HAL
open science

A single bacterial sulfatase is required for metabolism of colonic mucin O -glycans and intestinal colonization by a symbiotic human gut bacterium

Ana Luis, Chunsheng Jin, Gabriel Vasconcelos Pereira, Robert Glowacki, Sadie Gugel, Shaleni Singh, Dominic Byrne, Nicholas Pudlo, James London, Arnaud Baslé, et al.

To cite this version:

Ana Luis, Chunsheng Jin, Gabriel Vasconcelos Pereira, Robert Glowacki, Sadie Gugel, et al.. A single bacterial sulfatase is required for metabolism of colonic mucin O -glycans and intestinal colonization by a symbiotic human gut bacterium. 2020. hal-03025006

HAL Id: hal-03025006

https://hal.science/hal-03025006

Preprint submitted on 18 Dec 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A single bacterial sulfatase is required for metabolism of colonic mucin *O*-glycans and intestinal colonization by a symbiotic human gut bacterium

*[1,2]Ana S. Luis, [2]Chunsheng Jin, [1]Gabriel Vasconcelos Pereira, [1]Robert W. P. Glowacki, [1]Sadie Gugel, [1]Shaleni Singh, [3]Dominic P. Byrne, [1]Nicholas Pudlo, [3]James A London, [4]Arnaud Baslé, [5]Mark Reihill, [5]Stefan Oscarson, [3]Patrick A. Eyers, [6]Mirjam, Czjzek, [6]Gurvan Michel, [6]Tristan Barbeyron, [3]Edwin A Yates, [2]Gunnar C. Hansson, [2]Niclas G. Karlsson, *[,3]Alan Cartmell, *[1]Eric C. Martens

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

[2]*Department of Medical Biochemistry, Institute for Biomedicine, Sahlgrenska Academy, University of Gothenburg, Box 440, 405 30 Gothenburg, Sweden*

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

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

[5]*Centre for Synthesis and Chemical Biology, University College Dublin, Belfield, Dublin 4, Ireland.*

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

*Correspondence to:

emartens@umich.edu

Alan.Cartmell@liverpool.ac.uk

ana.luis@medkem.gu.se

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Summary

Humans have co-evolved with a dense community of microbial symbionts that inhabit the lower intestine. In the colon, secreted mucus creates a physical barrier that separates these microbes from the intestinal epithelium. Some gut bacteria are able to utilize mucin glycoproteins, the main mucus component, as a nutrient source. However, it remains unclear which bacterial enzymes initiate the degradation of the highly complex *O*-glycans found in mucins. In the colon, these glycans are heavily sulfated, but the specific sulfatases that are active on colonic mucins have not been identified. Here, we show that sulfatases are essential to the utilization of colonic mucin *O*-glycans by the human gut symbiont *Bacteroides thetaiotaomicron*. We have characterized the activity of 12 different sulfatases encoded by this species, showing that these enzymes collectively are active on all of the known sulfate linkages in colonic *O*-glycans. Crystal structures of 3 enzymes provide mechanistic insight into the molecular basis of substrate-specificity. Unexpectedly, we found that a single sulfatase is essential for utilization of sulfated *O*-glycans *in vitro* and also plays a major role *in vivo*. Our results provide insight into the mechanisms of mucin degradation by gut bacteria, an important process for both normal microbial gut colonization and diseases such as inflammatory bowel disease (IBD). Sulfatase activity is likely to be a keystone step in bacterial mucin degradation and inhibition of these enzymes may therefore represent a viable therapeutic path for treatment of IBD and other diseases.

Introduction

The human gut microbiota (HGM) significantly impacts several aspects of intestinal health and disease, including inflammatory bowel disease (IBD)[1] and colorectal cancer (CRC)[2]. In the colon, secreted mucus creates a physical barrier that separates gut microbes from the intestinal epithelium[3] preventing close contact that can lead to inflammation and eventual CRC if this barrier is either experimentally eliminated[4,5] or has reduced glycosylation[6-9]. A major component of the colonic mucus is mucin 2 (MUC2), a glycoprotein that contains up to 80% glycans by mass and more than 100 different glycan structures that are *O*-linked to serine or threonine residues[10]. Mucin glycosylation is variable along the gastrointestinal (GI) tract with a marked increase in sulfation in the

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

colon[11]. In mucins, *O*-linked sulfate may be attached to the 6-hydroxyl of *N*-acetyl-D-glucosamine (6S-GlcNAc) and non-reducing end D-galactose (Gal) sugars at hydroxyl positions 3-, 4- or 6- (3S-, 4S- and 6S-Gal, respectively)[11-13] (**Fig. 1a**). Sulfation often occurs as terminal caps that block enzymatic degradation of oligosaccharides. To degrade and utilize colonic mucin *O*-glycans, members of the HGM need to express appropriate carbohydrate sulfatases to remove these modifications. *Bacteroides thetaiotaomicron* (*Bt*) is a dominant member of the human gut microbiota that is able to utilize *O*-glycans as a sole nutrient source[14]. Underscoring the importance of sulfatases, *Bt* requires active sulfatases for competitive colonization of the wild-type mouse gut[15] and to induce inflammation in genetically-susceptible mice[16]. However, the specific sulfatases that mediate these effects remain unknown. Indeed, despite the critical roles of sulfatases in many biological processes, including several human diseases[17], a significant knowledge gap exists regarding the biochemical, structural and functional roles of these enzymes.

We hypothesized that specific *Bt* sulfatases play essential roles in initiation of *O*-glycan degradation. To test this, we measured the activities of 23 putative *Bt* sulfatases, determining that 12 enzymes are active on either model glycan substrates or purified colonic mucin *O*-glycans. Together with defining the specific activities of these enzymes, we determined the corresponding structures for 3 sulfatases, revealing the basis of substrate specificity. Using molecular genetics, we next assessed the contributions of these sulfatases to *Bt* fitness *in vitro* and *in vivo,* unexpectedly revealing that a single enzyme is essential for utilization of sulfated mucin *O*-glycans and plays a major role in competitive gut colonization. Identifying specific bacterial sulfatases that are critical for intestinal mucin degradation provides new potential targets, with a goal of blocking progression of diseases such as IBD and possibly other disorders that result from bacterial disruption of the mucus barrier.

Utilization of colonic mucins by HGM species

Several studies have identified HGM members that are able to utilize porcine gastric mucin *O*-glycans (gMO) as a sole carbon source[14,18,19]. However, this substrate does not adequately reflect the structural complexity of mucin *O*-glycans found in the

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

colon, especially those with increased sulfation that are lacking in gMO[11]. To identify HGM species that utilize sulfated colonic mucins, we measured the growth of 19 *Bacteroides* type strains, plus *Akkermansia muciniphil*a, on highly sulfated porcine colonic mucin oligosaccharides (cMO). We identified six *Bacteroides* strains that utilize cMO (**Fig. 1b, Extended Data Fig. 1**). Interestingly, two known mucin-degraders, *A. muciniphila*[20] and *B. massiliensis*[21], grew robustly on gMO but failed to utilize sulfated cMO as a substrate (**Extended Data Fig. 1**), highlighting the importance of employing colonic mucins as substrates to draw more physiologically-relevant conclusions about the metabolic targets of these organisms. *Bt*, the bacterium with the highest number of sulfatases (28), was one of the strains with the best growth on cMO (**Fig. 1b**), suggesting that some of these enzymes might play key roles in promoting this ability. Therefore, to understand the role of sulfatases in colonic mucin utilization by HGM bacteria we focused on the biochemical and genetic characterization of the *Bt* enzymes. See **Supplementary Discussion 1** for additional details of GI mucins and bacterial growth kinetics.

Substrate specificity of Bt sulfatases

Sulfatases are classified into four main families (S1 to S4) in the SulfAtlas database according to sequence similarity, catalytic mechanism and fold[22]. Family S1 is currently divided into 72 subfamilies (designated S1_X) and comprises the formylglycine sulfatases, which operate via a hydrolytic mechanism that utilizes a non-genetically coded formylglycine amino acid as its catalytic residue. In *Bt* and other anaerobic bacteria, this residue is introduced co-translationally by the anaerobic sulfatase maturating enzyme (anSME)[15], which converts a serine or cysteine, within the consensus sequence **C/S**-X-P/A/S-X-R, to formylglycine, which then serves as catalytic nucleophile[23]. The *Bt* genome encodes 28 S1 sulfatases classified into twelve different subfamilies (**Supplementary Table 1**). Four *Bt* sulfatases have been previously characterized and all act on glycosaminoglycans (GAGs) that are components of extracellular matrix (**Fig. 1c**)[24,25]. Interestingly, several of the uncharacterized S1 sulfatases are encoded within polysaccharide utilization loci (PULs) that are known to be upregulated *in vivo* or during growth on gMO[14] and encode other glycoside hydrolases enzymes potentially involved in degrading mucin *O*-glycans (**Extended Data Fig. 2**).

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To understand the role of sulfatases in mucin metabolism, we cloned and expressed in soluble form 23 of the remaining 24 uncharacterized sulfatases. The recombinant proteins were tested for activity against a panel of commercially available sulfated saccharides (**Supplementary Table 2**). In this initial screen, we identified activities for twelve sulfatases (**Fig. 1c, Extended Data Figs. 3, 4** and **Supplementary Table 3**). Among these, 5 of the enzymes represent the first activities reported for their respective subfamilies: two S1_20 members (BT1636 and BT1622) were determined to target 3S-Gal, with BT1622 also preferentially cleaving 3S-*N*-acetyl-D-galactosamine (3S-GalNAc), two S1_16 enzymes (BT3796 and BT3057) cleave 4S-Gal/4S-GalNAc and one S1_46 enzyme (BT1918) cleaves 3S-GlcNAc, using the *N*-acetyl group as an absolute specificity determinant. This represents the first report of a bacterial sulfatase active on 3S-GalNAc, indicating that this sulfation could exist as a yet unidentified modification of host glycans. Subsequently, we refer to these enzymes by their gene/locus tag number with the corresponding activity in superscript (*e.g.*, BT1636$^{3S\text{-}Gal}$).

In addition to assigning new catalytic activities associated with three subfamilies that previously lacked any characterization, we also identified sulfatases displaying novel activities inside previously characterized subfamilies. These include three S1_15 enzymes (BT1624$^{6S\text{-}Gal/GalNAc}$, BT3109$^{6S\text{-}Gal/GalNAc}$ and BT4631$^{6S\text{-}Gal/GalNAc}$) that extend this family, previously only known to include 6S-GalNAc sulfatases, to those cleaving 6S-Gal. Two members of S1_4 were active on 3S-Gal (BT4683$^{3S\text{-}Gal}$) or 6S-Gal (BT3487$^{6S\text{-}Gal}$), representing novel activities within this arylsulfatase subfamily. Finally, consistent with the activity previously described for S1_11 members, two enzymes were 6S-GlcNAc sulfatases (BT1628$^{6S\text{-}GlcNAc}$ and BT3177$^{6S\text{-}GlcNAc}$) (**Fig. 1c** and **Extended Data Fig. 3**).

Characterized sulfatases within the same subfamily, with the exception of two S1_4 members, cleaved the same sulfate ester linkages (**Fig. 1c**). However, despite these enzymes targeting the same linkages, their optimal activity depends on the surrounding glycan context. The activity of the 3S-Gal sulfatases is dependent on the linkage between Gal and GlcNAc with BT4683$^{3S\text{-}Gal}$ showing a preference for 3'-sulfate-*N*-acetyl-D-lactosamine (3'S-LacNAc, 3'S-D-Gal-β1,4-D-GlcNAc). In contrast, BT1622$^{3S\text{-}Gal/GalNAc}$ demonstrated enhanced activity towards 3'-sulfate-lacto-*N*-biose (3'S-LNB, 3'S-D-Gal-β1,3-D-GlcNAc) (**Extended Data Fig. 4a** and **Supplementary Table 3**).

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Furthermore, additional affinity/activity studies revealed that BT1622$^{3S\text{-}Gal/GalNAc}$ preferentially targets GalNAc and not Gal (**Extended data Fig. 4b,c** and **Supplementary Table 3**), suggesting that this sulfatase evolved to optimally target sulfate *O*3-linked to GalNAc. *Bt* sulfatase activity is also affected by the presence of terminal epitopes such as those that occur in Lewis antigens. Despite BT1636$^{3S\text{-}Gal}$ being equally active on 3'S-LacNAc and 3'S-LNB, this protein has a lower affinity and it is 100-fold less active when L-fucose (Fuc) is linked to GlcNAc (3'S-Lewis-a/x) (**Extended data Fig. 4a,b** and **Supplementary Table 3**). While BT1622$^{3S\text{-}Gal/GalNAc}$ is only weakly active on 3'S-Lewis-a antigen and not active at all on 3'S-Lewis-x, the reciprocal is true for BT4683$^{3S\text{-}Gal}$ (**Supplementary Table 3**). The subfamily S1_15 enzyme BT1624$^{6S\text{-}Gal/GalNAc}$ was only weakly active on 6S-Lewis-a/x antigens (**Extended Data Fig. 4a**), suggesting that these enzymes cannot accommodate Fuc linked to GlcNAc and that Fuc needs to be removed prior to sulfate cleavage. Additionally, affinity studies showed that BT3109$^{6S\text{-}Gal/GalNAc}$ has a strong affinity for Gal, while the previously characterized GAG sulfatase BT3333$^{6S\text{-}GalNAc}$ showed a preference for GalNAc, suggesting that optimal activity of S1_15 sulfatases likely depends on the glycan context and BT3333$^{6S\text{-}GalNAc}$ evolved to target sulfated linkages in GAGs, a substrate that contains GalNAc but not Gal.

Bt sulfatase activity on colonic mucin oligosaccharides (cMO)

We next tested the activity of *Bt* sulfatases on custom purified cMO, which we determined to contain at least 131 different oligosaccharides (**Supplementary Table 4)**. Only 4 of the 6 sulfatases tested displayed activity on cMOs (**Fig. 2a,b** and **Supplementary Table 4)**. BT1628$^{6S\text{-}GlcNAc}$ and BT3177$^{6S\text{-}GlcNAc}$ removed 6-*O*-sulfate from all GlcNAc structures but only when present at the non-reducing end of *O*-glycans confirming an exo-mode of action observed using commercial oligosaccharides (**Fig. 2** and **Extended Data Fig. 5a**). After incubation with BT1636$^{3S\text{-}Gal}$, we were able to detect 14 new oligosaccharides and an overall increase of non-sulfated glycans (**Fig. 2** and **Extended Data Fig. 5b)**. Compared to the non-enzyme treated control, 36 oligosaccharides could no longer be detected after incubation with BT1636$^{3S\text{-}Gal}$ (**Fig. 2a** and **Supplementary Table 4).** We determined the structures of 8 of these glycans and all present a terminal 3S-Gal (**Fig. 2)**. This sulfatase was active on 3'S-Gal-β1,3-GalNAc

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(core 1) and more complex sulfated structures built around other common core structures (**Fig. 2c)**, indicating that BT1636$^{3S\text{-}Gal}$ evolved to accommodate the various linkages and substitutions found in mucin *O*-glycans. After incubation with BT4683$^{3S\text{-}Gal}$, 9 oligosaccharides disappeared, indicating that this enzyme is active on a smaller subset of structures (**Fig. 2a** and **Supplementary Table 4)**. We were only able to determine the structure of one of those glycans and, unexpectedly, we found that this enzyme is endo-active on sialylated 3S-Gal (**Fig. 2b)**, which is consistent with the protein structural data discussed below. Additionally, the strong activity of BT1636$^{3S\text{-}Gal}$ on larger glycans with unknown structures (**Supplementary Table 4)** reveals that porcine cMOs are highly sulfated at the O3-position of galactose. Although previous studies have reported the presence of 3S-Gal in colonic mucins[26-28], the analysis of such complex samples is technically challenging and, until now, no precise enzymatic tools were available to probe these linkages. Together, these results represent the first report of HGM sulfatases active on colonic mucin *O*-glycans and highlight the possibility to use gut bacterial sulfatases as analytical tools in structural characterization of mucin glycans (see **Supplementary Discussion 2** for additional details of sulfatase activity on cMO).

Structural characterization of 3S-Gal/GalNAc sulfatases

To further understand the molecular details of carbohydrate recognition by S1 sulfatases, we determined the crystal structures of the three different 3S-Gal sulfatases that belong to two subfamilies. Consistent with previous structures of S1 sulfatases, all 3 enzymes display a $\alpha/\beta/\alpha$ topology with a C-terminal sub-domain, and the active site residues interacting with the sulfate group are fully conserved (**Extended data Fig. 6a,b**). The structure of BT1636$^{3S\text{-}Gal}$ in complex with the product LacNAc revealed that His177 coordinates with *O*4 of Gal and mutation to alanine ablates enzyme activity, suggesting that this residue is the major specificity determinant for Gal (presenting an axial *O*4) over glucose (equatorial *O*4) (**Fig. 3** and **Supplementary Table 3**). BT1636$^{3S\text{-}Gal}$ also makes strong interactions with O2 via R353 and E334. The essential His177 in BT1636$^{3S\text{-}Gal}$ is highly conserved (92%) within S1_20 sulfatases (**Extended data Fig. 7**). In BT1622$^{3S\text{-}Gal/GalNAc}$, mutation of the corresponding H176 to alanine causes a ~300-fold reduction in activity (**Supplementary Table 3**), further highlighting the importance of this residue in

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Gal recognition. In BT1622$^{3S\text{-}Gal/GalNAc}$, R353 and E334 are replaced by C357 and N334, amino acids that present shorter side chains that allow the accommodation of a *C*2-linked *N*-acetyl group found in GalNAc (**Fig. 3**). This observation is consistent with the ability of BT1622$^{3S\text{-}Gal/GalNAc}$ to cleave 3S-GalNAc and preferentially bind GalNAc over Gal, while BT1636$^{3S\text{-}Gal}$ only recognizes Gal (**Extended Data Fig. 4b,c**). Additionally, and consistent with *exo*-activity observed for both enzymes, the substrates are buried in a deep pocket and only *O1* is solvent exposed (**Fig. 3**).

For the S1_4 enzyme BT4683$^{3S\text{-}Gal}$, a structure solved in complex with LacNAc did not reveal any interaction between the protein and the O4 of Gal. In BT4683$^{3S\text{-}Gal}$, the interaction with Gal is driven by the residues R72 and E335, spatially equivalent to R353 and E334 in BT1636$^{3S\text{-}Gal}$, that form hydrogen bonds with *O2* of D-Gal (**Fig. 3**) and disruption of either of these residues eliminates activity (**Supplementary Table 3**). However, in BT4683$^{3S\text{-}Gal}$, a sulfatase that does not have any affinity for monosaccharides (**Extended data Fig. 4b**), the active site is located in an open cleft (**Fig. 3**) that allows the accommodation of additional substitutions on Gal (**Extended Data Fig. 6d**). This finding is consistent with the apparent endo-activity found using cMO. Together, these structures reveal the key specificity determinants in 3S-Gal/GalNAc sulfatases, highlighting that these enzymes have evolved to target sulfate groups in the different contexts in which they are found in complex host glycans. This is especially true for BT1636$^{3S\text{-}Gal}$ which utilizes high affinity interactions with both O2 and O4 to drive enhanced activity to remove terminal 3S-Gal linkages in cMOs (see **Supplementary Discussion 3** for additional details of sulfatase structural characterization and phylogeny).

Roles of sulfatases in B. thetaiotaomicron O-glycan utilization

Bt is able to utilize cMO as a sole carbon source (**Fig. 1b**), but the key enzymes involved in the degradation of these glycans remain unclear. Highlighting the importance of sulfatases to this symbiont's physiology, deletion of the gene encoding the only anaerobic sulfatase maturating enzyme (*anSME*) eliminates activation of all 28 S1 sulfatases[15] and the ability of *Bt* to grow efficiently on cMOs (**Fig. 4a**). Based on this observation we generated a series of strains with compounded gene deletions in which one or several groups of sulfatases were eliminated based on their biochemical activity.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Deletion of all 3S-Gal/GalNAc sulfatases (Δ*bt1636*$^{3S\text{-}Gal}$ + Δ*bt1622*$^{3S\text{-}Gal/GalNAc}$ + Δ*bt4683*$^{3S\text{-}Gal}$) resulted in a growth phenotype similar to Δ*anSME* (**Extended Data Fig. 8a**). Interestingly, we observed a similar growth defect when just BT1636$^{3S\text{-}Gal}$ was deleted, but not the other 3S-Gal sulfatases (**Fig. 4a** and **Extended Data Fig. 8a**), consistent with the prominent activity of the recombinant form of this enzyme on cMOs. In contrast, a strain with compounded deletions of eight other sulfatases (Δ*bt1622*$^{3S\text{-}Gal/GalNAc}$ + Δ*bt4683*$^{3S\text{-}Gal}$ + Δ*bt1624*$^{6S\text{-}Gal/GalNAc}$ + Δ*bt3109*$^{6S\text{-}Gal}$ + Δ*bt4631*$^{6S\text{-}Gal/GalNAc}$ + Δ*bt1628*$^{6S\text{-}GlcNAc}$ + Δ*bt3177*$^{6S\text{-}GlcNAc}$ + Δ*bt3051*$^{putative\_6S\text{-}GlcNAc}$) displayed a growth phenotype similar to wild-type (**Fig. 4a**), indicating that these enzymes are not essential for cMO utilization. However, a Δ*10X sulf* mutant, which included the deletion of BT1636$^{3S\text{-}Gal}$ and the two 4S-Gal/GalNAc sulfatases, showed a similar growth defect as Δ*anSME* and Δ*bt1636*. Complementation of this and other loss of function mutants with only *bt1636*$^{3S\text{-}Gal}$ restored growth on cMO to levels similar to wild-type (**Fig. 4a** and **Extended Data Fig. 8a**). Cellular localization experiments revealed that BT1636$^{3S\text{-}Gal}$ is located at the cell surface in *Bt* (**Fig. 4b**) and together these data suggest that this single 3S-Gal sulfatase is a critical cell surface enzyme involved in the utilization of sulfated *O*-glycans that are prominent in the colon.

