. 2011; 67: 282–292. DOI: 10.1107/S090744491003982X [PubMed: 21460446]
46. Long F, Vagin AA, Young P, Murshudov GN. BALBES: a molecular-replacement pipeline. Acta crystallographica. Section D, Biological crystallography. 2008; 64: 125–132. DOI: 10.1107/S0907444907050172 [PubMed: 18094476]

47. McCoy AJ. Solving structures of protein complexes by molecular replacement with Phaser. Acta crystallographica. Section D, Biological crystallography. 2007; 63: 32–41. DOI: 10.1107/S0907444906045975 [PubMed: 17164524]
48. Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of Coot. Acta Crystallogr D Biol Crystallogr. 2010; 66: 486–501. DOI: 10.1107/S0907444910007493 [PubMed: 20383002]
49. Murshudov GN, et al. REFMAC5 for the refinement of macromolecular crystal structures. Acta crystallographica. Section D, Biological crystallography. 2011; 67: 355–367. DOI: 10.1107/S0907444911001314 [PubMed: 21460454]
50. Lebedev AA, et al. JLigand: a graphical tool for the CCP4 template-restraint library. Acta crystallographica. Section D, Biological crystallography. 2012; 68: 431–440. DOI: 10.1107/S090744491200251X [PubMed: 22505263]
51. Chen VB, et al. MolProbity: all-atom structure validation for macromolecular crystallography. Acta crystallographica Section D, Biological crystallography. 2010; 66: 12–21. DOI: 10.1107/S0907444909042073 [PubMed: 20057044]
52. Potterton L, et al. CCP4i2: the new graphical user interface to the CCP4 program suite. Acta Crystallogr D Struct Biol. 2018; 74: 68–84. DOI: 10.1107/S2059798317016035 [PubMed: 29533233]
53. Collaborative Computational Project, N. The CCP4 suite: programs for protein crystallography. Acta crystallographica Section D, Biological crystallography. 1994; 50: 760–763. DOI: 10.1107/S0907444994003112 [PubMed: 15299374]
54. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002; 30: 3059–3066. DOI: 10.1093/nar/gkf436 [PubMed: 12136088]
55. Clamp M, Cuff J, Searle SM, Barton GJ. The Jalview Java alignment editor. Bioinformatics. 2004; 20: 426–427. DOI: 10.1093/bioinformatics/btg430 [PubMed: 14960472]
56. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014; 30: 1312–1313. DOI: 10.1093/bioinformatics/btu033 [PubMed: 24451623]
57. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. 1981; 17: 368–376. DOI: 10.1007/BF01734359 [PubMed: 7288891]
58. Le SQ, Gascuel O. An improved general amino acid replacement matrix. Mol Biol Evol. 2008; 25: 1307–1320. DOI: 10.1093/molbev/msn067 [PubMed: 18367465]
59. Felsenstein J. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. Evolution. 1985; 39: 783–791. DOI: 10.1111/j.1558-5646.1985.tb00420.x [PubMed: 28561359]
60. Varki A, et al. Symbol Nomenclature for Graphical Representations of Glycans. Glycobiology. 2015; 25: 1323–1324. DOI: 10.1093/glycob/cwv091 [PubMed: 26543186]