To further investigate the role of BT1636$^{3S\text{-}Gal}$ in *O*-glycan utilization, we analyzed the oligosaccharides present in the culture supernatant of the wild-type and Δ*bt1636*$^{3S\text{-}Gal}$ strains after growth on cMO. Consistent with a robust ability of *Bt* to degrade diverse colonic *O*-glycans, no oligosaccharides were detected in wild-type supernatant (**Supplementary Table 5**). Compared to the cMO used as substrate (control), the supernatant of Δ*bt1636*$^{3S\text{-}Gal}$ showed a 20-fold accumulation of terminal 3S-Gal capped glycans, suggesting that these could not be degraded. Indeed, the three most common 3S-Gal structures detected in the Δ*bt1636*$^{3S\text{-}Gal}$ supernatant accounted for 50% of the total oligosaccharides detected in this sample (**Fig. 4c**). These results reveal that deletion of BT1636$^{3S\text{-}Gal}$ results in loss of the ability to utilize 3S-Gal *O*-glycans. Interestingly, 49 of the 72 glycans detected in Δ*bt1636*$^{3S\text{-}Gal}$ supernatant were not present in the control sample (**Supplementary Table 5**), suggesting that although the mutant does not grow well on cMO it is able to modify these oligosaccharides to generate new glycans

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(**Supplementary discussion 5**). These data, combined with the cell surface location of BT1636$^{3S\text{-}Gal}$, suggest that this sulfatase is required early in *O*-glycan catabolism likely by cleaving 3S-Gal from *O*-glycans prior to importing them into the periplasm where these oligosaccharides will be sequentially degraded by additional enzymes and also serve as cues for activating transcription of the *O*-glycan PULs (**Extended Data Fig. 2)**. Although *Bt* does encode two additional 3S-Gal/GalNAc sulfatases, the low activity of these additional sulfatases on cMOs (**Fig. 2**) and their likely periplasmic location, suggests why these enzymes cannot compensate for loss of BT1636$^{3S\text{-}Gal}$. Interestingly, all of the *Bacteroides* species able to utilize cMO (**Fig. 1a**) have homologues of BT1636$^{3S\text{-}Gal}$, suggesting that this activity plays a key role in mucin utilization by other HGM members (**Supplementary Table 6**).

Finally, to investigate the requirement for specific sulfatases *in vivo*, we utilized gnotobiotic mice in which we competed individual mutants against the wild-type strain to evaluate their colonization fitness. It has been reported that mouse colonic Muc2 prominently displays 6S-GlcNAc modifications[29]. However, mutants lacking either just the two active 6S-GlcNAc sulfatases (Δ*6S-GlcNAc* double mutant), or these two enzymes, another putative 6S-GlcNAc sulfatase and all three 6S-Gal/GalNAc sulfatases (Δ*6S-GlcNAc* + Δ*6S-Gal/GalNAc* hexa mutant), competed equally with wild-type (**Extended Data Fig. 8c**), suggesting that neither of these two sulfatase activities are essential determinants *in vivo*. A significant defect was observed with a mutant lacking all 3S-Gal/GalNAc sulfatases (**Fig. 4d**). The fitness defect was exacerbated by eliminating 3S-Gal/GalNAc and 6S-GlcNAc sulfatase activities together (Δ*3S-Gal/GalNAc* + Δ*6S-GlcNAc*) (**Fig. 4d**), suggesting that they synergize *in vivo*. Consistent with its prominent role in cMO utilization *in vitro*, a mutant lacking just BT1636$^{Gal\text{-}3Sulf}$ displayed a significant defect that was similar or slightly more severe compared to the Δ*3S-Gal/GalNAc* mutant when competed with the wild-type strain (**Fig. 4d**), further suggesting that this enzyme plays an essential role in gut colonization by allowing *Bt* to access 3S-Gal *O*-glycans (see **Supplementary Discussion 4-5** for additional details regarding growth and competition of mutants *in vitro* and *in vivo*).

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Conclusion

To degrade the complex *O*-glycans found in mucins some HGM bacteria have evolved complex arsenals of degradative enzymes which include diverse sulfatases. Disarming all of the sulfatases in *Bt* via *anSME* deletion results in drastically reduced competitive colonization (**Extended Data Fig. 8c**)[15] and an inability to elicit colitis in an animal model of IBD[16]. While these findings support a critical role for active sulfatases in both fitness and promoting inflammation, they provide no insight into the complexity of catalytic modifications carried out by these enzymes. In this study, we reveal that *Bt* has a robust ability to grow on highly sulfated mucin oligosaccharides from colonic tissue and that it possesses active sulfatases capable of removing sulfate groups in all contexts in which sulfation is known to occur in mucin, including novel specificities. Surprisingly, we found that a single key sulfatase is essential for growth on colonic mucin *O*-glycans. This cell surface enzyme removes 3-sulfate capping Gal allowing the degradation of these glycans by additional enzymes. This critical role of BT1636$^{3S\text{-}Gal}$ supports the conclusion that keystone steps exist in the complex pathway of mucin degradation. Further delineation of these critical steps, along with identification of the corresponding enzymes(s), are a prerequisite to modulating such events and potentially inhibiting mucin-degrading activities in bacteria that contribute to disease.

Methods

Recombinant Protein Production

Genes were amplified by PCR using the appropriate primers and the amplified DNA cloned in pET28b using *NheI/XhoI* restriction sites or pETite (Expresso™ T7 cloning and expression system, Lucigen) generating constructs with either N- or C-terminal His_6 tags (**Supplementary Table 7**). The catalytic serine was mutated to cysteine since *Escherichia coli* in only able to convert cysteine to formylglycine. Recombinant genes were expressed in *Escherichia coli* strains BL21 (DE3) or TUNER (Novagen), containing the appropriate recombinant plasmid, and cultured to mid-exponential phase before induction with 1 mM (BL21(DE3)) or 0.2 mM (TUNER) of isopropyl β-D-1-

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

thiogalactopyranoside; cells were cultured for another 16 h at 16°C and 180 rpm. Recombinant protein were purified to >90% electrophoretic purity by immobilized metal ion affinity chromatography using a cobalt-based matrix (Talon, Clontech) and eluted with imidazole as described previously[24]. For the proteins selected for structural studies, another step of size exclusion chromatography was performed using a Superdex 16/60 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 respective molar extinction coefficient. When necessary, proteins were then concentrated by centrifugaton using a molecular mass cutoff of 30 kDa.

Site-Directed Mutagenesis

Site-directed mutagenesis was conducted using the PCR-based QuikChange kit (Stratagene) and conducted according to the manufacturer's instructions, using the appropriate plasmid as the template and primers (**Supplementary Table 8**). All mutations were confirmed by DNA sequencing.

Sources of purified carbohydrates

All carbohydrates were from Sigma, Carbosynth or Dextra Laboratories. All other chemical reagents were purchased from Sigma. The 3S-GalNAc was chemically synthesized as previously described[30].

Mucin purification

Gastric mucin oligosaccharides (gMO) were purified from commercial available porcine gastric mucins (type III, Sigma) as previously described[14]. Colonic mucins oligosaccharides (cMO) were purified from pig distal pig colons and rectum. Briefly, the tissue was open and the fecal contents were carefully removed. The mucosa was scrapped off and mucus was extracted by homogenizing the tissue in at least 5 times volume of extraction buffer (6 M guanidine chloride, 5 mM EDTA, 10 mM NaH_2PO_4, pH 6.5) and slow stirring at 4ºC for 16 h. The solution was spun down at 15,000 rpm and 10ºC for 30 min and supernatant was discharged. The pellets were resuspended in

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

extraction buffer and the process was repeated until the supernatant was clear for at least two extractions. After the extraction the mucins were solubilized by reducing the disulfide bonds. The pellets were resuspended in fresh reduction buffer (6 M guanidine chloride, 0.1 M Tris, 5 mM EDTA, pH 8.0) containing 25 mM of 1,4-dithiothreitol and slowly stirred at 37ºC for 5 h. After this incubation, 62.5 mM of iodoacetamide were added and the solution was stirred slowly in the dark at room temperature for 16 h. The solution was centrifuged at 10,000 rpm at 4ºC for 30 min and the supernatant containing the solubilized mucins was extensively dialysed into water. Samples were dissolved into 100 mM Tris-HCl pH 8.0 containing 1 mg/ml of trypsin and incubated slowly stirring at 37ºC for 16 h. The glycans were β-eliminated by adding 0.1 M NaOH and 1 M $NaBH_4$ and incubate the solution at 65ºC for 18 h. After cooling the solution to room temperature, the pH was adjusted to 7.0 with concentrated HCl and extensively dialysed in water. The released porcine colonic mucin glycans were recovered by lyophilization the solution until completely dry and used in further experiments.

HPLC and TLC sulfatase enzymatic assays

The sulfatase activity screen against commercially available sulfated oligosaccharides (**Supplementary Table 2**) was performed with 1 μM of recombinant enzyme and 1 mM of substrate in 10 mM MES pH6.5 with 5 mM CaCl2 for 16h at 37ºC. Sulfated *N*-acetyl-D-lactosamine and lacto-*N*-biose were generated by incubating the respective sulfated Lewis antigens with 1 μM of α-1,3/1,4-fucosidase BT1625[31] in the same conditions. Reactions were analysed by thin layer chromatography (TLC). Briefly, 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% H3PO3, 100 ml of ethyl acetate and 2 g diphenylamine) and heated at 100°C for 20 min. When relevant, the enzymatic activity was confirmed by high-performance anionic exchange chromatography (HPAEC) with pulsed amperometric detection using standard methodology. The sugars (reaction substrate/products) were bound to a Dionex CarboPac P100 column and eluted with an initial isocratic flow of 10 mM NaOH during 20 min then a gradient of 10-100 mM of NaOH for 20 min at a flow rate of 1.0 ml min^{-1}. The

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

reaction products were identified using the appropriated standards. All experiments were performed in triplicate.

Liquid Chromatograph-Electrospray Ionization Tandem Mass Spectrometry

Enzymatic reactions of sulfatases in colonic mucin oligosaccharides and culture supernatant were cleaned up with graphitized carbon[32]. Reactions with sulfated defined saccharides were reduced and desalted. Briefly, reactions were dried in Speed vac, reconstituted in 20 μL of 50 mM NaOH and 500 mM $NaBH_4$ and incubated at 50ºC for 3 h. Reactions were cool down on ice, neutralized with 1 μL of glacial acetic acid and desalted using a cation exchange column containing AG®50W-X8 resin. All cleaned and desalted reactions were reconstituted in water before analysis by liquid chromatograph-electrospray ionization tandem mass spectrometry (LC-ESI/MS). The oligosaccharides were separated on a column (10 cm × 250 μm) packed in-house with 5 μm porous graphite particles (Hypercarb, Thermo-Hypersil, Runcorn, UK). The oligosaccharides were injected on to the column and eluted with a 0-40 % acetonitrile gradient in 10 mM ammonium bicarbonate over 46 min at a flow rate of 10 μl/min.. A 40 cm × 50 μm i.d. fused silica capillary was used as transfer line to the ion source. Samples were analyzed in negative ion mode on a LTQ linear ion trap mass spectrometer (Thermo Electron, San José, CA), with an IonMax standard ESI source equipped with a stainless steel needle kept at –3.5 kV. Compressed air was used as nebulizer gas. The heated capillary was kept at 300°C, and the capillary voltage was –33 kV. Full scan (*m/z* 380-2,000, two microscan, maximum 100 ms, target value of 30,000) was performed, followed by data-dependent MS^2 scans (two microscans, maximum 100 ms, target value of 10,000) with normalized collision energy of 35%, isolation window of 2.5 units, activation q=0.25 and activation time 30 ms). The threshold for MS^2 was set to 300 counts. Data acquisition and processing were conducted with Xcalibur software (Version 2.0.7). Glycans were identified from their MS/MS spectra by manual annotation and validated by available structures stored in Unicarb-DB database (2020-01 version) (**Supplementary Fig. 1**)[33]. *O*-Glycan structural characterization was based on diagnostic fragment ions[34].The schematic glycosidic or cross-ring cleavages were assigned according to the Domon and Costello nomenclacture[35]. For comparison of glycan abundance between samples, the

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

individual glycan structures were quantified relative to the total content by integration of the extracted ion chromatogram peak area using Progenesis QI. The area under the curve (AUC) of each structure was normalized to the total AUC and expressed as a percentage.

Microfuidic-based enzymatic desulfation assays

Sulfated carbohydrates were labelled at their reducing end with BODIPY which has a maximal emission absorbance of ~503 nm, which can be detected by the EZ Reader via LED-induced fluorescence. Non-radioactive mobility shift carbohydrate sulfation assays were optimised in solution with a 12-sipper chip coated with CR8 reagent using a PerkinElmer EZ Reader II system using EDTA-based separation buffer. This approach allows real-time kinetic evaluation of substrate de-sulfation[36]. 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 product and unsulfated substrate 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 after assembly in a 384-well plate in a final volume of 80 µl in the presence of substrate concentrations between 0.5 and 20 µM with 100 mM Bis-Tris-Propane, 150 mM NaCl, 0.02% (v/v) Brij-35 and 5 mM $CaCl_2$. The degree of de-sulfation was calculated by peak integration using EZ Reader software, which measures the sulfated carbohydrate : unsulfated carbohydrate ratio at each individua 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 and to ensure assay linearity. k_{cat}/K_M values, using the equation $V_0=(V_{max}/K_M)/S$, were determined by linear regression analysis with GraphPad Prism software. Substrate concentrations were varied to ensure assay linearity, and substrate concentrations present were significantly $<K_M$.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

NMR desulfation assays

NMR experiments, monitoring the de-sulfation of 6S-D-galactose and 6S-*N*-acetyl-D-galactosamine, were conducted in D_2O with 50 mM sodium phosphate, pH 7.0, supplemented with 150 mM NaCl at 25°C on a 800MHz Bruker Avance III spectrometer equipped with a TCI CryoProbe and a 600MHz Bruker Avance II+ spectrometer, also fitted with a TCI CryoProbe. 1D and 2D proton and TOCSY spectra (mixing time 80 ms) were measured using standard pulse sequences provided by the manufacturer. Spectra were processed and analysed using TopSpin 3.4A and TopSpin 4.0 software (Bruker). Galactose integrals were recorded directly for the C(6)H_2-OH peak within the region 3.694 to 3.721ppm, referenced to the combined C(2) peaks of D-galactose and 6S-D-galactose with in the region 3.415 to 3.475ppm. Similarly, 6S-*N*-acetyl-D-galactosamine integrals were recorded directly for the C(6)H_2-OH peak within the region 3.674 to 3.747ppm, referenced to the combined C(4) peaks for *N*-acetyl-D-galactosamine and 6S-*N*-acetyl-D-galactosamine in the region 3.925 to 3.968ppm.

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[37] 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 to 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. When BT1622$^{3S\text{-}Gal/GalNAc}$ and BT1636$^{3S\text{-}Gal}$ were assessed against 3'S-LacNAc and 3'S-LNB 100 mM Hepes (pH 7.0) was employed instead of BTP, although no

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

difference in the T_m value of the proteins was observed. Three independent assays were performed for each protein and protein ligand combination.

Glycan labelling

Sulfated saccharide samples were labelled according to a modification of the method previously described reporting the formation of N-glycosyl amines for 4,6-O-benzilidene protected D-gluopyranose monosaccharides with aromatic amines[38]. Briefly, the lyophilised sugar (1 mg) was dissolved in 0.50 ml anhydrous methanol in a 1.5 ml screw-top PTFE microcentrifuge tube. 0.1 mg, BODIPY-FL hydrazide (4,4–difluoro-5,7–dimethyl-4-bora-3a,4a–diaza-s-indacene-3-propionic acid hydrazide, $\lambda_{ex./em.\ 505/513}$, extinction coefficient 80,000 M^{-1} cm^{-1}) was added and the mixture vortexed (1 min), then incubated in darkness at 65ºC for 24 h. The products were then cooled and a portion purified by TLC on silica coated aluminium plates and developed with methanol or 1:1 v/v ethyl acetate/methanol to provide R_f values suitable to allow separation of unreacted label from labelled glycan product. The unreacted BODIPY-FL label (orange on the TLC plate) was identified by reference to a lane containing the starting material (BODIPY-FL hydrazide) on the TLC plate, allowing differentiation from the putative labelled product (also orange). This latter band was scraped from the plates and extracted in fresh methanol (2 x 0.5 ml), spun for 3 min at 13,000 x *g* and the supernatant was recovered and dried (rotary evaporator) to recover the fluorescent-coloured product (bright green when dissolved in aqueous solution), which was then employed in subsequent experiments.

Anaerobic bacterial culture and genetic manipulation

All strains were anaerobic grown at 37 °C in a chamber (10% H_2, 5% CO_2, and 85% N_2; Coy Manufacturing, Grass Lake, MI). *Bacteroides* type strains were culture in either tryptone-yeast extract-glucose medium (TYG), brain heart infusion medium or minimal medium (MM) containing an appropriate carbon source. *Bacteroides massilliensis* and *Akkermansia muciniphila* were culture as described before[21,39]. *Bt* strains containing specific gene deletions or inactivated versions of enzymes (BT1636$^{3S\text{-}Gal}$ S77A) were made by counterselectable allelic exchange as previously described[40]. Complemention

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of deletion strains was performed using pNBU2 vector as previously described[40], containing a constitutive promotor used previously[41]. All primers used to generate the mutants and complementation are listed in **Supplementary Table 9**. Growth of the WT and mutants was measured on an automated plate reader by increase in absorbance at 600 nm in 96-well plates containing 200 μl of minimal media mixed with the respective filter-sterilised (monosaccharide and gMO) or autoclave-sterilised cMO as described before[39]. To achieve consistent growth, all carbon sources were used at 5 mg/ml with exception of gMO that was added in a final concentration of 10 mg/ml. All growth curves presented are averages and s.e.m of three technical replicates.

Crystallization of carbohydrate sulfatases

After purification, all proteins were carried forward in the same eluent as used for the size exclusion chromatography (see Recombinant Protein Production). Sparse matrix screens were set up in 96-well sitting drop TTP Labtech plates (400-nL drops). Initial hits crystals for all proteins were obtained between 20 and 35 mg/ml. For BT1622$^{3S\text{-}Gal/GalNAc}$ and BT1636$^{3S\text{-}Gal}$ wildtype *Bt* variants were used, having a Ser at the catalytic formylglycine position, whilst for for BT4683$^{3S\text{-}Gal}$ the S73C mutant was used. BT1622$^{3S\text{-}Gal/GalNAc}$ with 20 mM LNB crystallised in 20% PEG 3350 and 0.2 M sodium citrate tribasic dihydrate. BT1636$^{3S\text{-}Gal}$ with 20 mM LacNAc crystallised in 40% MPD and 0.2 M sodium cacodylate pH 6.5. for BT4683$^{3S\text{-}Gal}$ with 20 mM LacNAc crystallised in 20% PEG 3350, 0.2 M sodium iodide and BTP pH 8.5. All crystals were cryoprotected with the addition of the ligand they were crystallised with plus 20% PEG 400 and 20% glycerol was used as the cryoprotectant for BT4683$^{3S\text{-}Gal}$ and BT1622$^{3S\text{-}Gal/GalNAc}$, respectively. No cryoprotectant was added to BT1636$^{3S\text{-}Gal}$ crystals. Data were collected at Diamond Light Source (Oxford) on beamlines I03, I04, I04-1 and I24 at 100 K. The data were integrated with XDS[42], or Xia2[43] 3di or 3dii and scaled with Aimless[44]. 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[45] for all proteins except BT1622$^{3S\text{-}Gal/GalNAc}$. The phase problem for BT1622$^{3S\text{-}Gal/GalNAc}$ was initially solved using Molrep[46] and BT1636$^{3S\text{-}Gal}$ as the search model. This gave a partial solution, which could not be fully solved due to twinning. An acceptable model of BT1622$^{3S\text{-}Gal/GalNAc}$ was

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

559 constructed to be used to better solve the phase problem and the molecular replacement
560 was re-performed. Models underwent recursive cycles of model building in Coot[47] and
561 refinement cycles in Refmac5[48]. Bespoke ligands were generated using JLigand[49]. The
562 models were validated using Coot[47] and MolProbity[50]. Structural figures were made using
563 Pymol and all other programs used were from the CCP4[51] and CCP4i2 suite[52]. The data
564 processing and refinement statistics are reported in **Supplementary Table 10**.
565
566 **Immunolabelling of BT1636 in *Bt* cell surface**
567 For the fluorescence microscopy, *Bt* cells (Wild type (Δ*tdk*) and Δ*bt1636*$^{3S\text{-}Gal}$) were grown
568 to early exponential phase (Abs_{600nm} 0.25–0.35) in rich TYG medium. One ml of the
569 cultures was collected, centrifuged at 13,000 x *g*, and subsequently washed three times
570 in MM with no carbon source. *Bt* cells incubated with cMO for four hours and fixed in 4.5%
571 formalin overnight at 4ºC with gentle rocking. Cells were stained with a polyclonal
572 antibody raised in rabbit against purified recombinant BT1636 (BT1636^{Ab}, Cocalico
573 Biologicals) and detected with an Alexa Fluor® 488-conjugated goat anti-rabbit IgG
574 secondary antibody (Molecular Probes). Images were taken with Zeiss Apotome using
575 the same exposure time between samples.
576
577 **Gnotobiotic Mouse Experiments**
578 All experiments involving animals, including euthanasia via carbon dioxide asphyxiation,
579 were approved by the University Committee on Use and Care of Animals at the University
580 of Michigan (NIH Office of Laboratory Animal Welfare number A3114-01) and overseen
581 by a veterinarian. Groups of 3 to 5, 6-8 week old germfree Swiss Webster mice were
582 randomly assigned to each experiment. 7 days prior gavage the animals diet was
583 switched to a fiber-free diet (Envigo-Teklad TD 130343) that was maintained through all
584 the experiment. At day 0, mice were gavage with equal amount of *Bt* WT strain and mutant
585 and fecal samples were collected at day 2 and every 5 days until day 42. At the end-point
586 of the experiment distal small intestine and cecal contents were also collected. The
587 bacteria gDNA extraction and quantification by qPCR of the relative abundance of each
588 strain on the various samples was carried out as described previously[39].
589

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Phylogenetic analysis

To maximise sequence coverage, and avoid repetition, we selected 800 and 920 representative sequences of subfamily S1_20 (composed of 1356 sequences) and S1_4 (composed of 1895 sequences), respectively. The sequences were aligned by MAFFT v.7[53] using L-INS-i algorithm. The multiple sequence alignment was visualized by Jalview software v.11.0[54] and non-aligned regions were removed. 404 and 364 positions were used for the S1_4 and S1_20 phylogeny, respectively. Phylogeny was made using RAxML v. 8.2.4[55]. The phylogenetic tree was build with the Maximum Likelihood method[56] and the LG matrix as evolutive model[57] 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 1,000 resamplings of the dataset[58]. All the final global phylogenetic trees were obtained with MEGA v.7[59]. Fifteen S1_0 sequences from the sulfAtlas database were used as an outgroup.