Europe PMC Funders Author Manuscripts

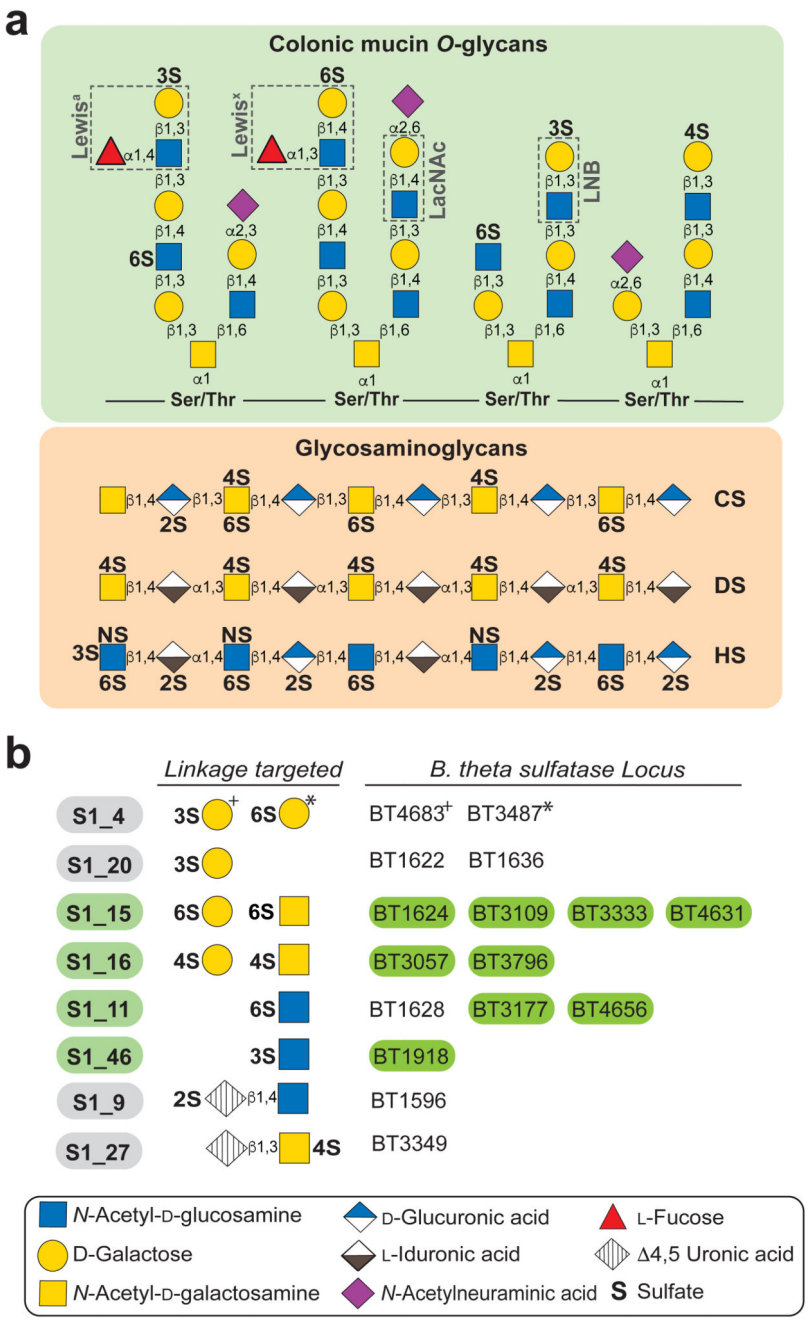


Fig. 1. Schematic representation of sulfated host carbohydrates found in the colon

a, Schematic representation of sulfated host glycans found in mucin *O*-glycans (green box) and glycosaminoglycans that are integral to the extracellular matrix and glycocalyx (peach box). These represent a constant, host-derived, nutrient source for the colonic microbiota. **b,** Subfamily and substrates targeted by *B. theta* sulfatases. Green boxes identify the sulfatases which were analysed in this study. + and * in S1_4 correlate the substrate target with the specific sulfatase. Monosaccharide symbols are shown according to the Symbol Nomenclature for Glycan system[60].

Europe PMC Funders Author Manuscripts

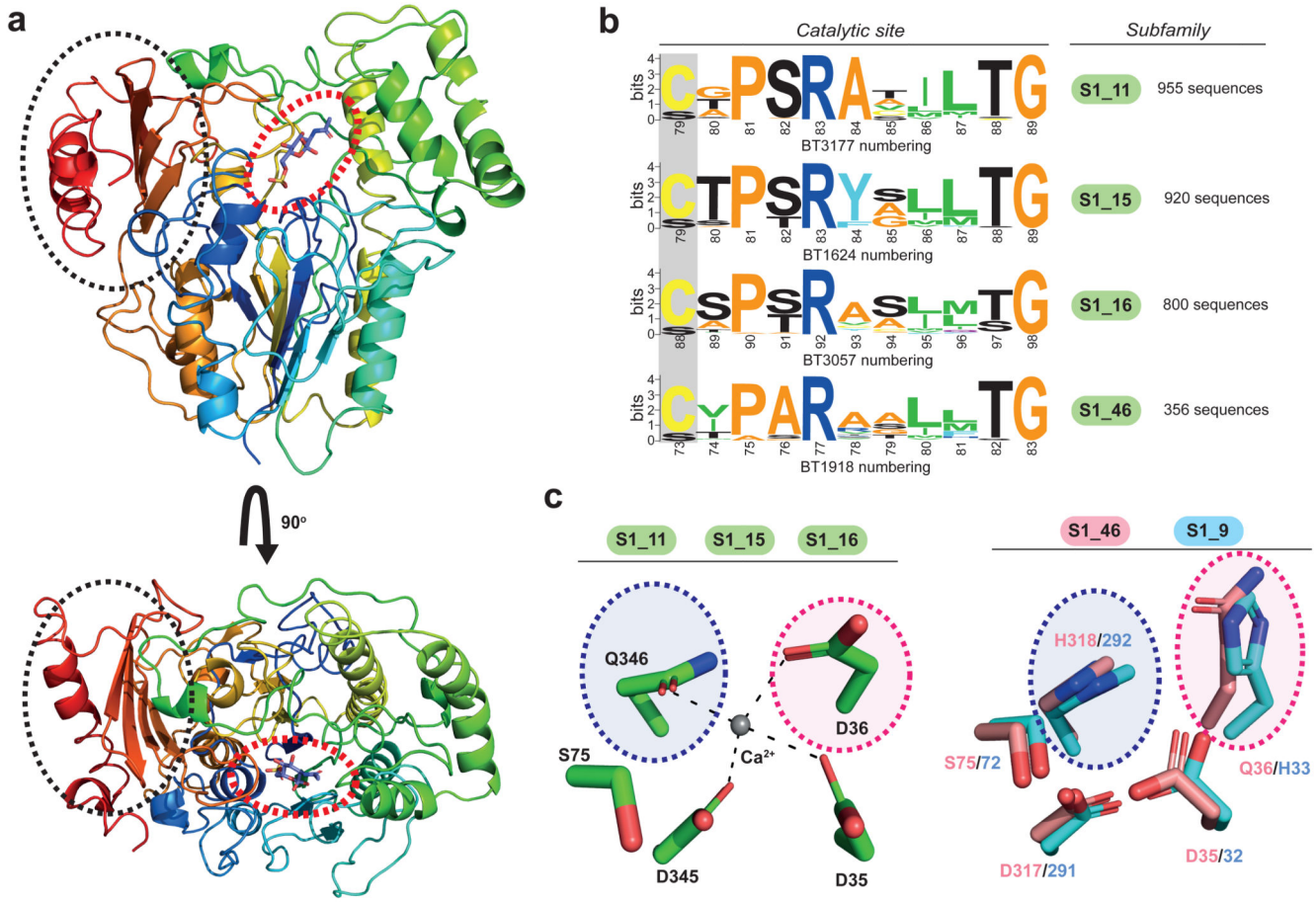


Fig. 2. S1 carbohydrate sulfatases share a conserved α/β/α fold, sulfate binding site, and catalytic apparatus.

a, Cartoon representation of BT3177$^{6S\text{-}GlcNAc}$ colour ramped from blue (N-terminal domain) to red (C-terminal domain). The black dashed circle indicates the β-sheet C terminal sub-domain and the red dashed circle the active site location on the core α/β/α domain. **b,** Conservation of the consensus active site sequence **C/S**-X-P/A-S/X-R for S1_11, S1_15, S1_16 and S1_46 subfamilies. The position of formylglycine installation is highlighted with a grey box. The number of sequences analysed with each subfamily is displayed on the right side if the respective subfamily number **c,** Stick representation of examples of the different sulfatases calcium-binding site. On left side (green), representation of BT3177$^{6S\text{-}GlcNAc}$ calcium site as an example of the typical site observed in most sulfatase structures described to date. On right side, display of alternate sites observed in BT1918$^{3S\text{-}GlcNAc}$ (salmon) and BT1596$^{2S\text{-}\Delta4,5UA}$ (cyan). The variable areas are highlighted in blue and pink dashed circles.