Quantification and statistical analysis

For *in vivo* competitions, when three or more fecal samples were collected, Student's t tests (one-tailed, paired) were performed for each time point in Excel. When necessary, the statistical analysis for remaining samples in stated in the respective figure legend.

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: 7ANB, 7ANA, 7AN1 and 7ALL. The LC-MS/MS raw files and annotated structures are submitted to the Glycopost (https://glycopost.glycosmos.org/preview/12430260615f9d5733a1a5d, code 1955) and Unicarb-DB database, respectively.

Code availability statement: No new codes were developed or compiled in this study

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Competing interests statement

The authors declare no competing interests.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement N° 748336. This work was supported by National Institutes of Health grants (DK118024 and DK125445 awarded to ECM, U01AI095473 awarded to GCH), the European Research Council ERC (694181), The Knut and Alice Wallenberg Foundation (2017.0028), Swedish Research Council (2017-00958) 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. Mass spectrometry of glycans was performed in the Swedish infrastructure for biologic mass spectrometry (BioMS) supported by the Swedish Research Council. We are also grateful for Dr. Erwan Corre's help regarding bioinformatics analyses (ABIMS platform, Station Biologique de Roscoff, France).

Author contributions

ASL, AC and ECM designed experiments and wrote the manuscript. ASL and AC cloned, generated proteins mutants, expressed and purified sulfatases. ASL and AC performed enzyme assay, binding and kinetic analyses with assistance of DPB and PAE. MR and SO synthesized the O3-sulfated N-acetylgalactosamine. EAY labelled glycan substrates. AC grew, solved, collected and analysed crystallographic data with assistance from AB. AC and JAL carried out NMR kinetic analyses. CJ, ASL, GCH and NGK performed and interpreted analytical glycobiology experiments. ASL, GP, RWPG, SG, SS and NAP performed bacterial growth experiments and analysed *in vivo* competition data. MC, GM and TB performed sulfatase phylogenetic analyses. All authors read and approved the manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure legends

Figure 1. Bacterial growth on colonic mucin and *Bt* sulfatase activities. a, Schematic representation of mucin *O*-glycans and relevant terminal epitopes (dashed boxes). Sugars are shown according to the Symbol Nomenclature for Glycan system[60]. **b**, Growth of *Bacteroides* type strains and A*kkermansia muciniphila* on colonic mucin *O*-glycans (cMO) and number of respective encoded S1 sulfatases. The bars represent the average of two independent experiments with different batches of cMO. Bacterial species able to utilize gastric mucin glycans are highlighted in blue. Maximum absorbance is the difference of the maximum absorbance value (Abs_{600nm}) for each culture and the initial absorbance at time 0 (T_{0h}). Graphic shows the example of growth curves for *B. fragilis* (*Bf*), *B. thetaiotaomicron* (*Bt*), *B. vulgatus* (*Bv*) and *B. massilliensis* (*Bm*). **c**, Phylogeny of *Bt* sulfatases showing the 28 S1 sulfatases and their respective substrates where known, including this study. Enzymes are color coded according the respective subfamilies with sulfatases characterized in this study highlighted in bold. * indicates sulfatase activity previously characterized and arrows point the substrate preferentially targeted by the respective enzyme. Sulfatases on a shared branch that share more than 86% and 39-58% of sequence identity are highlighted in blue and green background, respectively. Data from biological replicates n = 3 to 6 and error bars denote s.e.m..

Figure 2. Activity of *Bt* sulfatases on colonic mucin *O*-glycans. a, Representation of *O*-glycans detected by mass spectrometry in cMO (control) and after sulfatase treatment from the lower (top) to the higher (bottom) mass range. **b,** Relative abundance and putative structures for the specific m/z shown in panel a. Remaining structures are shown in **Extended figure 5a**. **c,** Schematic representation of the putative structures that were not detected after treatment with $BT1636^{3S\text{-}Gal}$.

Figure 3. Crystal structures of 3S-Gal/GalNAc sulfatases. a, Schematic representation of the residues interacting with targeted sugars, including the putative catalytic residues (in dark red), the calcium ion (grey sphere) and subsites S, 0 and +1 highlighted in red. $BT1636^{3S\text{-}Gal}$ and $BT4683^{3S\text{-}Gal}$ in complex with LacNAc (D-Gal-β1,4-

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

D-GlcNAc) and BT1622$^{3S\text{-}Gal/GalNAc}$ in complex with GalNAc. **b,** Surface representation of the active pocket. The equivalent Gal/GalNAc specificity residues in BT1636$^{3S\text{-}Gal}$ and BT4683$^{3S\text{-}Gal}$ are highlighted in red and blue. The open active site of BT4683$^{3S\text{-}Gal}$ is highlighted in purple. In all structures the amino acids and ligands are represented as stick.

Figure 4. BT1636$^{3S\text{-}Gal}$ activity is required for the utilization of cMO and competitive fitness *in vivo*. **a**, Growth of *Bt* wild-type Δ*tdk* (WT), different sulfatase gene-deletion mutants (named "Δ*btXXXX*") and strains complemented with *bt1636$^{3S\text{-}Gal}$* on colonic or gastric mucin O-glycans (cMO and gMO, respectively) (line represents the average of biological replicates (n = 3) and error bars denote s.e.m.) **b,** Immunofluorescent and differential interference contrast (DIC) microscopy of *Bt* WT and sulfatase mutant staining with polyclonal antibody (Ab) against BT1636$^{3S\text{-}Gal}$ (green) and DNA staining with DAPI (blue). **c,** Relative abundance of different *O*-glycans detected by mass spectrometry in Δ*bt1636$^{3S\text{-}Gal}$* culture supernatant or cMO in minimal media without bacteria (control), after 96h in anaerobic conditions. The mass and associated structure of the 3 more abundant glycans in both samples are shown. **d,** *in vivo* competitions in gnotobiotic mice (n = 5-9 separate mice from two separate experiments, except competition of the Δ*6S-GlcNAc* mutant that showed no defect in one experiment) fed fiber-free diet and inoculated with WT and mutants. The fecal relative abundance of each strain was determined along the time course and in small intestine and cecum at day 42 (experimental endpoint). The relative abundance in each mouse is represented in the respective light colour. The error bars denote s.e.m. Significant differences between wild-type and mutant strain were compared at each time point using student's t-test (paired, one tail) and * indicates sample days in which the mutant was significantly different ($p < 0.01$) from the wild-type.

Extended Data 1. Growth of *Bacteroides* type strains and A*kkermansia muciniphila* in different mucin *O*-glycans. The graphics show the growth of strains able to utilize *O*-glycans in minimal media containing the indicated carbon source (biological replicates

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

n =3, error bars denote the s.e.m.). cMO, colonic mucin *O*-glycans; gMO, gastric mucin *O*-glycans.

Extended Data 2. Schematic representation of polysaccharide utilization loci (PULs) encoding sulfatases (sulf). Genes are colour coded according to the predicted function of the respective proteins. Glycoside hydrolases (GH) in known families are indicated by GHXX or GH*, where XX and * indicates the respective family number or non-classified, respectively.

Extended Data 3. Enzymatic screen of *Bt* sulfatases using sulfated monosaccharides. Recombinant enzymes (1 μM) were incubated with 1 mM of substrate in 10 mM MES pH6.5 with 5 mM $CaCl_2$ for 16 h at 37ºC. Reactions were analyzed by thin layer chromatography (left side) or HPAEC with pulsed amperometric detection (right side). Control reactions without sulfatases were carried out in the same conditions. The standards in TLC and HPAEC-PAD are labelled on the left side and top, respectively. The different panel represent activities found for sulfatases targeting: **(a)** 4S-Gal/GalNAc; **(b)** 6S-Gal/GalNAc; **(c)** 6S-GlcNAc and; **(d)** 3S-GlcNAc. The data shown are a representative from biological replicates (n = 3).

Extended Data 4. Activity and affinity of sulfatases to targeted substrates. a, Recombinant enzymes (1 μM) were incubated with 1 mM of substrate in 10 mM MES pH6.5 with 5 mM $CaCl_2$ for 16h at 37 ºC. Sulfated disaccharides were generated by adding 1 μM of a characterized α1,3/1,4-fucosidase (BT1625) in the enzymatic reaction. Control reactions without sulfatases were carried in the same conditions. Samples were analysed by mass spectrometry and the intensity of the substrate and reaction products was used for comparison of the relative abundance of these sugars after incubation with the respective enzymes. **b,** Affinity studies looking at the effect of ligand binding on the melting temperature of 3S and 6S-Gal sulfatases. All reactions were performed in 100 mM BTP, pH 7.0 with 150 mM NaCl. For sample melting temperatures see **Table S11**. **c,** Activity of 3S-Gal/GalNAc sulfatases (10 μM) against 3S-GalNAc (10 mM). Reactions

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

were performed in 10 mM Hepes, pH 7.0, with 150 mM NaCl and 5 mM $CaCl_2$. The data shown are one representative from the biological replicates conducted (n = 3).

Extended Data 5. Activity of *Bt* sulfatases against colonic mucin *O*-glycans (cMO) analysed by mass spectrometry. a, Relative abundance of defined oligosaccharides after incubation of cMO with different sulfatases or no enzyme (control). The putative structure for the different mass is shown on the right side of the graphic. The reactions were performed with 1 μM of enzyme and 0.5% cMO in 10 mM MES pH 6.5 with 5 mM $CaCl_2$ for 16 h at 37ºC. **b,** Relative abundance of structures detected in different samples organized by sulfate-linkage (top panel) or presence of one or several sugar substitutions such as sulfate, sialic acid and fucose (bottom panel). The colour-coded bars represent the relative abundance and the total number of the structures containing the specific linkage/substitution. The complete dataset is provided in **Supplementary Table 4**.

Extended Data 6. Schematic representation of 3S-Gal/GalNAc sulfatases. a, Cartoon representation colour ramped from blue (α/β/α N-terminal domain) to red (β-sheet C-terminal domain). **b**, Overlay of the active site S residues of $BT1636^{3S\text{-}Gal}$ (green) $BT1622^{3S\text{-}Gal/GalNAc}$ (blue) and $BT4683^{3S\text{-}Gal}$ (pink). The putative catalytic residues are shown in bold. The calcium ion is represented as a grey sphere and its polar interactions indicated as dashed lines. **c,** Ligand density of maps for LacNAc in $BT1636^{3S\text{-}Gal}$ and $BT4683^{3S\text{-}Gal}$, and GalNAc in $BT1622^{3S\text{-}Gal/GalNAc}$, contoured at 1σ (0.33 e/A^3, 0.37 e/A^3 and 0.18e/A^3, respectively); **d,** Docking of putative structures of *O*-glycans targeted by $BT4683^{3S\text{-}Gal}$ using the LacNAc as reference point showing that this structure can accommodate a sialic acid in -1 subsite and additional sugars in positive subsites (left hand side). The docking sugars are schematic shown as sticks (middle panel) and schematic represented inside the dashed box (right hand side). Using the LacNAc product as an 'anchor' additional sugars were built in manually with Coot 0.9 and regularized to low energy conformations.

Extended Data 7. Phylogenetic tree of S1_20 and S1_4 sulfatases. The radial trees were constructed using the branched trees shown in **Supplementary Figs. 2 and 3**. For

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

clarity, all labels and sequence accession codes have been omitted. Red filled circles designate sequences from *B. thetaiotaomicron* sulfatases. The residue is written in black without any attributes if present in the sequence, in grey and italics if the residue is mutated to any type in that sequence, or to a specific residue type if given in brackets. **a,** Radial representation of the phylogenetic tree constructed with representative sequences of the sulfatase S1_4 subfamily. The colour code is given as a pattern of presence or absence of the residues R72, E335 and W505, which are crucial in substrate recognition by BT4683 (acc-code Q89YP8, coloured red). A grey X in italics specifically designates that the residue W505 is absent in that sequence, and no obvious orthologous residue can be found from the alignment. **b,** Radial representation of the phylogenetic tree constructed with representative sequences of the sulfatase S1_20 subfamily. The colour code is given as a pattern of presence or absence of the residues E100, Q173 H177, E334, R353, which are crucial in substrate recognition by BT1636 (acc-code Q8A789, coloured red). A grey X in italics specifically designates that the residue E100 is absent in that sequence, and no obvious ortologous residue can be found from the alignment.

Extended Data 8. Sulfatase activity is required for growth in cMO and *in vivo* fitness. a, Growth curves of *Bt* wild-type Δ*tdk* (WT), different sulfatase mutants (Δ*btXXX*) and complemented strains on glucose, colonic or gastric mucin *O*-glycans (cMO and gMO, respectively). The curves represent the average of biological replicates (n = 3) and the error bars denote s.e.m. **b,** Relative abundance of oligosaccharides detected by mass spectrometry in culture supernatant of WT and Δ*bt1636*$^{3S\text{-}Gal}$ after growth in cMO for 96h at anaerobic conditions. The control corresponds to cMO incubated in the same conditions without bacterium. The colours represent the relative abundance of structures grouped according to the presence of epitopes (sulfate, fucose and sialic acid) and the numbers represent the total number of structures that contain the respective substitution. **c**, Colonization of gnotobiotic mice fed a fiber-free diet by *Bt* WT and mutants lacking the full (Δ*anSME*, no S1 sulfatases active) or specific sulfatase activity (Δ*6S-GlcNAc* and Δ*6S-GlcNAc*+Δ*6S-Gal/GalNAc*). The fecal relative abundance of each strain was determined in regular intervals until day 42. The relative abundance of time 0 represents the abundance in gavaged inoculum. At the experimental endpoint the relative abundance

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

was also determined in small intestine and cecum. The graphics represent the average of n =3 and the error bars denote the s.e.m. The relative abundance in each individual animal is represented in a lighter colour in each of the respective graphics.

References

1 Packey, C. D. & Sartor, R. B. Commensal bacteria, traditional and opportunistic pathogens, dysbiosis and bacterial killing in inflammatory bowel diseases. *Current opinion in infectious diseases* **22**, 292-301 (2009).

2 Sears, C. L. & Garrett, W. S. Microbes, Microbiota, and Colon Cancer. *Cell Host Microbe* **15**, 317-328, doi:10.1016/j.chom.2014.02.007 (2014).

3 Johansson, M. E. *et al.* The inner of the two Muc2 mucin-dependent mucus layers in colon is devoid of bacteria. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 15064-15069, doi:10.1073/pnas.0803124105 (2008).

4 Van der Sluis, M. *et al.* Muc2-deficient mice spontaneously develop colitis, indicating that Muc2 is critical for colonic protection. *Gastroenterology* **131**, 117-129, doi:Doi 10.1053/J.Gastro.2006.04.020 (2006).

5 Velcich, A. *et al.* Colorectal cancer in mice genetically deficient in the mucin Muc2. *Science* **295**, 1726-1729, doi:10.1126/science.1069094 (2002).

6 Bergstrom, K. *et al.* Core 1- and 3-derived O-glycans collectively maintain the colonic mucus barrier and protect against spontaneous colitis in mice. *Mucosal immunology*, doi:10.1038/mi.2016.45 (2016).

7 Larsson, J. M. *et al.* Altered O-glycosylation profile of MUC2 mucin occurs in active ulcerative colitis and is associated with increased inflammation. *Inflamm Bowel Dis* **17**, 2299-2307, doi:10.1002/ibd.21625 (2011).

8 Fu, J. *et al.* Loss of intestinal core 1-derived O-glycans causes spontaneous colitis in mice. *J Clin Invest* **121**, 1657-1666, doi:10.1172/JCI45538 (2011).

9 Kudelka, M. R. *et al.* Cosmc is an X-linked inflammatory bowel disease risk gene that spatially regulates gut microbiota and contributes to sex-specific risk. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 14787-14792, doi:10.1073/pnas.1612158114 (2016).

10 Larsson, J. M., Karlsson, H., Sjovall, H. & Hansson, G. C. A complex, but uniform O-glycosylation of the human MUC2 mucin from colonic biopsies analyzed by nanoLC/MSn. *Glycobiology* **19**, 756-766, doi:10.1093/glycob/cwp048 (2009).

11 Holmen Larsson, J. M., Thomsson, K. A., Rodriguez-Pineiro, A. M., Karlsson, H. & Hansson, G. C. Studies of mucus in mouse stomach, small intestine, and colon. III. Gastrointestinal Muc5ac and Muc2 mucin O-glycan patterns reveal a regiospecific distribution. *American journal of physiology. Gastrointestinal and liver physiology* **305**, G357-363, doi:10.1152/ajpgi.00048.2013 (2013).

12 Thomsson, K. A. *et al.* Detailed O-glycomics of the Muc2 mucin from colon of wild-type, core 1- and core 3-transferase-deficient mice highlights differences

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

compared with human MUC2. *Glycobiology* **22**, 1128-1139, doi:10.1093/glycob/cws083 (2012).

13 Robbe, C. *et al.* Evidence of regio-specific glycosylation in human intestinal mucins - Presence of an acidic gradient along the intestinal tract. *Journal of Biological Chemistry* **278**, 46337-46348, doi:DOI 10.1074/jbc.M302529200 (2003).

14 Martens, E. C., Chiang, H. C. & Gordon, J. I. Mucosal Glycan Foraging Enhances Fitness and Transmission of a Saccharolytic Human Gut Bacterial Symbiont. *Cell Host Microbe* **4**, 447-457, doi:DOI 10.1016/j.chom.2008.09.007 (2008).

15 Benjdia, A., Martens, E. C., Gordon, J. I. & Berteau, O. Sulfatases and a radical S-adenosyl-L-methionine (AdoMet) enzyme are key for mucosal foraging and fitness of the prominent human gut symbiont, Bacteroides thetaiotaomicron. *J Biol Chem* **286**, 25973-25982, doi:10.1074/jbc.M111.228841 (2011).

16 Hickey, C. A. *et al.* Colitogenic Bacteroides thetaiotaomicron Antigens Access Host Immune Cells in a Sulfatase-Dependent Manner via Outer Membrane Vesicles. *Cell Host Microbe* **17**, 672-680, doi:10.1016/j.chom.2015.04.002 (2015).

17 Diez-Roux, G. & Ballabio, A. Sulfatases and human disease. *Annu Rev Genomics Hum Genet* **6**, 355-379, doi:10.1146/annurev.genom.6.080604.162334 (2005).

18 Wlodarska, M. *et al.* Indoleacrylic Acid Produced by Commensal Peptostreptococcus Species Suppresses Inflammation. *Cell Host Microbe* **22**, 25-37 e26, doi:10.1016/j.chom.2017.06.007 (2017).

19 Tramontano, M. *et al.* Nutritional preferences of human gut bacteria reveal their metabolic idiosyncrasies. *Nat Microbiol* **3**, 514-522, doi:10.1038/s41564-018-0123-9 (2018).

20 Derrien, M., Vaughan, E. E., Plugge, C. M. & de Vos, W. M. Akkermansia muciniphila gen. nov., sp. nov., a human intestinal mucin-degrading bacterium. *Int J Syst Evol Microbiol* **54**, 1469-1476 (2004).

21 Pudlo, N. A. *et al.* Symbiotic Human Gut Bacteria with Variable Metabolic Priorities for Host Mucosal Glycans. *mBio* **6**, e01282-01215, doi:10.1128/mBio.01282-15 (2015).

22 Barbeyron, T. *et al.* Matching the Diversity of Sulfated Biomolecules: Creation of a Classification Database for Sulfatases Reflecting Their Substrate Specificity. *PLoS ONE* **11**, e0164846, doi:10.1371/journal.pone.0164846 (2016).

23 Benjdia, A. *et al.* Anaerobic sulfatase-maturating enzymes, first dual substrate radical S-adenosylmethionine enzymes. *J Biol Chem* **283**, 17815-17826, doi:10.1074/jbc.M710074200 (2008).

24 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* **114**, 7037-7042, doi:10.1073/pnas.1704367114 (2017).

25 Ndeh, D. *et al.* Metabolism of multiple glycosaminoglycans by Bacteroides thetaiotaomicron is orchestrated by a versatile core genetic locus. *Nature communications* **11**, 646, doi:10.1038/s41467-020-14509-4 (2020).