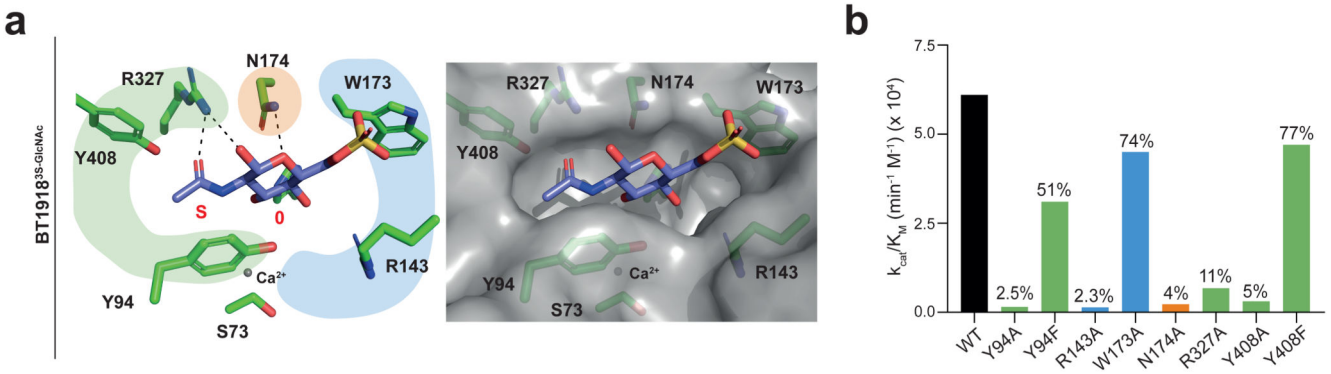


Fig. 3. S1_46 members from the HGM require recognition of the *N*-acetyl group for activity.

a, BT1918$^{3S\text{-}GlcNAc}$ structural data, modelled at 1.95 Å, as stick representation showing carbohydrate-binding interactions of the 0 subsite of BT1918$^{3S\text{-}GlcNAc}$ (left panel) and surface representation showing the 0 subsite pocket (right panel). The critical residues interacting with the *N*-acetyl group, the sulfate flanking residues, and sugar ring-only interactions are highlighted in green, blue and orange, respectively. **b,** Effects of alanine scanning on BT1918$^{3S\text{-}GlcNAc}$ catalysis. Percentage values above the bar indicate relative activity to wild-type (WT). Experiments are technical triplicates and calculated errors represent SEM.

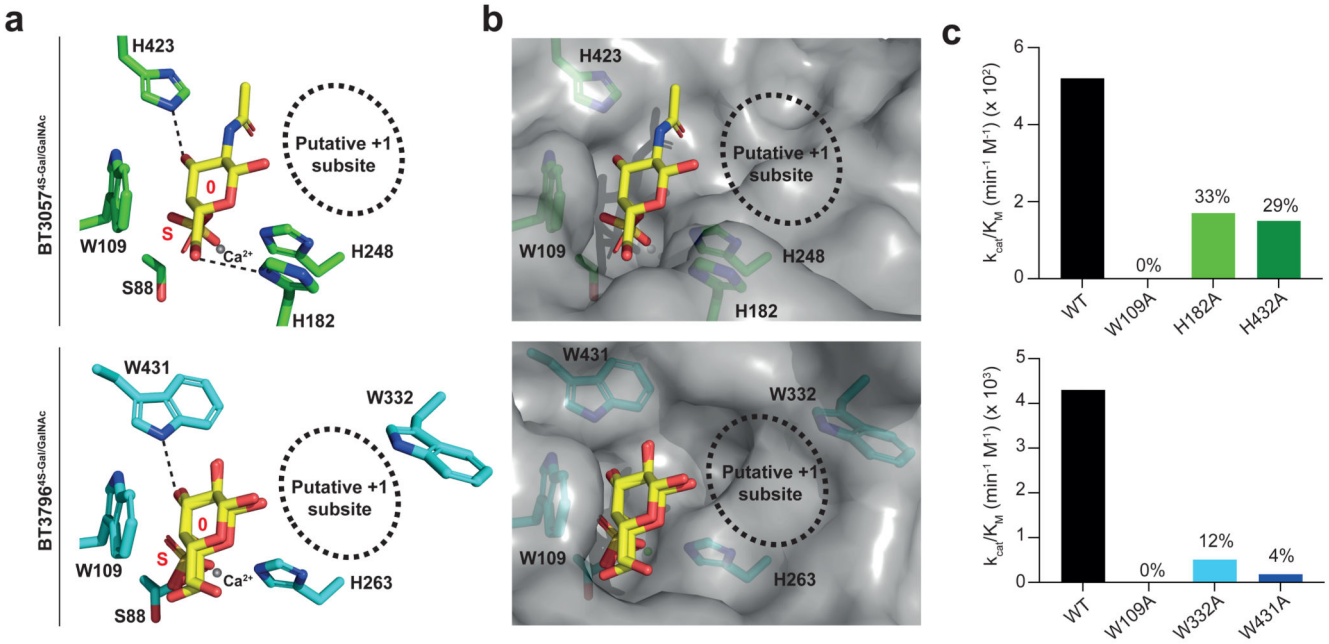


Fig. 4. Aromatic stacking analysis at the 0 subsite in S1_16 members from the HGM