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

26 Thomsson, K. A., Karlsson, H. & Hansson, G. C. Sequencing of sulfated oligosaccharides from mucins by liquid chromatography and electrospray ionization tandem mass spectrometry. *Analytical chemistry* **72**, 4543-4549, doi:10.1021/ac000631b (2000).

27 Thomsson, K. A., Backstrom, M., Holmen Larsson, J. M., Hansson, G. C. & Karlsson, H. Enhanced detection of sialylated and sulfated glycans with negative ion mode nanoliquid chromatography/mass spectrometry at high pH. *Analytical chemistry* **82**, 1470-1477, doi:10.1021/ac902602e (2010).

28 van der Post, S. & Hansson, G. C. Membrane protein profiling of human colon reveals distinct regional differences. *Mol Cell Proteomics* **13**, 2277-2287, doi:10.1074/mcp.M114.040204 (2014).

29 Tobisawa, Y., Imai, Y., Fukuda, M. & Kawashima, H. Sulfation of colonic mucins by N-acetylglucosamine 6-O-sulfotransferase-2 and its protective function in experimental colitis in mice. *J Biol Chem* **285**, 6750-6760, doi:10.1074/jbc.M109.067082 (2010).

30 Egan, M., Jiang, H., O'Connell Motherway, M., Oscarson, S. & van Sinderen, D. Glycosulfatase-Encoding Gene Cluster in Bifidobacterium breve UCC2003. *Appl Environ Microbiol* **82**, 6611-6623, doi:10.1128/AEM.02022-16 (2016).

31 Briliute, J. *et al.* Complex N-glycan breakdown by gut Bacteroides involves an extensive enzymatic apparatus encoded by multiple co-regulated genetic loci. *Nat Microbiol* **4**, 1571-1581, doi:10.1038/s41564-019-0466-x (2019).

32 Packer, N. H., Lawson, M. A., Jardine, D. R. & Redmond, J. W. A general approach to desalting oligosaccharides released from glycoproteins. *Glycoconjugate journal* **15**, 737-747, doi:10.1023/a:1006983125913 (1998).

33 Hayes, C. A. *et al.* UniCarb-DB: a database resource for glycomic discovery. *Bioinformatics* **27**, 1343-1344, doi:10.1093/bioinformatics/btr137 (2011).

34 Everest-Dass, A. V., Abrahams, J. L., Kolarich, D., Packer, N. H. & Campbell, M. P. Structural feature ions for distinguishing N- and O-linked glycan isomers by LC-ESI-IT MS/MS. *J Am Soc Mass Spectrom* **24**, 895-906, doi:10.1007/s13361-013-0610-4 (2013).

35 Domon, B. & Costello, C. E. Structure elucidation of glycosphingolipids and gangliosides using high-performance tandem mass spectrometry. *Biochemistry* **27**, 1534-1543, doi:10.1021/bi00405a021 (1988).

36 Byrne, D. P. *et al.* New tools for carbohydrate sulfation analysis: heparan sulfate 2-O-sulfotransferase (HS2ST) is a target for small-molecule protein kinase inhibitors. *Biochem J* **475**, 2417-2433, doi:10.1042/BCJ20180265 (2018).

37 Byrne, D. P. *et al.* cAMP-dependent protein kinase (PKA) complexes probed by complementary differential scanning fluorimetry and ion mobility-mass spectrometry. *Biochem J* **473**, 3159-3175, doi:10.1042/BCJ20160648 (2016).

38 Das, T. M., Rao, C. P. & 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. *Carbohydrate research* **334**, 261-269, doi:10.1016/s0008-6215(01)00202-6 (2001).

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

39 Desai, M. S. *et al.* A Dietary Fiber-Deprived Gut Microbiota Degrades the Colonic Mucus Barrier and Enhances Pathogen Susceptibility. *Cell* **167**, 1339-1353 e1321, doi:10.1016/j.cell.2016.10.043 (2016).

40 Koropatkin, N. M., Martens, E. C., Gordon, J. I. & Smith, T. J. Starch catabolism by a prominent human gut symbiont is directed by the recognition of amylose helices. *Structure* **16**, 1105-1115, doi:10.1016/j.str.2008.03.017 (2008).

41 Degnan, P. H., Barry, N. A., Mok, K. C., Taga, M. E. & Goodman, A. L. Human gut microbes use multiple transporters to distinguish vitamin B(1)(2) analogs and compete in the gut. *Cell Host Microbe* **15**, 47-57, doi:10.1016/j.chom.2013.12.007 (2014).

42 Kabsch, W. Xds. *Acta Crystallogr D Biol Crystallogr* **66**, 125-132, doi:10.1107/S0907444909047337 (2010).

43 Winter, S. E., Lopez, C. A. & Baumler, A. J. The dynamics of gut-associated microbial communities during inflammation. *EMBO Rep* **14**, 319-327, doi:10.1038/embor.2013.27 (2013).

44 Evans, P. R. & Murshudov, G. N. How good are my data and what is the resolution? *Acta Crystallogr D Biol Crystallogr* **69**, 1204-1214, doi:10.1107/S0907444913000061 (2013).

45 Long, F., Vagin, A. A., Young, P. & Murshudov, G. N. BALBES: a molecular-replacement pipeline. *Acta Crystallogr D Biol Crystallogr* **64**, 125-132, doi:10.1107/S0907444907050172 (2008).

46 Vagin, A. & Teplyakov, A. Molecular replacement with MOLREP. *Acta Crystallogr D Biol Crystallogr* **66**, 22-25, doi:10.1107/S0907444909042589 (2010).

47 Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* **66**, 486-501, doi:10.1107/S0907444910007493 (2010).

48 Murshudov, G. N. *et al.* REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr D Biol Crystallogr* **67**, 355-367, doi:10.1107/S0907444911001314 (2011).

49 Lebedev, A. A. *et al.* JLigand: a graphical tool for the CCP4 template-restraint library. *Acta Crystallogr D Biol Crystallogr* **68**, 431-440, doi:10.1107/S090744491200251X (2012).

50 Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* **66**, 12-21, doi:10.1107/S0907444909042073 (2010).

51 Collaborative computational project, n. The CCP4 suite: Programs for protein crystallography. *Acta Crystallographica* **D50**, 760-763 (1994).

52 Potterton, L. *et al.* CCP4i2: the new graphical user interface to the CCP4 program suite. *Acta Crystallogr D Struct Biol* **74**, 68-84, doi:10.1107/S2059798317016035 (2018).

53 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* **30**, 3059-3066, doi:10.1093/nar/gkf436 (2002).

54 Clamp, M., Cuff, J., Searle, S. M. & Barton, G. J. The Jalview Java alignment editor. *Bioinformatics* **20**, 426-427, doi:10.1093/bioinformatics/btg430 (2004).

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

55 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313, doi:10.1093/bioinformatics/btu033 (2014).

56 Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution* **17**, 368-376, doi:10.1007/BF01734359 (1981).

57 Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol Biol Evol* **25**, 1307-1320, doi:10.1093/molbev/msn067 (2008).

58 Felsenstein, J. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution* **39**, 783-791, doi:10.1111/j.1558-5646.1985.tb00420.x (1985).

59 Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol* **33**, 1870-1874, doi:10.1093/molbev/msw054 (2016).

60 Neelamegham, S. *et al.* Updates to the Symbol Nomenclature for Glycans guidelines. *Glycobiology* **29**, 620-624, doi:10.1093/glycob/cwz045 (2019).

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

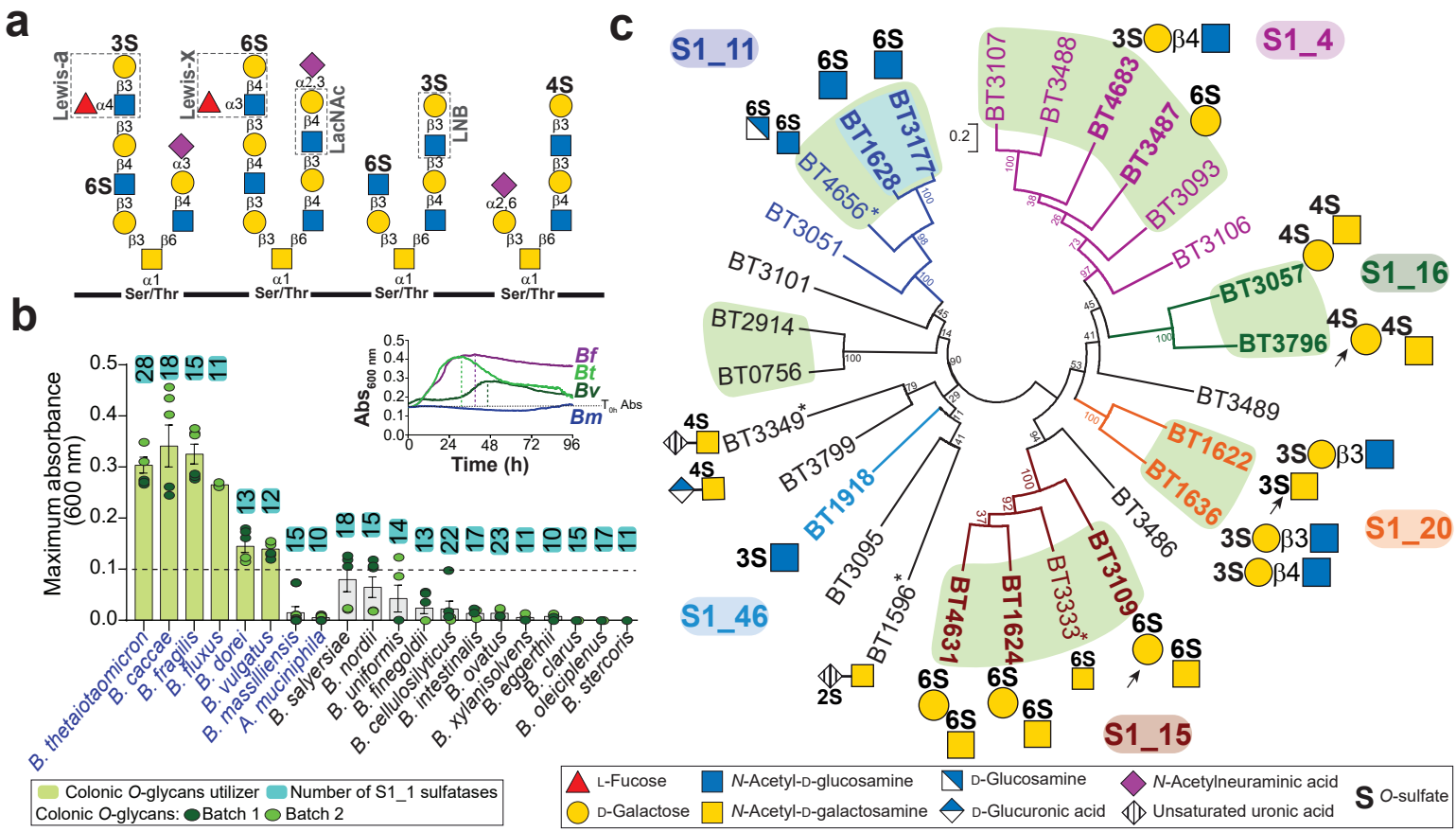


Fig 1.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

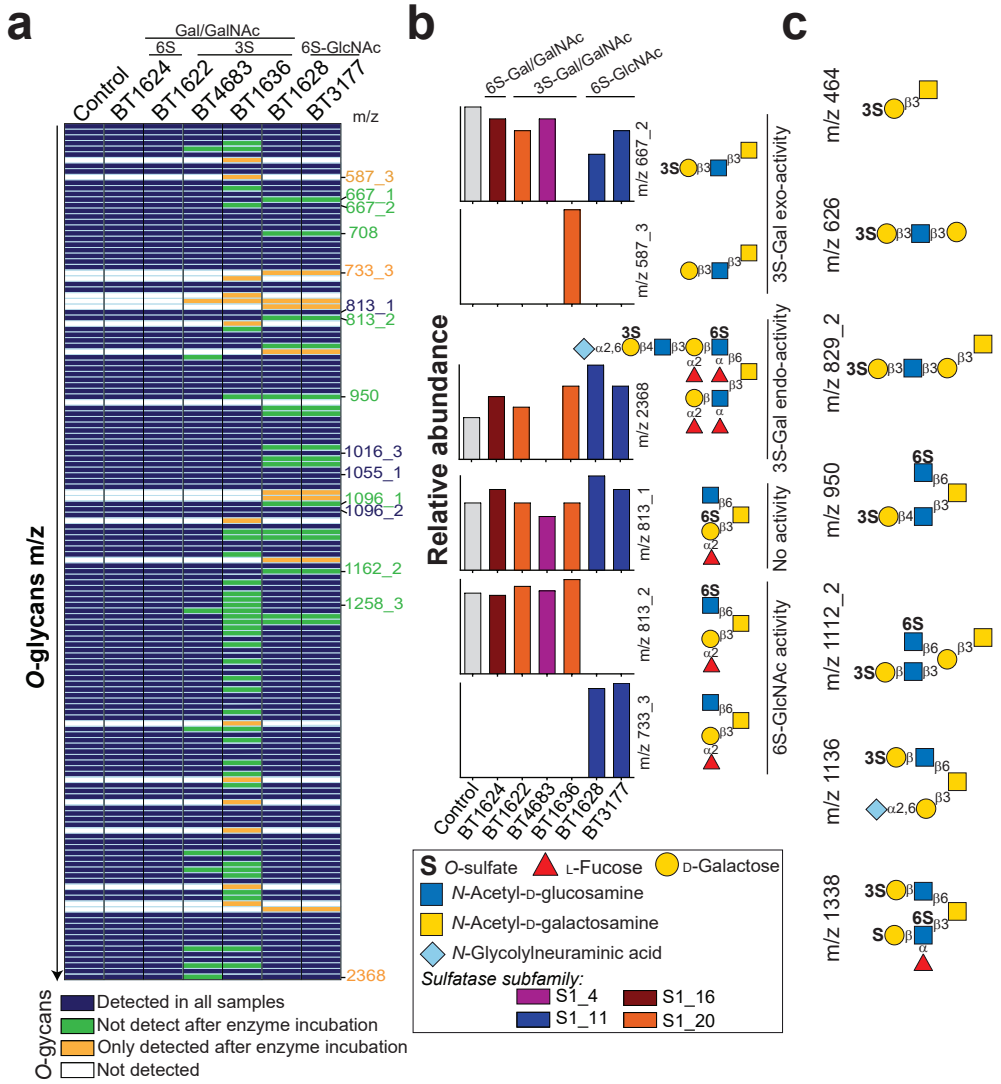


Fig 2.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

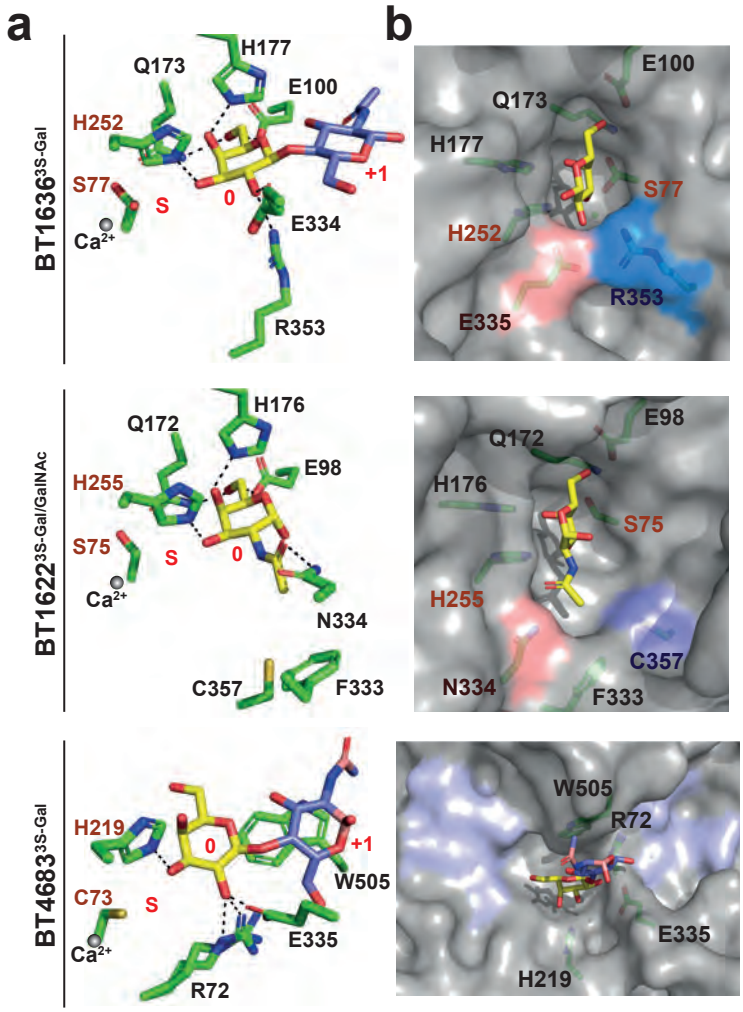


Fig 3.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

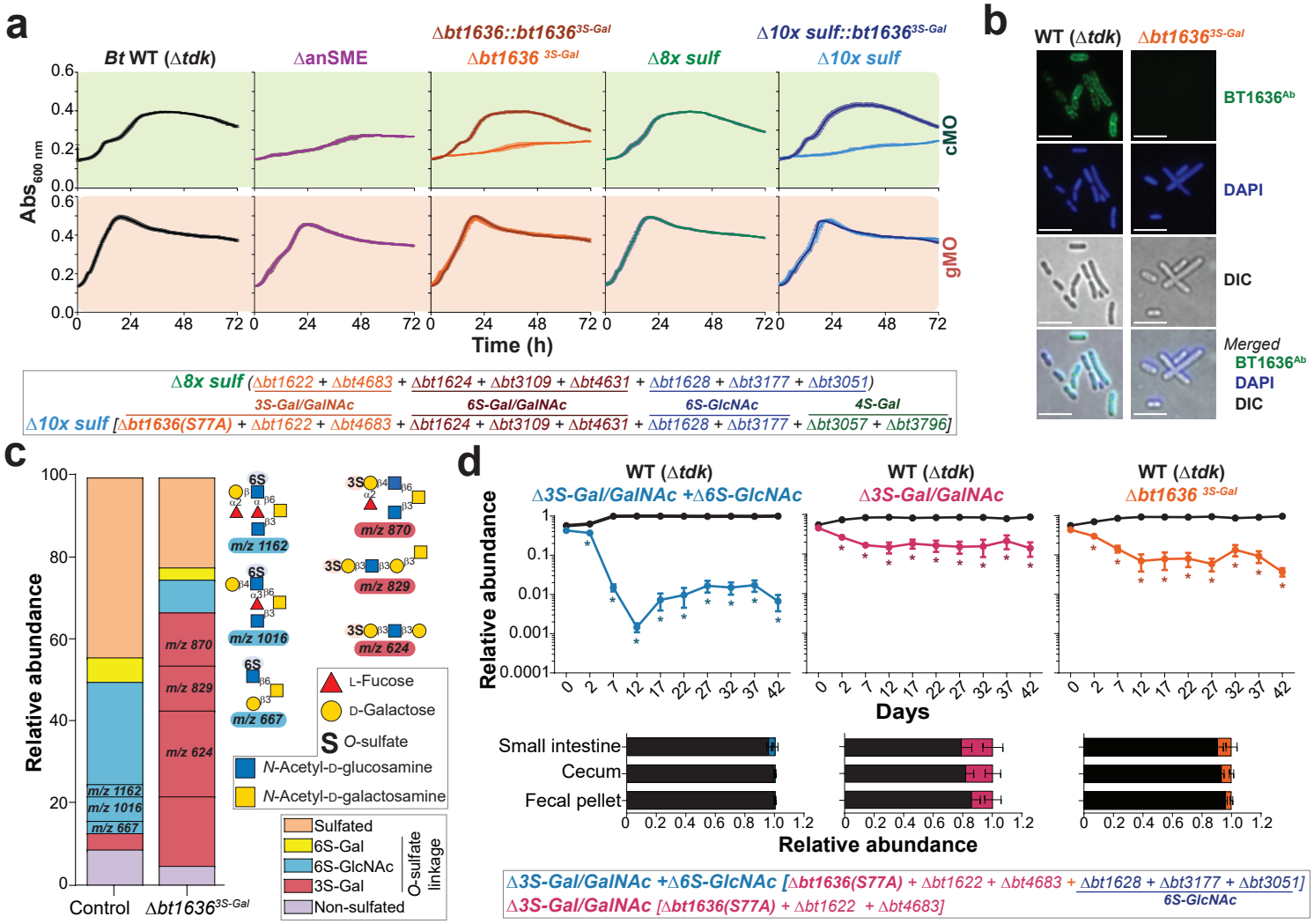
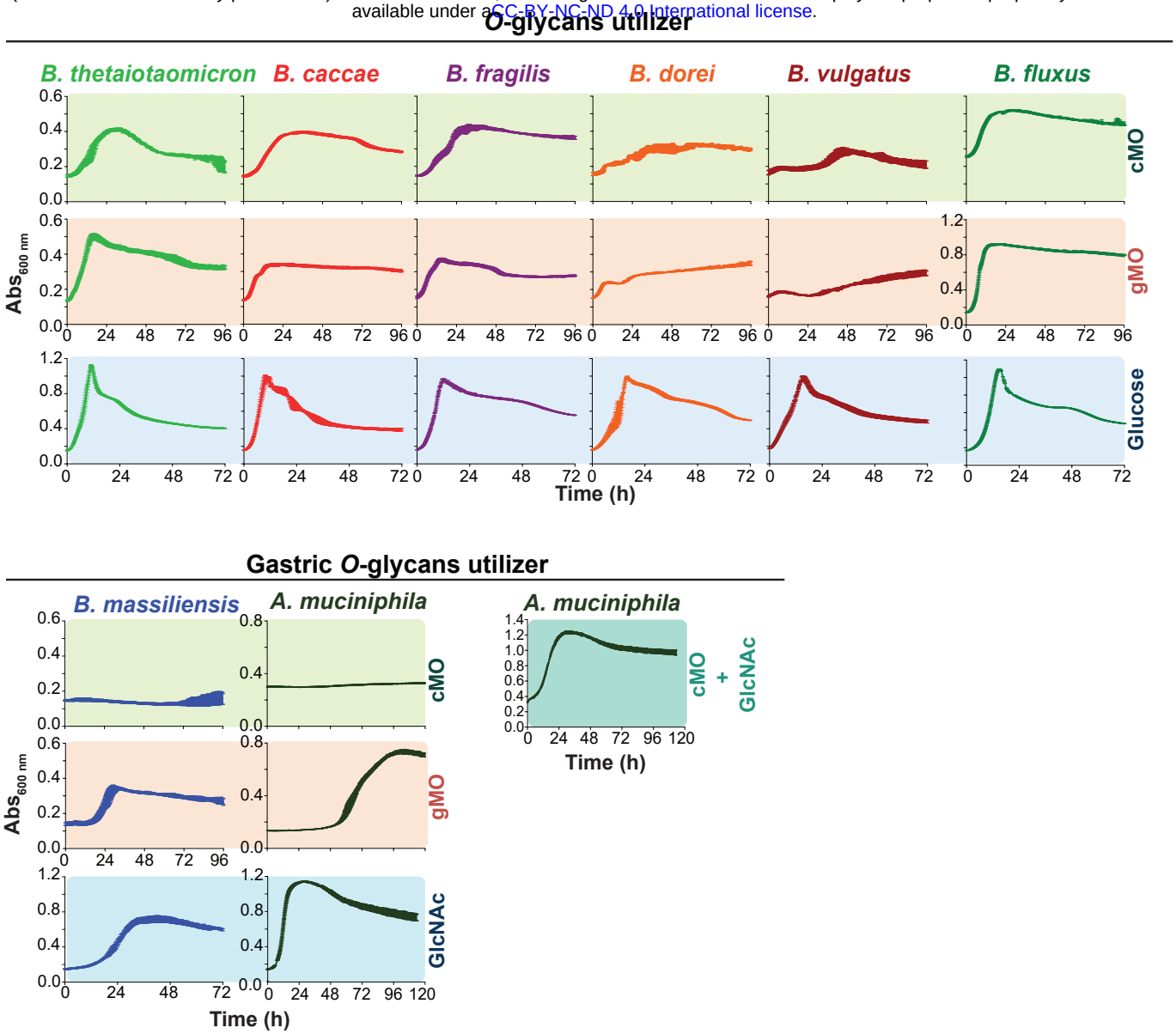


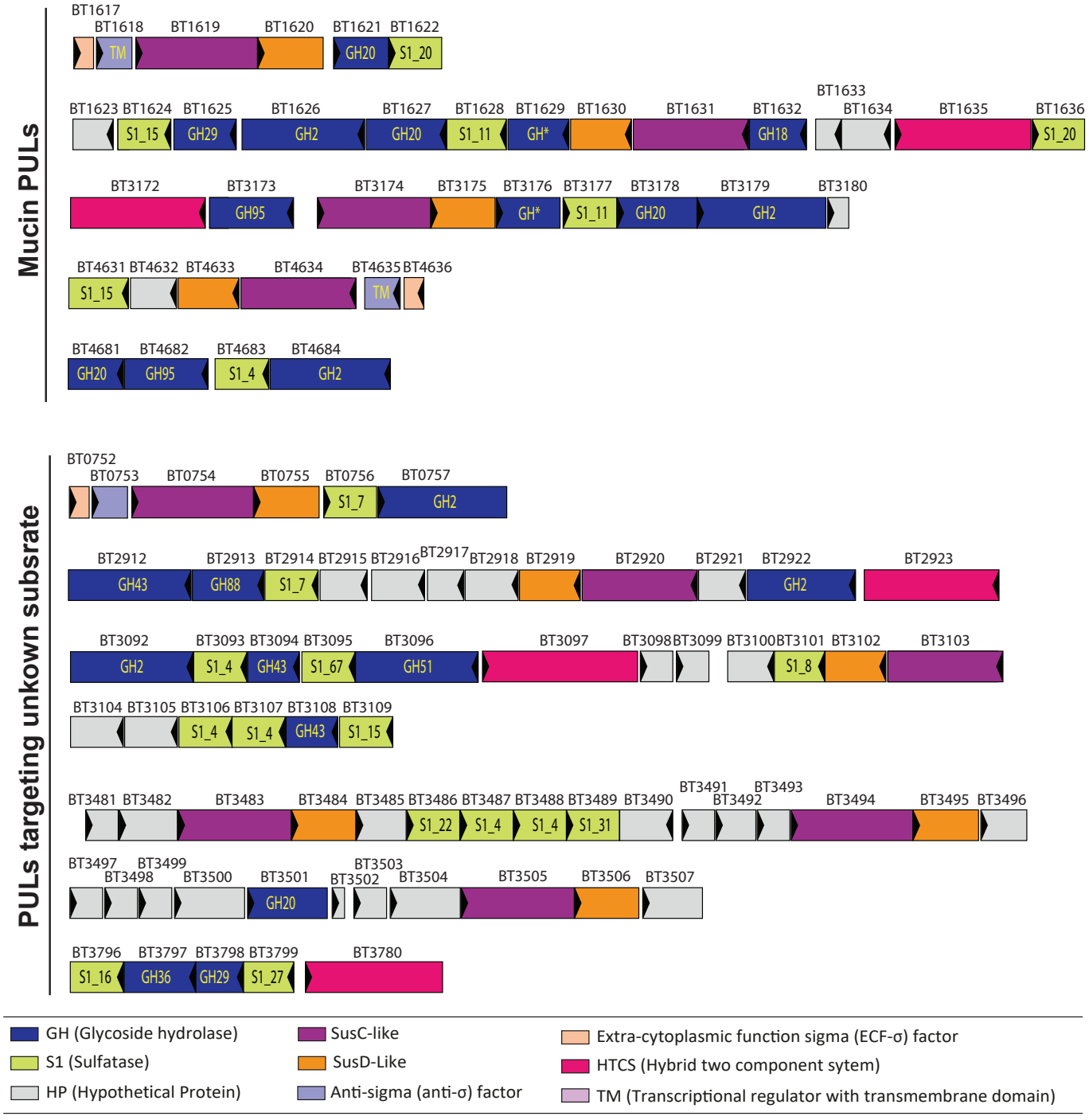
Fig 4.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