a, Stick representations of carbohydrate-binding interactions of the 0 subsite BT3057$^{4S\text{-}Gal/GalNAc}$ and BT3796$^{4S\text{-}Gal/GalNAc}$, modelled at 1.91 Å and 1.50 Å, respectively. Although aromatic stacking is critical, hydrogen bonding to *O*3 via a secondary amine is also found in both structures. **b,** surface representations of the 0 subsite of BT3057$^{4S\text{-}Gal/GalNAc}$ (top) and BT3796$^{4S\text{-}Gal/GalNAc}$(bottom). **c,** Effects of alanine scanning on sulfatase activities with the percentages above the bar indicating the relative activity to wild-type (WT) enzyme. Experiments are technical triplicates and calculated errors represent SEM.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

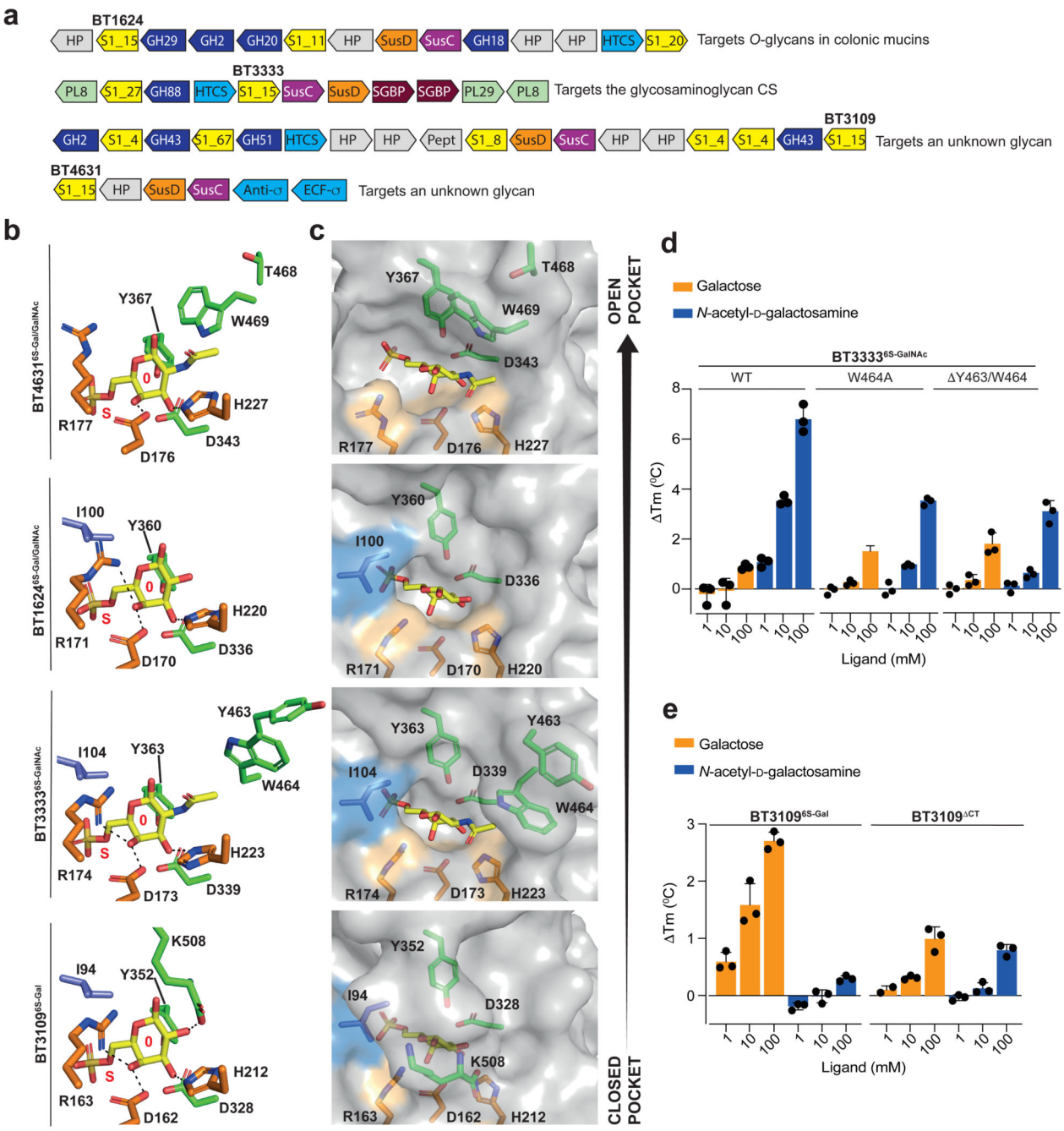


Fig. 5. Specificity of S1_15 subfamily members drives for 6S-Gal and 6S-GalNAc is determined by the openness of the S and 0 subsites

a, Schematic representation of the PULs encoding the described four S_15 sulfatases. HP (protein of unknown function), S1 (sulfatase S1 with the respective subfamily number superscript), GHXX (glycoside hydrolase with X representing the family number), PLXX (polysaccharide lyase with X representing the family number), HTCS (Hybrid two component receptor), ECF-σ (Extracytoplasmic factor sigma), anit-σ (Anti sigma factor), SusC (Starch utilisation system like C), and SusD (Starch utilisation system like D). **b,**