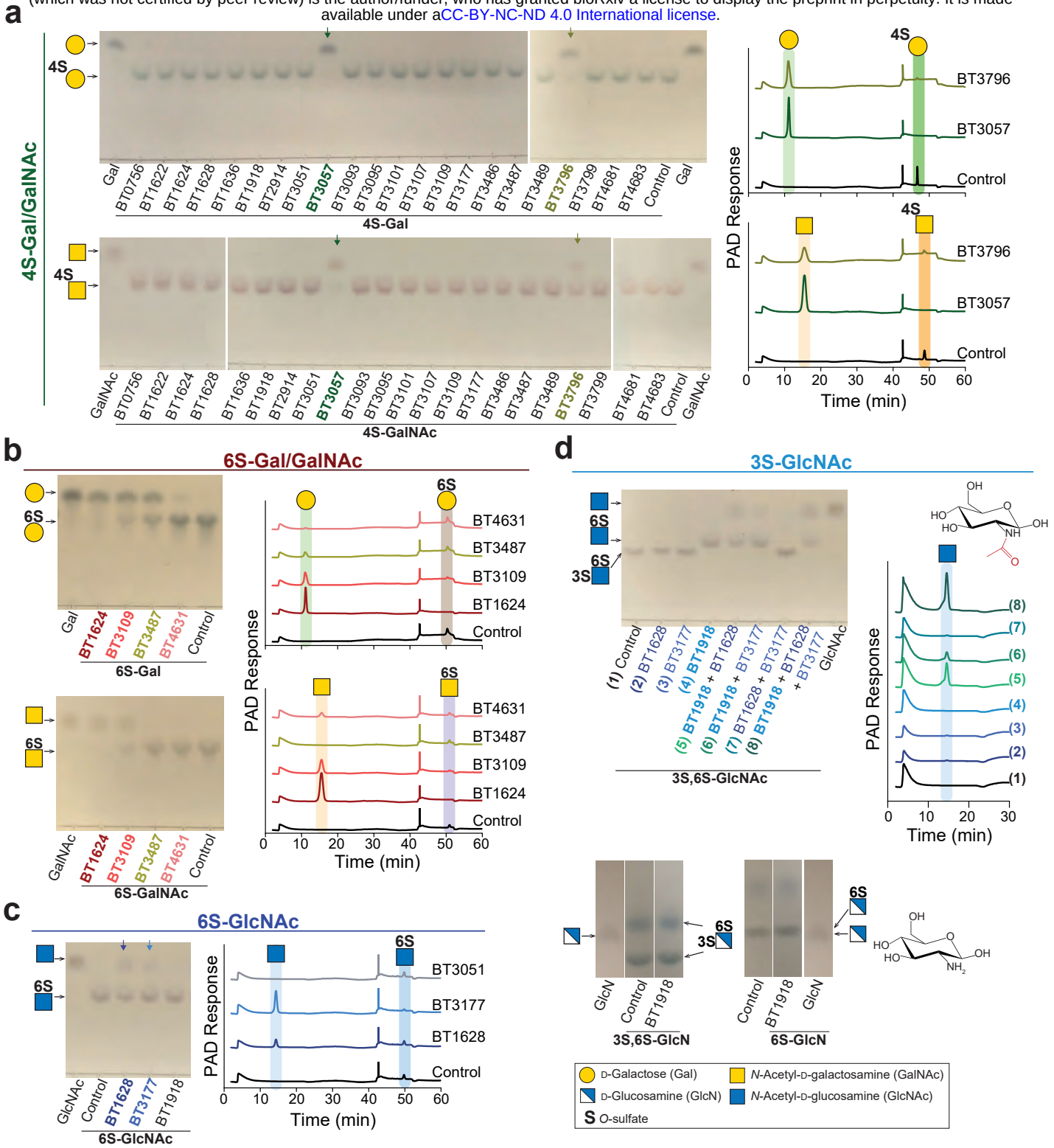
Extended Data Figure 1

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



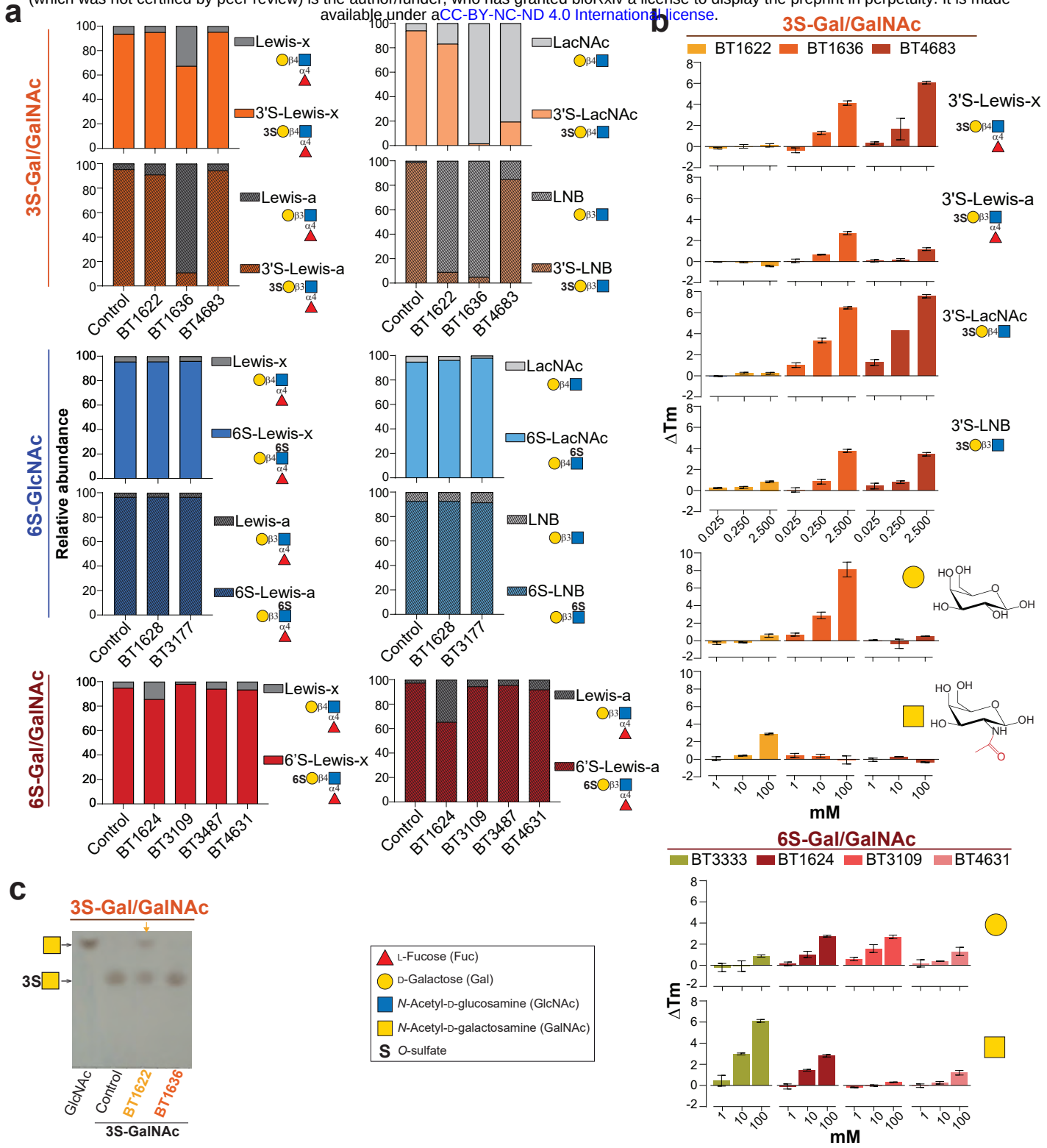
Extended Data Figure 2

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



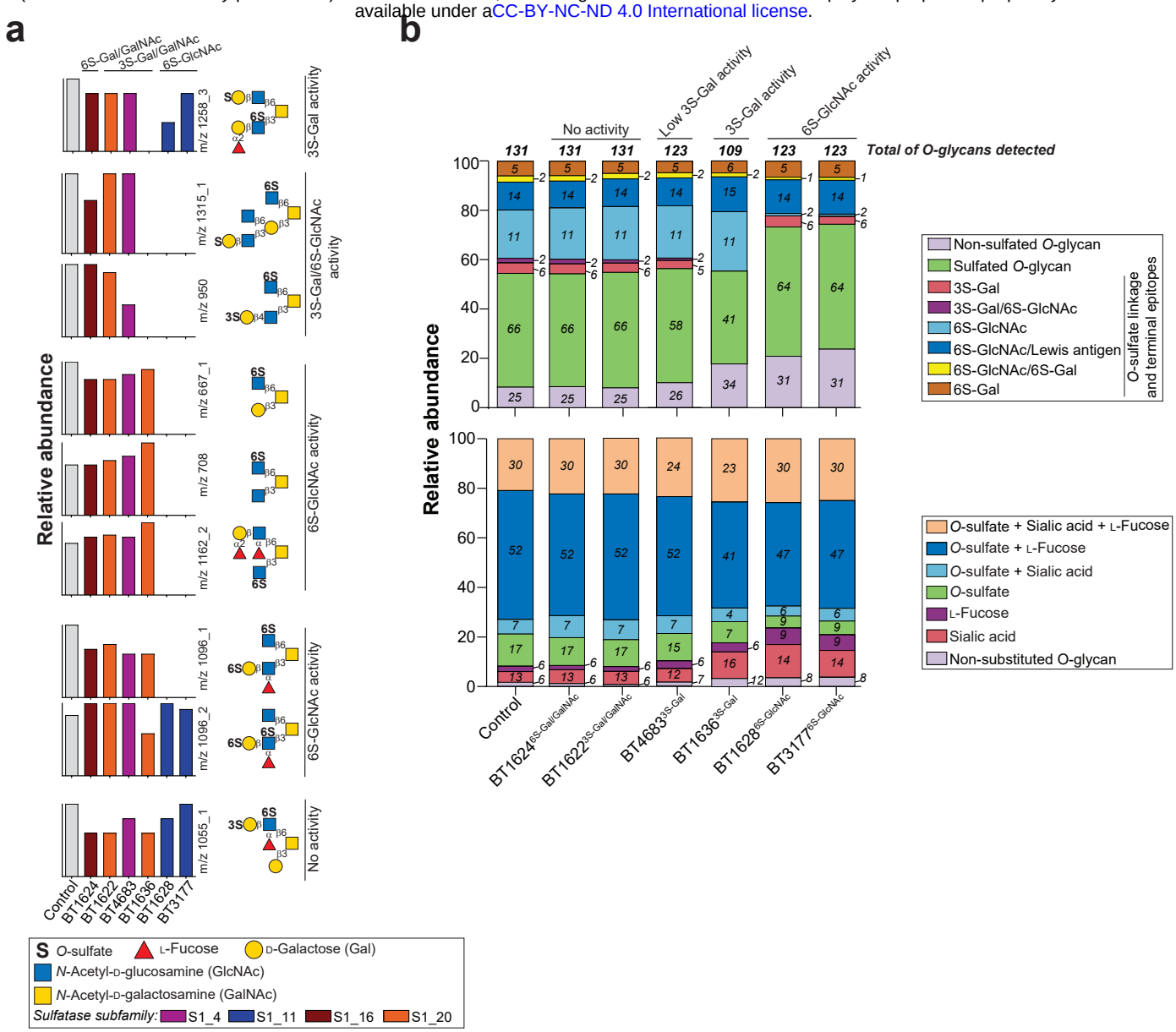
Extended Data Figure 3

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



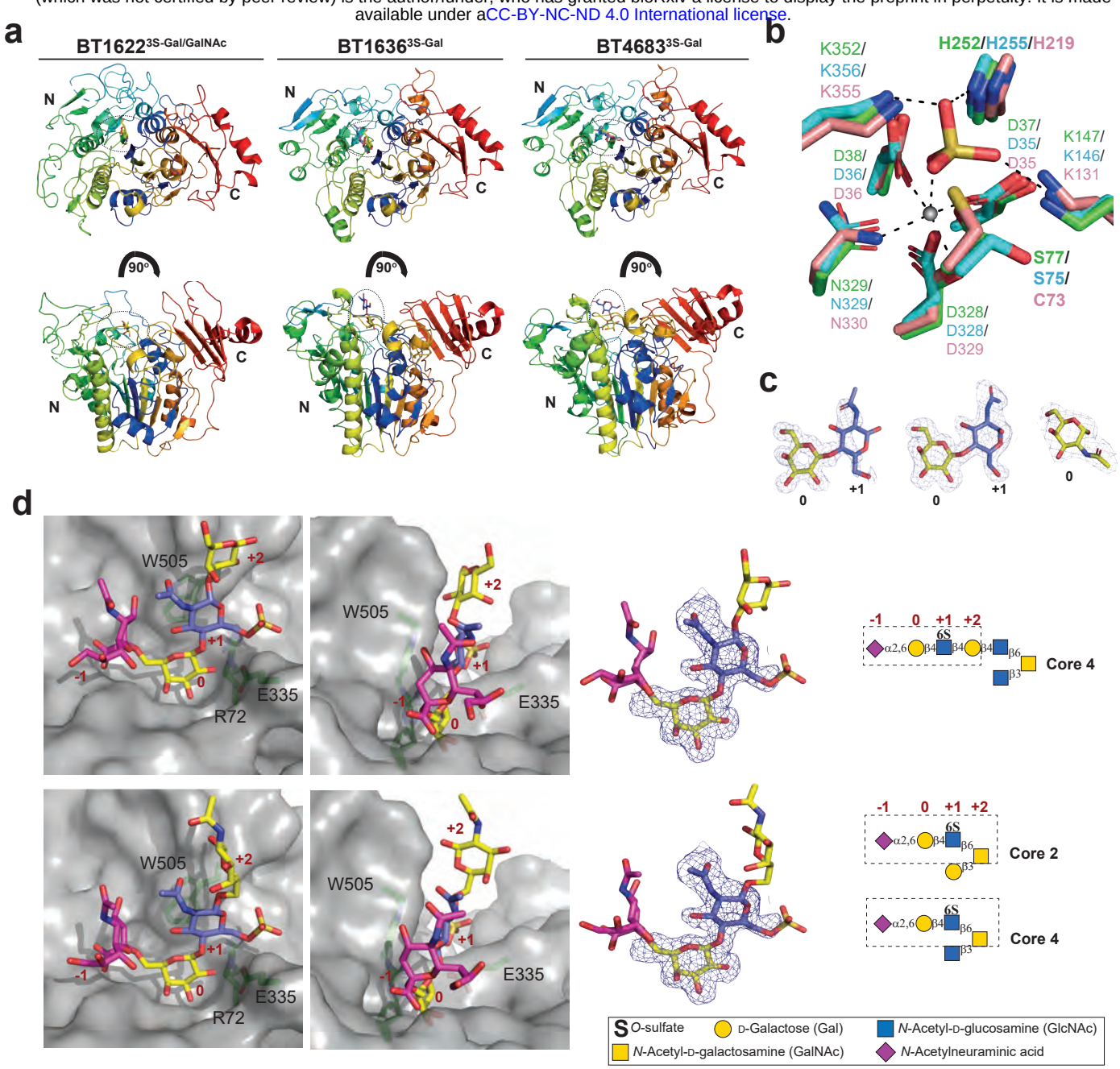
Extended Data Figure 4

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



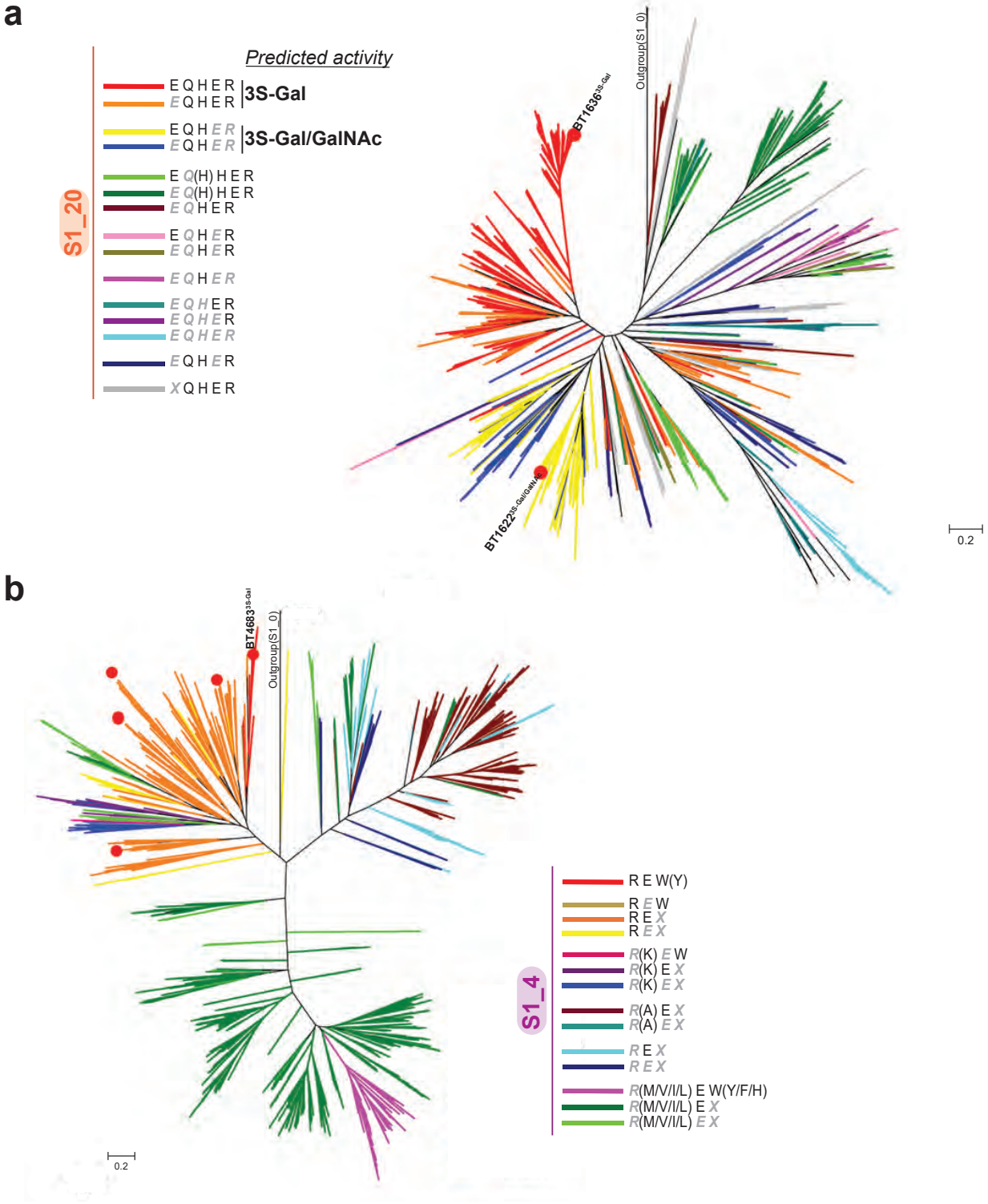
Extended Data Figure 5

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



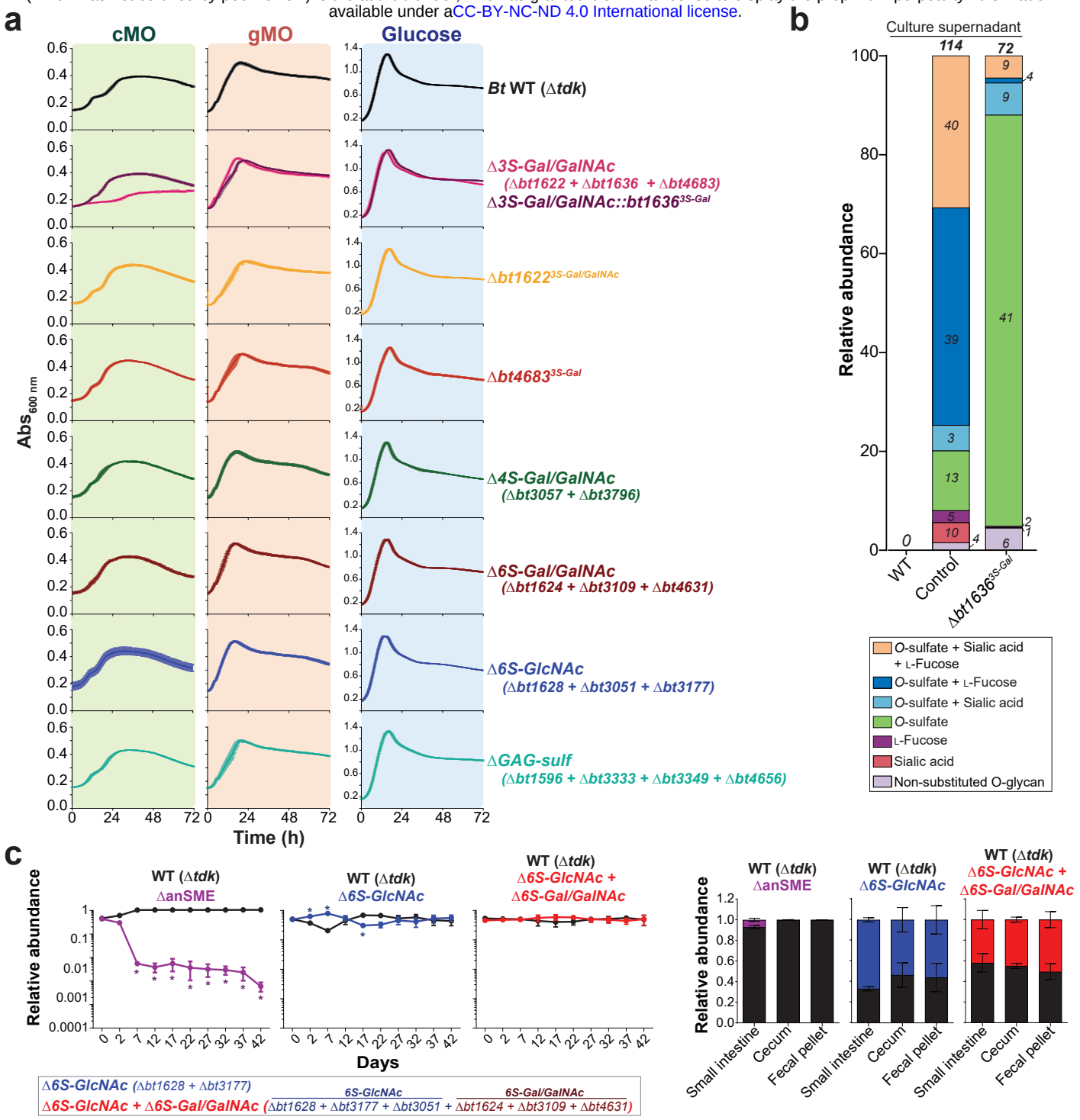
Extended Data Figure 6

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Extended Data Figure 7

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Extended Data Figure 8

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplemental Discussion

1. *Utilization of different mucin O-glycans sources by HGM*

Mucin composition varies throughout the gastrointestinal (GI) tract, with the stomach having mainly MUC5AC and the colon mainly MUC2[1]. The glycosylation of these respective mucins also varies along the GI tract with higher levels of sulfated and sialylated structures observed in the distal colon compared to the upper GI tract[2]. Among the 20 bacterial strains tested for growth, 12 failed to grow on gastric mucin *O*-glycans (gMO) or colonic mucin *O*-glycans (cMOs) (**Fig. 1b**). Only 6 bacteria were able to utilize both *O*-glycans substrates but growth was variable. In both *O*-glycan substrates, *Bacteroides thetaiotaomicron* (*Bt*), *B. caccae*, *B. fragilis* and *B. fluxus* grew better than *B. dorei* and B. *vulgatus* (**Fig. 1b** and **Extended Data Fig. 1**). The differences observed in the growth profiles were reproducible in two different batches of purified cMOs (**Fig. 1b**) Indeed, it is likely that different HGM members have evolved to target different (or only a subset) of the available *O*-glycans and this fine-tuning of host glycan utilization may have important implications in gut colonization and symbiosis. Additionally, *B. massiliensis* and *Akkermansia mucinipila* grew on gMO but failed to utilize cMO (**Fig. 1b** and **Extended Data Fig. 1**). Both strains were able to grow on *N*-acetylglucosamine (GlcNAc) and *Akkermansia mucinipila* grew on GlcNAc in the presence of cMO suggesting that these *O*-glycans do not inhibit the growth of this bacterium. Previous studies have determined that *B. massiliensis* and *Akkermansia mucinipila* are mucin-degraders by demonstrating growth on gastric mucins[3,4]. However, the lack of growth in colonic *O*-glycans suggests that these bacteria are not able to initiate the degradation of more complex, sulfated colonic glycans. This finding highlights the importance of taking into account *O*-glycosylation differences along the GI tract and the need to utilize colonic mucins to draw conclusions regarding the full mucin-degrading potential of the colonic HGM.

2. **Sulfatase activity in cMO**

Despite all 12 sulfatases being active on defined oligosaccharides, of those tested on cMO, BT1622$^{3S\text{-}Gal/GalNAc}$ (S1_20 subfamily) and BT1624$^{6S\text{-}Gal/GalNAc}$ (S1_15 subfamily) did not show any activity on this complex substrate (**Fig. 2, Extended Data Fig. 5** and **Supplementary Table 4)**. These findings are consistent with the results

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

observed in defined commercial oligosaccharides where BT1622$^{3S\text{-}Gal/GalNAc}$ showed a preference for sulfated GalNAc over Gal glycans (**Extended Data Fig. 4**) and BT1624$^{6S\text{-}Gal/GalNAc}$ activity is blocked by the presence of additional substitutions (such as Lewis antigens) (**Extended Data Fig. 4a**). Additionally, 2 of 6 detected 6S-Gal structures contained a capping sialic acid and a terminal blood group H type 2 [Fuc-α1,2-(6S)Gal-β1,4-GlcNAc-] (**Fig. 2**). The lack of activity of BT1624$^{6S\text{-}Gal/GalNAc}$ towards such structures confirms an exo-mode of action that we describe for this sulfatase using commercial substrates.

Overall, when compared to the non-enzyme treated control, we detected an increase of non-sulfated structures and decrease of sulfated oligosaccharides in all samples with the active enzymes (BT4683$^{3S\text{-}Gal}$, BT1636$^{3S\text{-}Gal}$, BT1628$^{6S\text{-}GlcNAc}$ and BT3177$^{6S\text{-}GlcNAc}$) (**Extended Data Fig. 5b**). BT1636$^{3S\text{-}Gal}$ (S1_20) was active towards all detected 3S-Gal structures with the exception of glycan 1055_1 that is a doubly sulfated 3S-Gal/6S-GlcNAc fucosylated structure (**Extended Data Fig. 5a**). As we observed using commercial substrates, the presence of Lewis-a/x epitopes leads to a decrease in the activity of this sulfatase (**Extended Data Fig. 4a** and **Supplementary Table 4**) and the presence of a second sulfate group might exacerbate this negative effect leading to the lack of activity towards this complex sulfated *O*-glycan. The incubation of the 6S-GlcNAc sulfatases BT1628$^{6S\text{-}GlcNAc}$ and BT3177$^{6S\text{-}GlcNAc}$ with cMO suggests that these enzymes are redundant, but because they are encoded in different PULs they could be expressed in response to different activating cues (**Fig. 2** and **Supplementary Table 4)**. Compared to the non-enzyme treated control, 16 glycans were not detected after incubation with these sulfatases, 14 of these structures have a terminal 6S-GlcNAc (**Fig. 2** and **Supplementary Table 4)**. BT1628$^{6S\text{-}GlcNAc}$ and BT3177$^{6S\text{-}GlcNAc}$ were active in 6S-GlcNAc core 3 (GlcNAc-β1,3-GalNAc) and core 4 (GlcNAc-β1,6-GalNAc) structures (**Extended Data Fig. 5a**), suggesting that these sulfatases are well suited to accommodate the variations in linkages/sugars found in mucin *O*-glycans. Additionally, we also detect 7 new glycans that are likely to be reaction products of BT1628$^{6S\text{-}GlcNAc}$ and BT3177$^{6S\text{-}GlcNAc}$ (**Fig. 2** and **Supplementary Table 4)**.

The identification and characterization of the first sulfatases active on mucin *O*-glycans creates the opportunity to improve our understanding of *O*-glycan structures by using these enzymes as analytical tools. After the treatment with BT1636$^{3S\text{-}Gal}$

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

several oligosaccharides predicted to contain a terminal sulfate linked to Gal were not detected. Although we could not determine the specific sulfate linkage by mass spectrometry, the activity of the 3S-sulfatase suggests that these oligos contain a terminal 3S-Gal (**Extended Data Fig. 5a** and **Supplementary Table 4**). The specificity of the 6S-GlcNAc sulfatases for non-fucosylated *O*-glycans also illuminates their potential use as tools to characterize the structure of these complex structures since it allows the differentiation of different isomers. For example, we detect two oligosaccharides with mass 1096, however, after incubation with BT1628$^{6S\text{-}GlcNAc}$ or BT3177$^{6S\text{-}GlcNAc}$, only the isomer 1096_2 was detected, indicating that the isomer 1096_1 contain a terminal 6S-GlcNAc (**Extended Data Fig. 5a** and **Supplementary Table 4**).

3. ***Conserved structural features of the S1 formylglycine family***

Protein fold and subsites nomenclature

S1 sulfatases comprise the most common and largest family of sulfatases, currently encompass 36,816 members in sulfAtlas and are found in all domains of life[5]. S1 sulfatases are part of the alkaline phosphatase superfamily and adopt an alkaline phosphatase-like fold. This is an N-terminal α/β/α domain with S1 sulfatases also possessing a smaller C-terminal 'sub domain'. The active site is located in the N-terminal domain that has a large mixed β-sheet composed of ~10 β strands, sandwiched between α helices above and below. The C-terminal 'sub-domain' is composed of a 4 stranded antiparallel β-sheet and a single amphipathic terminal helix. This C-terminal domain abuts the N-terminal domain through the antiparallel β-sheet with loops from the β strands sometimes contributing to the active site architecture (**Extended data Fig. 6a**). The subsite nomenclature for carbohydrate sulfatases is such that the invariant sulfate binding site is denoted as the S site. The S site sulfate is appended to the 0 subsite sugar. Subsites then increase in number (i.e. +1, +2, +3) as the sugar moves toward the reducing end (free O1) and decreases in number as the sugar chain moves towards the non-reducing end (i.e. -1, -2, -3)[6].

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

S1 formylglycine active site conservation

The sulfate binding site (S site) is invariant across the S1 family and comprises the catalytic residues (nucleophile and catalytic acid) and a calcium binding site (**Extended data Fig. 6b**). An invariant histidine is likely the potential catalytic acid but a lysine has also been suggested to possibly fulfil this role[7]. The pKa of His is ~6.0, whilst Lys has a pKa of >10, making it more chemically feasible that His performs the role of the catalytic acid. Homologues of these residues (H252 and K352 in BT1636$^{3S\text{-}Gal}$) make hydrogen bonds to the scissile sulfoester linkage (**Extended data Fig. 6b**). Previously published work with BT1596 and BT4656, which are 2S-Uronic acid and 6S-GlcNAc sulfatases, respectively, showed that the mutation of either residue to alanine causes inactivation[7]. Consistent with this work, a BT4683$^{3S\text{-}Gal}$ H219A mutant was inactive. However, in BT1622$^{3S\text{-}Gal/GalNAc}$, the mutation of H255 to Ala caused only a ~30-fold decrease in activity (**Supplementary Table 3**). Thus, it is possible that in BT1622$^{3S\text{-}Gal/GalNAc}$ the loss of H255 is compensated by the invariant residue K356 and interestingly BT1622$^{3S\text{-}Gal/GalNAc}$ has a pH optimum ~2 units higher than most sulfatases assayed (**Supplementary Fig. 4**).

The calcium binding site is located at the base of the S site interacting with the sulfate group. This calcium ion is an essential component of the catalytic mechanism helping to stabilise negative charges that occur during the catalysis. All three of the solved structures had occupation for calcium. In BT1636$^{3S\text{-}Gal}$ D328 and the sulfate group of the substrate coordinate above and below the calcium with D37, D38, N329 and the formylglycine binding in a plane completing an octahedral coordination (**Extended Data Fig. 6b**). These three Asp and the Asn coordinated with calcium are structurally conserved in all 3S-Gal/GalNAc sulfatases structures (**Extended data Fig. 6b**).

The solved structures of BT1636$^{3S\text{-}Gal}$ and BT1622$^{3S\text{-}Gal/GalNAc}$ were native *Bt* proteins having a Ser at the formylglycine position. However, the structure of BT4683$^{3S\text{-}Gal}$ was obtained with the active protein where S73 was mutated to Cys (as *E. coli* can only convert Cys, not Ser, to formylglycine). The analysis of BT4683$^{3S\text{-}Gal}$ reveals that the crystallized protein still has the Cys and not formyglycine indicating poor installation of the formylglycine. This observation means the kinetic data, (although the rates are significant and readily measurable) may be an underestimation of true catalytic performance. This will affect the k_{cat} component of the k_{cat}/K_M

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

measurement and thus the k_{cat}/K_M reported in **Supplementary Table 3** is an underestimate of the true activity.

Additional 3S-Gal/GalNAc specificity determinants based on structures

BT1636$^{3S\text{-}Gal}$ was solved in complex with the product LacNAc; the Gal at 0 subsite is well ordered and makes extensive interactions, whilst the +1 GlcNAc is highly disordered and appears to make no interactions with the protein (**Fig. 3** and **Extended Data Fig. 6c**). O2 of the Gal hydrogen bonds with Oε1 of E334 and NH2 of R353. Mutation of these residues to Ala causes ~300 and ~60-fold reductions in k_{cat}/K_M, respectively. The O6 group of Gal potentially coordinates with Oε2 of E100 and Nε2 of Q173 and mutations of these residues to Ala cause ~80 and ~50-fold decreases in k_{cat}/K_M, respectively (**Fig. 3** and **Supplementary Table 3**). Comparison of the BT1622$^{3S\text{-}Gal/GaNAc}$ structure with BT1636$^{3S\text{-}Gal}$ shows that E98 and Q172 (which correspond to E100 and Q173 in BT1636$^{3S\text{-}Gal}$) are conserved **(Fig. 3)** and mutating E98 to Ala caused only a 15-fold decrease in k_{cat}/K_M (**Supplementary Table 3**). Additionally, in BT1622$^{3S\text{-}Gal/GaNAc}$ the hydrophobic interactions with the *N*-acetyl group, and the more open pocket, may offset the effects H176A (300-fold loss in activity) when compared to H177A (complete loss in activity) in BT1636$^{3S\text{-}Gal}$ (**Fig. 3** and **Supplementary Table 3**).

BT4683$^{Gal\text{-}3S}$ also displayed the same 3S-Gal activity as the S1_20 enzymes but showed a preference for 3'S-LacNAc, reciprocal to BT1622$^{3S\text{-}Gal}$. BT4683$^{3S\text{-}Gal}$ bound the O2 Gal of LacNAc via Oε2 of E335 (equivalent to E334 in BT1636$^{3S\text{-}Gal}$) and through either Nε or NH1 of R72 **(Fig. 3)**. Although R72 is sequentially distal to R353 in BT1636$^{3S\text{-}Gal}$ it is spatially similar and likely contributes in a similar capacity **(Fig. 3)**. Despite the mutations R72A and E335A resulting in loss of activity, the Glu and Arg are only conserved in 62 % and 19 % of S1_4 sequences, respectively, suggesting there is a significant but not absolute selection for an equatorial O2 in this subfamily (**Extended Data Fig. 7** and **Supplementary Fig. 2**). Uniquely among the 3S-Gal sulfatases identified, BT4683$^{3S\text{-}Gal}$ utilises a hydrophobic stacking interaction through W505 to provide a platform for the +1 GlcNAc and partially the 0 Gal. Mutation of W505 to Ala almost completely abolishes activity on 3'S-LacNAc (**Supplementary Table 3**) but surprisingly this residue is not conserved in our phylogenetic analyses of S1_4 being present in only 8 other sequences (**Extended Data 7** and **Supplementary**

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig. 2). It is important to note, that W505 is not well conserved; potential equivalent aromatic residues can be found in some additional clades, which are coloured light brown (or bronze), pink or dark red, but it is not evident from the alignment that these are functional equivalents. Future structural work is needed to confirm if other aromatic residues take equivalent positions in those sulfatases. Additionally, the BT4683$^{3S\text{-}Gal}$ activity against defined sulfated saccharides was suggestive of an exo-acting enzyme that cleaves terminal 3S-Gal (**Extended Data Fig. 4**). However, a close analysis of this sulfatase structure shows that the active site is located in an open cleft characteristic of an endo-active enzyme[6]. This more open cleft of BT4683$^{3S\text{-}Gal}$ may allow additional sugars/sulfates to be accommodated on the O6 of both the 0 Gal and +1 GlcNAc. Indeed, the activity determined in cMO shows that this sulfatase can act on sialylated *O*-glycan (**Fig. 2**). Further modelling of different *O*-glycan structures (using the crystallographically solved LacNAc as an 'anchor') indicate that this enzyme can accommodate complex *O*-glycans with internal sulfation **(Fig. 3 and Extended Data 6d)**. Together, these results suggest that BT4683$^{Gal\text{-}3S}$, and its close homologues, could be endo 3S sulfatases where the 0 subsite specificity for Gal is driven by glycan context and/or distal subsites such as -1 and +2, rather than an axial O4 as in S1_20.

Additionally, it is unclear why BT1636$^{Gal\text{-}3S}$ acts better on LacNAc substrates than BT1622$^{3S\text{-}Gal/GalNAc}$. It is interesting to note, however, that both BT1636$^{3S\text{-}Gal}$ and BT4683$^{3S\text{-}Gal}$ perform well on LacNAc configured substrates and utilise an Arg and Glu to coordinate O2 whilst BT1622$^{3S\text{-}Gal}$ lacks these residues **(Fig. 3)**. These residues may lead to the enhanced activity on LacNAc (β1,4 glycan) vs. LNB (β1,3 substrate). Another thing to note is that a β1,4 vs β1,3 linkage will rotate the GlcNAc ~60° but switch the position of the Fuc residue from being on the '*N*-acetyl side' of the glycosidic bond to the 'O6 side' of the glycosidic bond, and this may also be the cause of the differential activities on β1,4 vs β1,3 linked substrates.

Phylogenetic analyses of S1_20 specificity determinants

The essential His that acts as a key specificity determinant of galacto- over gluco-substrates (H177 and H176 in BT1636$^{3S\text{-}Gal}$ and BT1622$^{3S\text{-}Gal/GalNAc}$, respectively) is highly conserved (92% of S1_20 sequences) (**Extended data Fig. 7** and **Supplementary Fig. 3**). The Gln (Q173 and Q172 in BT1636$^{3S\text{-}Gal}$ and BT1622$^{3S\text{-}}$

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$^{Gal/GalNAc}$, respectively) is only conserved in 66% of sequences and in 25% of the cases is substituted with a histidine, a residue that can also fulfill the same role of Gln interacting with Gal O6. Indeed, these conserved residues are located in a highly conserved domain with the consensus sequence [CDNS]-[QH]-[RVF]-[QHLD]-[AG]-H-[NRST]-[YHF]-[YF]-P (Prosite syntax). With H177 targeting the axial O4 of Gal directly, a Q173 may function indirectly to select for an axial O4 and thus these residues may operate as a selectivity 'dyad' for Gal with S1_20. Additionally, the residues implicated in recognition of Gal over GalNAc, E335 and R353 in BT1636$^{3S\text{-}Gal}$ are conserved in 64 and 74% of S1_20 sequences, whilst the residue that allows the accommodation of O2 N-acetyl and activity in GalNAc (N334 in BT1622$^{3S\text{-}Gal/GalNAc}$) is only found in 8% of members of this family (**Extended Data Fig. 7** and **Supplementary Fig. 3**). This observation suggests that the majority of the S1_20 sulfatases evolved to target sulfated Gal and only a subset of this subfamily's members can actually also be active on GalNAc. Interestingly, all of the close homologs of BT1636$^{3S\text{-}Gal}$ and BT1622$^{3S\text{-}Gal/GalNAc}$ that share the critical specificity determinants of these proteins (**Supplementary Tables 12 and 13**) were isolated from mammals at body regions rich in mucins, highlighting the role of these sulfatases in accessing sulfated host glycans.

4. **Growth of sulfatase mutants on *O*-glycans**

The deletion strain lacking 4S-Gal/GalNAc sulfatases (*Δbt3057* + *Δbt3796*) did not show any phenotype in cMO (**Extended Data Fig. 8a**), a result that is consistent with the lack of these sulfated linkages in colonic mucins (**Supplementary Table 4**). Unexpectedly, the deletion strains lacking the identified 6S-Gal/GalNAc sulfatases (*Δbt1624* + *Δbt3109* + *Δbt4631*) and 6S-GlcNAc sulfatases (*Δbt1628* + *Δbt3051* + *Δbt3177*) also did not show any growth defect on cMOs (**Extended Data Fig. 8a**). Analysis of cMO by mass spectrometry showed that this substrate contains a low abundance of 6S-Gal but a relatively high abundance of 6S-GlcNAc, especially in shorter structures (**Supplementary Table 4**). Although the low abundance of O6-sulfated Gal could explain the lack of phenotype of the 6S-Gal/GalNAc sulfatase deficient strain, the lack of effect in the Δ*6S-GlcNAc* mutant in cMO was unexpected (**Extended Data Fig. 8a**). Due to the limitations of the mass spectrometry technique it is not possible to analyse sulfation in longer oligos, making the real complexity of glycans found in colonic mucins unclear. Indeed, the lack of phenotype of Δ*6S-GlcNAc*

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

mutant in cMO suggests that 6S-GlcNAc might not be a major terminal epitope in colonic mucins. It is also important to note that the mutant *Δ6S-GlcNAc* is the deletion of two characterized 6S-GlcNAc sulfatases active on cMO (BT1628$^{6S\text{-}GlcNAc}$ and BT3177$^{6S\text{-}GlcNAc}$) and a third closely related S1_11 sulfatase (BT3051$^{putative_6S\text{-}GlcNAc}$) for which no activity was found. This putative 6S-GlcNAc sulfatase was deleted to avoid possible compensation of function after loss of BT1628$^{6S\text{-}GlcNAc}$ and BT3177$^{6S\text{-}GlcNAc}$ activities.

The deletion of previously characterized GAG-specific sulfatases[8] (*Δbt1596* + *Δbt3333* + *Δbt3349* + *Δbt4656*) did not result in any observable phenotype in cMO (**Extended Data Fig. 8a**), indicating that this substrate was not contaminated with additional endogenous host glycans. Additionally, despite some mutants exhibiting growth defects on sulfated cMO, all of the mutants grew well on gMO and glucose (**Fig. 4a** and **Extended Data Fig. 8a**), suggesting that the phenotypes observed are dependent on the mucin source (colon) and cannot be observed utilizing mucins from other regions of the gastrointestinal tract. Together these results highlight the contribution of sulfatases in utilization of colonic mucins by the HGM.

5. **Analysis of *Δbt1636$^{3S\text{-}Gal}$* culture supernatant by MS**

The analysis of the oligosaccharides present in *Δbt1636$^{3S\text{-}Gal}$* culture supernatant after 96h incubation revealed that the detected glycans are different from the cMO profile in the starting material (**Fig. 4c**, **Extended Data Fig. 8b** and **Supplementary Table 5**). We detected 114 glycans in the cMO sample, of which 39 were sulfated and fucosylated (44% total) (**Extended Data Fig. 8b**) and the three most common structures (12% total) were 6S-GlcNAc oligosaccharides (**Fig. 4c** and **Supplementary Table 5**). In the control sample, the levels of sulfation, sialylation and fucosylation were 92%, 40% and 77%, respectively (**Supplementary Table 5**). In the *Δbt1636$^{3S\text{-}Gal}$* culture supernatant, we detected 72 glycans, of which 41 were substituted only with *O*-sulfate (84% total) (**Extended Data Fig. 8b**). In the mutant supernatant the levels of sulfation (95%) were similar to cMO, however the levels of sialylation (11%) and fucosylation (5%) decreased substantially (**Supplementary Table 5**), suggesting that this mutant is not able to utilize sulfated structures and these accumulate in culture media.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Additionally, a total of 98 of the 114 structures present in cMO were not detected in *Δbt1636*$^{3S\text{-}Gal}$ culture supernatant whereas in mutant supernatant, we detected 49 glycans that were not detected in the initial substrate (**Supplementary Table 5**). This suggests that some of the oligosaccharides present in cMO can support the limited growth of *Δbt1636*$^{3S\text{-}Gal}$ and, although this mutant is not able to utilize many sulfated cMO structures, it can still modify the glycans to create novel structures. It remains unclear which enzymes are encoded by the mutant to modify the *O*-glycans, but the presence of a cell surface sialidase[9] can explain the decrease of sialylation levels in structures found in *Δbt1636*$^{3S\text{-}Gal}$ supernatant. Additionally, the presence of surface endo-acting glycoside hydrolases able to cleave *O*-glycans into shorter oligosaccharides[10] can also contribute to new glycan structures in the mutant culture supernatant. Together these results show that *Δbt1636*$^{3S\text{-}Gal}$ is not able to utilize most sulfated *O*-glycans explaining the limited growth of this mutant in cMO.

References

1 Arike, L. & Hansson, G. C. The Densely O-Glycosylated MUC2 Mucin Protects the Intestine and Provides Food for the Commensal Bacteria. *J Mol Biol* **428**, 3221-3229, doi:10.1016/j.jmb.2016.02.010 (2016).

2 Arike, L., Holmen-Larsson, J. & Hansson, G. C. Intestinal Muc2 mucin O-glycosylation is affected by microbiota and regulated by differential expression of glycosyltranferases. *Glycobiology* **27**, 318-328, doi:10.1093/glycob/cww134 (2017).

3 Derrien, M., Vaughan, E. E., Plugge, C. M. & de Vos, W. M. *Akkermansia muciniphila* gen. nov., sp. nov., a human intestinal mucin-degrading bacterium. *Int J Syst Evol Microbiol* **p54**, 1469-1476, doi:10.1099/ijs.0.02873-0 (2004).

4 Pudlo, N. A. *et al.* Symbiotic Human Gut Bacteria with Variable Metabolic Priorities for Host Mucosal Glycans. *MBio* **6**, e01282-01215, doi:10.1128/mBio.01282-15 (2015).

5 Barbeyron, T. *et al.* Matching the Diversity of Sulfated Biomolecules: Creation of a Classification Database for Sulfatases Reflecting Their Substrate Specificity. *PLoS One* **11**, e0164846, doi:10.1371/journal.pone.0164846 (2016).

6 Hettle, A. G. *et al.* The Molecular Basis of Polysaccharide Sulfatase Activity and a Nomenclature for Catalytic Subsites in this Class of Enzyme. *Structure* **26**, 747-758 e744, doi:10.1016/j.str.2018.03.012 (2018).

7 Cartmell, A. *et al.* How members of the human gut microbiota overcome the sulfation problem posed by glycosaminoglycans. *Proc Natl Acad Sci U S A* **114**, 7037-7042, doi:10.1073/pnas.1704367114 (2017).

8 Ndeh, D. *et al.* Metabolism of multiple glycosaminoglycans by Bacteroides thetaiotaomicron is orchestrated by a versatile core genetic locus. *Nat Commun* **11**, 646, doi:10.1038/s41467-020-14509-4 (2020).

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

9 Briliute, J. *et al.* Complex N-glycan breakdown by gut Bacteroides involves an_SI. *Nat Microbiol* **4**, 1571-1581, doi:10.1038/s41564-019-0466-x (2019).

10 Crouch, L. I. *et al.* Prominent members of the human gut microbiota express endo-acting O-glycanases to initiate mucin breakdown. *Nat Commun* **11**, 4017, doi:10.1038/s41467-020-17847-5 (2020).

Supplementary Table 1

Family S1 sulfatase subfamiles encoded in the genomes of different *Bacteroides* type strains and *Akkermansia*

	SulfAtlas S1 subfamily	S1_4	S1_7	S1_8	S1_9	S1_11	S1_14	S1_15	S1_16	S1_19	S1_20	S1_22	S1_24	S1_27	S1_28	S1_31	S1_34	S1_43	S1_46	S1_50	S1_55	S1_62	S1_67	
	BT Locus	BT3093 BT3106 BT3107 **BT3487** BT3488 **BT4683**	BT0756 BT2914	BT3101	BT1596*	**BT1628** BT3051 **BT3177** BT4656*		**BT1624** **BT3109** BT3333* **BT4631**	**BT3057** **BT3796**		**BT1622** **BT1636**	BT3486		BT3349* BT3799		BT3489			**BT1918**				BT3095	
	Predicted function and activities identified on this study	Aryl-sulfatase/ **3S-Gal/** **6S-Gal**	Iduronate 2-sulfatase Endo-4S-kappa-carrageenan sulfatase	*N*-S-glucosamine sulfohydrolase	Heparin/ heparan sulfate 2-*O*-S	**6S-GlcNAc/** Di-6S-GlcN	UNK	**6S-Gal/** **6S-GalNAc**	**4S-Gal/** **4S-GalNAc**	Endo-4S-kappa-carrageenan	**3S-Gal**	UNK	UNK	Chondroitin sulfate/ dermatan sulfate 4-*O*-endosulfatase	UNK	UNK	UNK	UNK	**3S-GlcNAc**	UNK	UNK	UNK	UNK	
Bacteroides type strain	***B. thetaiotaomicron* VPI-5482 (*Bt*)**	**6**	**2**	**1**	**1**	**4**		**4**	**2**		**2**	**1**		**2**		**1**			**1**				**1**	**28**
	B. caccae ATCC 43185 (*Bc*)		1		1	3		5	3	1	2			1					1					18
	B. cellulosilyticus DSM 14838 (*Bcel*)			2	2	5	1	1	3	1	2			3				1					1	22
	B. clarus DSM 22519 (*Bcla*)		1	1	4	2	2	1	1		2			1										15
	B. dorei DSM 17855 (*Bd*)		1			3		2	2		2			1					2					13
	B. eggerthii DSM 20697 (*Be*)			1	3	2	1	1						1			1							10
	B. finegoldii DSM 17565 (*Bfin*)				1	2		4	1		1		1	1						1		1		13
	B. fluxus DSM 22534 (*Bflu*)			1	1	1		3	1		2			1									1	11
	B. fragilis ATCC 25285 (*Bf*)	1				1		3	4		4			1					1					15
	B. intestinalis DSM 17393 (*Bi*)	1		3	3	3	1	1	1		2			1									1	17
	B. massilliensis DSM 17679 (*Bm*)					2		4	3		3			1	1				1					15
	B. nordii CL02T12C05 (*Bn*)	1		2	1	2		3	3		2			1										15
	B. oleiciplenus DSM 22535 (*Bole*)			1	3	2	1	2	3		2			1						1			1	17
	B. ovatus ATCC 8483 (*Bo*)	4	1	2	1	2		4	1		2	2		2		1							1	23
	B. plebeius DSM 17135 (*Bp*)	2	1	3		2	1	1	3		3			2										18
	B. salyersiae DSM 18765 (*Bsal*)	1		1		2		2	3		3	1		2					1	1	1			18
	B. stercoris ATCC 43183 (*Bste*)			1	2	3		1	2		1			1										11
	B. uniformis ATCC 8492 (*Bu*)	2		3	1	1	1	1	1	1		1			1								1	14
	B. vulgatus ATCC 8482 (*Bv*)	1	1			2		2	2		2			1					1					12
	B. xylanisolvens XB1A (*Bx*)	1		1	1	2		2			1			1		1							1	11
	Total number of sulfatases	20	8	23	25	46	8	47	39	3	38	5	1	25	2	3	1	1	8	3	1	1	8	**316**
Other bacterium (Verrucomicrobia)																								
	Akkermansia muciniphila ATCC BAA-835 (*Akk*)	1				2		1	2	1	3													10

* Previously characterized sulfatases; (S) sulfate; (Gal) D-Galactose; (GlcNac) *N*-acetyl-D-glucosamine; (GlcN) D-glucosamine; (GalNac) *N*-acetyl-D-galactosamine; (UNK) function unknown. *Bacteroides dorei, Bacteroides massilliensis* and *Bacteroides vulgatus* were recently assigned to Phocaeicola genus.

Supplementary Table 2

List of sulfated saccharides used in the initial sulfatase activity screen

Sulfated sacharide	Structure	Abreviation	Activity
3'-*O*-Sulfated Lewis-a	3SGal-β1,3-(Fuc-α1,4)-GlcNAc	3'-Lewis^{a}	+
3'-*O*-Sulfated Lewis-x	3SGal-β1,4-(Fuc-α1,3)-GlcNAc	3'-Lewis^{x}	+
3'-*O*-Lacto-*N*-biose*	3SGal-β1,3-GlcNAc	3'-LNB	+
3'-*O*-*N*-acetyllactosamine*	3SGal-β1,4-GlcNAc	3'-LacNAc	+
6-*O*-Sulfated Lewis-a	Gal-β1,3-(Fuc-α1,4)-6SGlcNAc	6-Lewis^{a}	+
6-*O*-Sulfated Lewis-x	Gal-β1,4-(Fuc-α1,3)-6SGlcNAc	6-Lewis^{x}	+
6-*O*-Lacto-*N*-biose*	Gal-β1,3-6S-GlcNAc	6-LNB	+
6-*O*-*N*-acetyllactosamine*	Gal-β1,4-6SGlcNAc	6-LacNAc	+
6'-*O*-Sulfated Lewis-a	6SGal-β1,3-(Fuc-α1,4)-GlcNAc	6'-Lewis^{a}	+
6'-*O*-Sulfated Lewis-x	3SGal-β1,4-(Fuc-α1,3)-GlcNAc	6'-Lewis^{x}	+
D-Galactose-4-*O*-sulfate	4SGal	4S-Gal	+
D-Galactose-6-*O*-sulfate	6SGal	6S-Gal	+
N-Acetyl-D-galactosamine-4-*O*-sulfate	4SGalNAc	4S-GalNAc	+
N-Acetyl-D-galactosamine-6-*O*-sulfate	6SGalNAc	6S-GalNAc	+
N-Acetyl-D-glucosamine-3-6-di-*O*-sulfate	3S/6SGlcNAc	3S/6S-GlcNAc	+
N-Acetyl-D-glucosamine-6-*O*-sulfate	6SGlcNAc	6S-GlcNAc	+
D-Glucosamine-2-*N*, 3-*O*, 6-*O*-trisulfate			-
D-Glucosamine-2-*N*, 3-*O*-disulfate			-
D-Glucosamine-2-*N*, 6-*O*-disulfate			-
D-Glucosamine-2-*N*-sulfate			-
D-Glucosamine-3, 6-di-*O*-sulfate			-
D-Glucosamine-3,4,6-tri-*O*-sulfate			-
D-Glucosamine-3-*O*-sulfate			-
D-Glucosamine-6-*O*-sulfate			-
D-Galactosamine-2-*N*-sulfate			-

(+) At least one sulfatase was found active on this saccharide; (-) No sulfatase activity detected by thin layer chromatography; * Sulfated disaccharides generated by incubating the respective Lewis antigen with a α-1,3/1,4-fucosidase (BT1625)

Supplementary Table 3

Sulfatase kinetics for WT and mutants against different saccharides

		k_{cat}/K_M (min^{-1} M^{-1})					
		3'S-Lewis-a	**3'S-Lewis-x**	**3'S-LacNAc**	**3'S-LNB**	**3S-GalNAc**	**3S,6S-GlcNAc**
WT	BT1622$^{3S\text{-}Gal/GalNAc}$ (S75C)	$(1.8 \pm 0.04) \times 10^2$	NQ	$(1.9 \pm 0.04) \times 10^3$	$(3.6 \pm 0.13) \times 10^4$	$(5.1 \pm 0.07) \times 10^5$	NA
Mutant	BT1622$^{3S\text{-}Gal/GalNAc}$ E98A	-	-	-	$(2.4 \pm 0.42) \times 10^3$	-	-
Mutant	BT1622$^{3S\text{-}Gal/GalNAc}$ H176A	-	-	-	$(1.2 \pm 0.80) \times 10^2$	-	-
Mutant	BT1622$^{3S\text{-}Gal/GalNAc}$ H255A	-	-	-	$(1.3 \pm 0.03) \times 10^3$	-	-
Mutant	BT1622$^{3S\text{-}Gal/GalNAc}$ N334A	-	-	-	$(1.7 \pm 0.08) \times 10^5$	-	-
WT	BT1636$^{3S\text{-}Gal}$ (S77A)	$(6.4 \pm 0.40) \times 10^1$	$(4.7 \pm 0.70) \times 10^1$	$(1.0 \pm 0.15) \times 10^4$	$(5.5 \pm 0.14) \times 10^3$	-	NA
Mutant	BT1636$^{3S\text{-}Gal}$ E100A	-	-	$(1.2 \pm 0.3) \times 10^2$	-	-	-
Mutant	BT1636$^{3S\text{-}Gal}$ Q173A	-	-	$(2.1 \pm 0.07) \times 10^2$	-	-	-
Mutant	BT1636$^{3S\text{-}Gal}$ H177A	-	-	NQ	-	-	-
Mutant	BT1636$^{3S\text{-}Gal}$ E334A	-	-	$(3.1 \pm 0.3) \times 10^1$	-	-	-
Mutant	BT1636$^{3S\text{-}Gal}$ R353A	-	-	$(1.7 \pm 0.2) \times 10^2$	-	-	-
WT	BT4683$^{3S\text{-}Gal}$ (S73C)	NA	$(2.0 \pm 0.1) \times 10^1$	$(1.4 \pm 0.09) \times 10^3$	$(1.4 \pm 0.08) \times 10^2$	-	NA
Mutant	BT4683$^{3S\text{-}Gal}$ R72A	-	-	NQ	-	-	-
Mutant	BT4683$^{3S\text{-}Gal}$ H219A	-	-	NA	-	-	-
Mutant	BT4683$^{3S\text{-}Gal}$ E334A	-	-	NA	-	-	-
Mutant	BT4683$^{3S\text{-}Gal}$ W505A	-	-	6.7 ± 0.7	-	-	-
WT	BT1918$^{3S\text{-}Gal}$ (S73C)	NA	NA	-	-	-	$(6.1 \pm 0.4) \times 10^4$

		k_{cat}/K_M (min^{-1} M^{-1})				
		6S-Gal*	**6S-GalNAc***	**6S-GlcNAc**	**4S-Gal**	**4S-GalNAc**
WT	BT1624$^{6S\text{-}Gal/GalNAc}$ (S79C)	3.9 ± 0.55	1.95 ± 0.6	NA	NA	NA
WT	BT3109$^{6S\text{-}Gal/GalNAc}$ (S79C)	2.73 ± 0.74	NQ	NA	NA	NA
WT	BT3333$^{6S\text{-}Gal/GalNAc}$ (S79C)	NQ	3.1 ± 0.02	NA	NA	NA
WT	BT4631$^{6S\text{-}Gal/GalNAc}$ (S79C)	NQ	NQ	NA	NA	NA
WT	BT3177$^{6S\text{-}GlcNAc}$ (S79C)	NA	NA	(2.6 ± 0.2) x 10^4	NA	NA
WT	BT3057$^{4S\text{-}Gal/GalNAc}$ (S88C)	NA	NA	NA	NQ	(5.2 ± 0.1) x 10^2
WT	BT3796$^{4S\text{-}Gal/GalNAc}$ (S88C)	NA	NA	NA	NQ	(4.3 ± 0.09) x 10^3

Reactions performed at 28°C and in 100 mM Bis-Tris-Propane, at the appropriate pH optimum (**Supplementary Fig. 4**), supplemented with 150 mM NaCl, 5 mM $CaCl_2$ and 0.02 % Brijj. 6S-Gal/GalNAc activity was analysed in 50 mM sodium phosphate with 150 mM NaCl at 25°C by NMR. (WT) wild-type sulfatases with catalytic Ser mutated to Cys. (NQ) Active but not possible to quantify, (NA) Not active; (LacNAc) *N*-acetyl-D-lactosamine; (LNB) lacto-*N*-biose; (GlcNAc) *N*-acetyl-D-glucosamine; (Gal) D-galactose; (GalNAc) *N*-acetyl-D-galactosamine; (GlcNAc) *N*-acetyl-D-glucosamine. All values were performed in triplicate and represent averaged data with the standard error of the mean, unless otherwise stated. * These values are duplicates of specific activities at 500 μM substrate and are in units of h^{-1}.