Europe PMC Funders Author Manuscripts

Stick and **c,** Surface representation showing the carbohydrate binding interactions of the 0 subsite of BT4631$^{6S\text{-}Gal/GalNAc}$, BT1624$^{6S\text{-}Gal/GalNAc}$, BT3333$^{6S\text{-}GalNAc}$, and BT3109$^{6S\text{-}Gal}$ (from top to bottom). BT4631$^{6S\text{-}Gal/GalNAc}$, BT1624$^{6S\text{-}Gal/GalNAc}$, and BT3109$^{6S\text{-}Gal}$ were modelled at 1.35 Å, 2.70 Å, and 1.75 Å, respectively. BT3333$^{6S\text{-}GalNAc}$ was solved previously with PDB code 6S20. The arrow on the right side indicates the degree of openness of the S and 0 subsites that increases from bottom to top and drives the altered substrate specificity. In panels b and c, the galacto-recognition residues have been coloured orange whilst the Ile, which is absent in BT4631$^{6S\text{-}Gal/GalNAc}$, is coloured in blue. **d,** DSF analysis of putative GalNAc specificity determinants (Y463 and W464) in BT3333$^{6S\text{-}GalNAc}$. **e,** DSF analysis of the effect of deleting the C-terminal extension (VEEEPLK) which drives specificity towards Gal in BT3109$^{6S\text{-}Gal}$. Experiments are technical triplicates and error bars represent SEM.

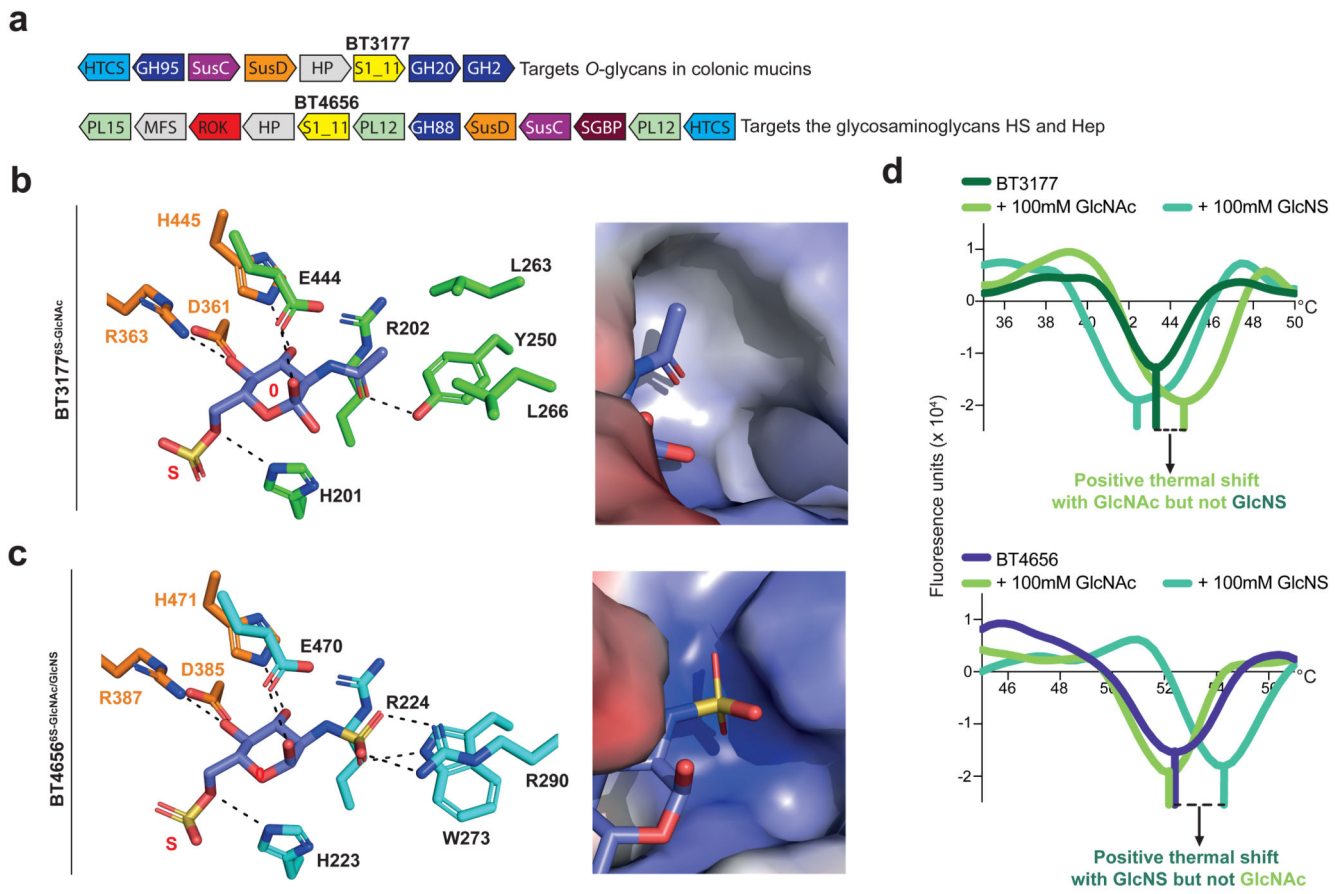


Fig. 6. Variations in the non-conserved region of S1_11 members, dictated by PUL context, drive recognition of *N*-sulfate or *N*-acetyl

a, Schematic representation of the PULs in which the two relevant S_11 sulfatases are encoded. HP (protein of unknown function), S1 (sulfatase S1 with the respective subfamily number superscript), GHXX (glycoside hydrolase with X representing the family number), PLXX (polysaccharide lyase with X representing the family number), HTCS (Hybrid two component receptor), SusC (Starch utilisation system like C), and SusD (Starch utilisation system like D). **b,** Carbohydrate interactions of BT3177$^{6S\text{-}GlcNAc}$, modelled at a resolution of 0.97 Å. **c,** Carbohydrate interactions of BT4656$^{6S\text{-}GlcNAc/GlcNS}$ (PDB code: 5G2V). In both panels the left is a stick representation of the carbohydrate interactions at the S and 0 subsites with the gluco-recognition triad shown in orange. The right panel is a surface representation with blue and red representing the positive and negative surface charge, respectively. **d,** DSF data showing the second derivative of thermal melting profiles with protein (no monosaccharide), 100 mM GlcNAc, and 100 mM GlcNS in the presence of 5 μM BT3177$^{6S\text{-}GlcNAc}$ (top), or 5 μM BT4656$^{6S\text{-}GlcNAc/GlcNS}$ (bottom). BT3177$^{6S\text{-}GlcNAc}$ interaction with GlcNAc is more thermoprotective than with GlcNS, whilst the reciprocal is true for BT4656$^{6S\text{-}GlcNAc/GlcNS}$, indicating different affinities by these enzymes, for these carbohydrates. Experiments are technical triplicates and error bars represent SEM.