Supplementary Table 4

LC-MS analysis of colonic mucin oligosaccharides (cMO)

Provided as a separate multi-tab Excel file

Supplementary Table 5

LC-MS analysis of O-glycans in culture supernatant of Δ*bt1636*$^{3S\text{-}Gal}$ mutant by LC-MS/MS

Provided as a separate multi-tab Excel file

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Table 6

Conservation of S1_20 3S-Gal/GalNAc specificity residues in *Bacteroides* type strains and *Akkermansia muciniphila*

	Growth in O-glycans	Locus Tag	SulfAtlas	E98 / E100	Q172 / Q173	H176 / H177	N334 / E334	C357 / R353	BT1622[3S-Gal] / BT1636[3S-Gal/GalNAc]		
					Gal/GalNAc vs Glc		Gal vs GalNAc		Specificity residues		
		Amuc_0451	B2UNH7	F	H	H	E	R			
		Amuc_0953	B2UQQ0	F	H	H	E	R			
		BACINT_04793	B3CI95	C	H	P	Q	G			
		BACCELL_01168	E2NA65	C	Y	P	Q	G			
		HMPREF9445_01772	F3PID0	C	Y	P	Q	G			
		BACSTE_00358	B0NLM9	C	Y	P	Q	G			
		Amuc_0491	B2UNV9	R	Q	H	D	G			
		BACPLE_01001	B5CWB1	S	Q	H	C	F			
		HMPREF9447_05285	K9DX80	A	E	H	E	R			
		HMPREF1071_03416	I9STT0	S	A	N	E	R			
		BACPLE_01018	B5CWC8	P	M	H	E	R			
BT1636 homologue		BF4472	Q64MS8	E	Q	H	E	R	39-88% identity	72-88% identity	79-88% identity
BT1636 homologue		HMPREF1534_03795	U6R7U0	E	Q	H	E	R	39-88% identity	72-88% identity	79-88% identity
BT1636 homologue		BACDOR_01990	B6VXI1	E	Q	H	E	R	39-88% identity	72-88% identity	79-88% identity
BT1636 homologue		BVU_1160	A6KZI7	E	Q	H	E	R	39-88% identity	72-88% identity	79-88% identity
BT1636 homologue		**BT1636**	Q8A789	E	Q	H	E	R	39-88% identity	72-88% identity	79-88% identity
BT1636 homologue		BACFIN_06228	C9KV90	E	Q	H	E	R	39-88% identity	72-88% identity	79-88% identity
BT1636 homologue		BACCAC_03738	A5ZLD8	E	Q	H	E	R	39-88% identity	72-88% identity	79-88% identity
BT1636 homologue		BACOVA_00697	A7LSB2	E	Q	H	E	R	39-88% identity	72-88% identity	79-88% identity
BT1636 homologue		BXY_12830*	D6CW71	E	Q	H	E	R	39-88% identity	72-88% identity	79-88% identity
BT1636 homologue		BACCELL_01473*	E2NB17	E	Q	H			39-88% identity	72-88% identity	
BT1636 homologue		BACINT_00870	B3C7H9	E	Q	H	E	R	39-88% identity	72-88% identity	
BT1636 homologue		HMPREF9447_02064	K9E4G4	E	Q	H	E	R	39-88% identity	72-88% identity	
BT1636 homologue		BACPLE_03515	B5D3C1	E	Q	H	E	R	39-88% identity	72-88% identity	
BT1636 homologue		HMPREF9446_03566	F3PXR8	E	Q	H	E	R	39-88% identity	72-88% identity	
BT1636 homologue		HMPREF1068_03212	I8XBU5	E	Q	H	E	R	39-88% identity		
BT1636 homologue		HMPREF1071_03543	I8Y9I8	E	Q	H	E	R	39-88% identity		
BT1636 homologue		BF3248	Q64R89	E	Q	H	E	R	39-88% identity		
BT1636 homologue		BF0570	Q64YV7	E	Q	H	E	R	39-88% identity		
BT1622 homologue		BACOVA_03124	A7LZ49	E	Q	H	N	G	54-76% identity	70-76% identity	
BT1622 homologue		BACCAC_03468	A5ZKM2	E	Q	H	N	A	54-76% identity	70-76% identity	
BT1622 homologue		HMPREF1534_02027	U6RDB9	E	Q	H	N	C	54-76% identity	70-76% identity	
BT1622 homologue		BF3255	Q64R82	E	Q	H	N	C	54-76% identity	70-76% identity	
BT1622 homologue		HMPREF1534_01949	U6RFP1	E	Q	H	N	C	54-76% identity	70-76% identity	
BT1622 homologue		**BT1622**	Q8A7A3	E	Q	H	N	C	54-76% identity	70-76% identity	
		HMPREF9446_00139	F3PN53	A	Q	H	V	R			
		HMPREF1071_03422	I8YBF9	H	Q	H	E	R			
		HMPREF1068_00776	I9SCK1	G	Q	H	E	R			
		HMPREF9445_02406	F3PJS7	K	Q	H	E	R			
		BACDOR_00391	B6VSZ0	K	Q	H	E	R			
		BVU_0164	A6KWS8	K	Q	H	E	R			

Colonic and gastric *O*-glycans utilizer | Gastric *O*-glycans utilizer | Non-mucin utilizer

* Protein without critical calcium binding residues (predicted to be inactive); Proteins locus tag: BTX (*Bacteroides thetaiotaomicron*), BACINT_X (*B. intestinalis*), BACCELL_x (*B. cellulosilyticus*), HMPREF9445_X (*B. clarus*), BACSTE_X (*B. stercoris*), BACPLE_X (*B. plebeius*), HMPREF9447_X (*B. oleiciplenus*), HMPREF1071_X (*B. salyersiae*), BFX (*B. fragilis*), HMPREF1534_X (*B. massilliensis*), BACDOR_X (*B. dorei*), BVU_X (*B. vulgatus*), BACFIN_X (*B. finegoldii*), BACCAC_X (*B. caccae*), BACOVA_X (*B. ovatus*), BXY_X (*B. xylanisolvens*), HMPREF1068_X (*B. nordii*), HMPREF9446_X (*B. fluxus*), Amuc_X (*Akkermansia muciniphila),* (Gal) D-Galactose, (GalNAc) *N*-acteyl-D-galactosamine

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Table 7

Primers designed to clone *Bt* sulfatases

Gene	Vector	Primer name	Primer sequence (3' to 5')
BT0756	pETite_Ntag	BT0756_Fw	catcatcaccaccatcacgagaacctgtacttccagggc cagcacaatgtctccgatgtttcc
		BT0756_Rv	gtggcggccgctctatta ttatttttttaatgatgatttcttcac
BT1622	pETite_Ntag	BT1622_Fw	catcatcaccaccatcacgagaacctgtacttccagggc caaacccaacccgcatctccc
		BT1622_Rv	gtggcggccgctctatta aaagtccggaccgtttctagc
	pET28b _Ntag	BT1622_Fw_NheI	ggaata**gctagc** caaacccaacccgcatctcccaac
		BT1622_Rv_XhoI	gagata**ctcgag** ttaaaagtccggaccgtttctagc
BT1624	pETite_Ctag	BT1624_Fw	gaaggagatatacatatg caggagaaacaaacccagcgtcc
		BT1624_Rv	gtgatggtggtgatgatg tagtggcagaccgtagcgtc
	pET28b _Ntag	BT1624_Fw_NheI	ggaata**gctagc** aaacaaacccagcgtccgaatgtt
		BT1624_Rv_XhoI	gagata**ctcgag** ttatagtggcagaccgtagcgtcc
BT1628	pETite_Ctag	BT1628_Fw	gaaggagatatacatatg acaaacaagcagaaaacgacagaacaaaagccg
		BT1628_Rv	gtgatggtggtgatgatg ttccttgtccctttccggaga
BT1636	pETite_Ctag	BT1636_Fw	gaaggagatatacatatg cagaaaaacaacacaaagccgaatatccctc
		BT1636_Rv	gtgatggtggtgatgatg tttcttctccggtagtgtaacgg
	pET28b _Ntag	BT1636_Fw_NheI	ggaata**gctagc** cagaaaaacaacacaaagccgaat
		BT1636_Rv_XhoI	gagata**ctcgag** ttatttcttctccggtagtgtaac
BT1918	pETite_Ctag	BT1918_Fw	gaaggagatatacatatg caaaaacaaccgcacattatcttaattatg
		BT1918_Rv	gtgatggtggtgatgatg tcgcttttctttcggatagttag
BT2914	pET28a_Ntag	BT2914_Fw_NheI	ggaata**gctagc** cagaataataagatgaatgttttg
		BT2914_Rv_XhoI	gagata**ctcgag** ctataattgagatttccggatctt
BT3051	pETite_Ctag	BT3051_Fw	gaaggagatatacatatg caacctgccggacaacgcccc
		BT3051_Rv	gtgatggtggtgatgatg cttattatagtatttttcaagcagtttctgc
BT3057	pETite_Ctag	BT3057_Fw	gaaggagatatacatatg caggaacgcccgaatattattgtcttcc
		BT3057_Rv	gtgatggtggtgatgatg aaagtaaacaaaaggcttcaactcatttcc
	pET28a_Ntag	BT3057_Fw_NheI	ggaata**gctagc** caggaacgcccgaatattattgtct
		BT3057_Rv_XhoI	gagata**ctcgag** ttaaaagtaaacaaaaggcttcaa
BT3093	pETite_Ctag	BT3093_Fw	gaaggagatatacatatg caagaaaagcctaatatcattgttatattggcgg
		BT3093_Rv	gtgatggtggtgatgatg cttatctctatattctttaaagtccttccgg
BT3095	pETite_Ctag	BT3095_Fw	gaaggagatatacatatg cagcgttttgtattacgggattggc
		BT3095_Rv	gtgatggtggtgatgatg ttgttcgggttggagataattataagg

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

BT3101	pETite_Ctag	BT3101_Fw	gaaggagatatacatatg cagcaaatcacgcagcctaatttgg
		BT3101_Rv	gtgatggtggtgatgatg cttttgtttggaaggcatatgatc
BT3106	pETite_Ctag	BT3106_Fw	gaaggagatatacatatg cagcagaatgtatcaaaacctaatattgtcc
		BT3106_Rv	gtgatggtggtgatgatg tttgagatattcttctacaggagtatacc
BT3107	pET28a_Ntag	BT3107_Fw_NheI	ggaata**gctagc** atgatggatggtaattttatatct
		BT3107_Rv_XhoI	gagata**ctcgag** ccatttcctaataataaaaaataa
BT3109	pETite_Ctag	BT3109_Fw	gaaggagatatacatatg tgcacttcacagcaggtagaagaaaagccg
		BT3109_Rv	gtgatggtggtgatgatg ttttaatggttcttcttccacctcac
BT3177	pETite_Ctag	BT3177_Fw	gaaggagatatacatatg acttcggtaaaaaaagaggctgcacagaagcc
		BT3177_Rv	gtgatggtggtgatgatg ttccttgtctctctcaggggag
	pET28a_Ntag	BT3177_Fw_NheI	ggaata**gctagc** atgccggaaggccatcgtgcaaca
		BT3177_Rv_XhoI	gagata**ctcgag** ttattccttgtccctttccggaga
BT3486	pETite_Ctag	BT3486_Fw	gaaggagatatacatatg gatacgaatcccaatgtagttatcatttatatcg
		BT3486_Rv	gtgatggtggtgatgatg gaggtctgtttgcaatcctttatc
BT3487	pETite_Ctag	BT3487_Fw	gaaggagatatacatatg aactcaaaagggaaaaaaactgataaaagaccc
		BT3487_Rv	gtgatggtggtgatgatg tttctttttttgtcccggatagggttc
BT3488	pETite_Ctag	BT3488_Fw	gaaggagatatacatatg caggagaaaccgaatgtcgtactgattctgg
		BT3488_Rv	gtgatggtggtgatgatg ttttttcggatgtttcggtaaaacaaag
BT3489	pET28a_Ntag	BT3489_Fw	ggaata**gctagc** cagcaggtagaacgccctaatatc
		BT3489_Rv	gagata**ctcgag** tcagcgtgttactttccactccac
BT3796	pETite_Ctag	BT3796_Fw	gaaggagatatacatatg cagacacagcgtgtggacaaacg
		BT3796_Rv	gtgatggtggtgatgatg cttaatttcgtccggccaagg
	pET28b _Ntag	BT3796_Fw_NheI	ggaata**gctagc** cagacacagcgtgtggacaaacg
		BT3796_Rv_XhoI	gagata**ctcgag** cttaatttcgtccggccaagg
BT3799	pETite_Ctag	BT3799_Fw	gaaggagatatacatatg cagacgtctccgaatcttgtcttc
		BT3799_Rv	gtgatggtggtgatgatg tttagttacggtttcaagataatgggc
BT4631	pETite_Ntag	BT4631_Fw	catcatcaccaccatcacgagaacctgtacttccagggc tctttgcccgcgtatgcc
		BT4631_Rv	gtggcggccgctctatta tcactttttccctttttgctctttcctg
BT4683	pETite_Ctag	BT4683_Fw	gaaggagatatacatatg agtactcctcccaatattgtattgattttatgtg
		BT4683_Rv	gtgatggtggtgatgatg ttctattccgctttgatccgg
	pET28b _Ntag	BT4683_Fw_NheI	ggaata**gctagc** agtactcctcccaatattgtattg
		BT4683_Rv_XhoI	gagata**ctcgag** ttattctattccgctttgatccgg

Restriction enzyme sequences and cloned gene flanking sequences in bold and underlined, respectively.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Table 8

Primers designed to generate the site-directed mutants of *Bt* sulfatases

Sulfatase Mutant	Primer name	Primer sequence (5' to 3')
BT0756_S82C*	BT0756_S82C_Fw	gcaatgttccggtgtgcggcgcttcccgtgc
	BT0756_S82C_Rv	gcacgggaagcgccgcacaCcggaacattgc
BT1622_S75C*	BT1622_S75C_Fw	cggttcccccgtatgcgctccggcacgct
	BT1622_S75C_Rv	agcgtgccggagcgcatacgggggaaccg
BT1622_E98A	BT1622_E98A_Fw	atccgcttcaacaatgccttagccgaacgcggagcagta
	BT1622_E98A_Rv	tccgcgttcggctaaggcattgttgaagcggatttgcgc
BT1622_H177A	BT1622_H177A_Fw	tgccagcgccagtcggccacatactatcctccttttctg
	BT1622_ H177A _Rv	aggaggatagtatgtggccgactggcgctggcaattgta
BT1622_H255A	BT1622_H255A_Fw	accactcccctgccggccgtctcgcttcaggcacccgaa
	BT1622_H255A_Rv	tgcctgaagcgagacggccggcaggggagtggtccacat
BT1622_N334A	BT1622_N334A_Fw	aatggtcctacgttcgccggaggcagcgattctccctgg
	BT1622_N334A_Rv	agaatcgctgcctccggcgaacgtaggaccattatctga
BT1624_S79C*	BT1624_S79C_Fw	cagtgcagctgctacctgcacgccttcccgctatgcg
	BT1624_S79C_Rv	cgcatagcgggaaggcgtgcaggtagcagctgcactg
BT1628_S84C*	BT1628_S84C_Fw	gttgccaattcattatgtggtcccagccgtgcctg
	BT1628_S84C_Rv	caggcacggctgggaccacataatgaattggcaac
BT1636_S77C*	BT1636_S77C_Fw	caggaagtcccgtctgcgctccttcccgagc
	BT1636_S77C_Rv	gctcgggaaggagcgcagacgggacttcctg
BT1636_E100A	BT1636_E100A_Fw	gtacggggcaacaaggcctattggacgaacgcccccaca
	BT1636_E100A_Rv	ggcgttcgtccaataggccttgttgccccgtacttcgca
BT1636_Q173A	BT1918_Q173A_Fw	ttcggctacatctgtgccttccaggcccacctttactac
	BT1918_Q173A_Rv	aaggtgggcctggaaggcacagatgtagccgaaatattc
BT1636_H177A	BT2914_H177A_Fw	tgtcagttccaggccgccctttactaccctaatttcctc
	BT2914_H177A_Rv	attagggtagtaaagggcggcctggaactgacagatgta
BT1636_R353A	BT3051_R353A_Fw	ctgcgtggcttgaaagcccaatgctatgaaggaggtatc
	BT3051_R353A_Rv	tccttcatagcattgggctttcaagccacgcagtttacc
BT1636_E334A	BT3057_E334A_Fw	aacggccctcacgaagccggcggagctgatcctacattc
	BT3057_E334A_Rv	aggatcagctccgccggcttcgtgagggccgttgtcgct
BT1918_S73C*	BT1918_S73C_Fw	cttcgtgtcccagctgcacccctgcccgtg
	BT1918_S73C_Rv	cacgggcaggggtgcagctgggacacgaag
BT2914_S75C*	BT2914_S75C_Fw	catattgtaacattcctgtatgtggggcttccagggcaagtc
	BT2914_S75C_Rv	gacttgccctggaagccccacatacaggaatgttacaatatg
BT3051_S79C*	BT3051_S79C_Fw	gcaccaactctatctgcgcacccagccgtgcc
	BT3051_S79C_Rv	ggcacggctgggtgcgcagatagagttggtgc
BT3057_S88C*	BT3057_S88C_Fw	cgcacagagcgtgtgctccccttccagagc
	BT3057_S88C_Rv	gctctggaaggggagcacacgctctgtgcg
BT3093_S69C*	BT3093_S69C_Fw	gtataacagtgcgcggtgctgtccgtcacgtgcc
	BT3093_S69C_Rv	ggcacgtgacggacagcaccgcgcactgttatac
BT3095_S73C*	BT3095_S73C_Fw	ctgtagcaccggcctgctctcctgcccgctg
	BT3095_S73C_Rv	cagcgggcaggagagcaggccggtgctacag

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

BT3101_S74C*	BT3101_S74C_Fw	caagctgctcctatgtgctcgcctacccgcc
	BT3101_S74C_Rv	ggcgggtaggcgagcacataggagcagcttg
BT3106_S76C*	BT3106_S76C_Fw	tataaccaggctcggtgcgcgccaaccagggcttc
	BT3106_S76C_Rv	gaagccctggttggcgcgcaccgagcctggttata
BT3107_S78C*	BT3107_S78C_Fw	caattttataatacttctcgttgttgcccaacccgggcaagtc
	BT3107_S78C_Rv	gacttgcccgggttgggcaacaacgagaagtattataaaattg
BT3109_S73C*	BT3109_S73C_Fw	gctacttcggctacttgcacccccagtcgttatgc
	BT3109_S73C_Rv	gcataacgactgggggtgcaagtagccgaagtagc
BT3177_S79C*	BT3177_S79C_Fw	gtggctaattctctgtgtggacccagccgtgcc
	BT3177_S79C_Rv	ggcacggctgggtccacacagagaattagccac
BT3486_S73C*	BT3486_S73C_Fw	cttccgcaccggtttgctcaccctcccgttg
	BT3486_S73C_Rv	caacgggagggtgagcaaaccggtgcggaag
BT3487_S77C*	BT3487_S77C_Fw	ttataatgcgagtcgttgctgtcccacccgtgcttc
	BT3487_S77C_Rv	gaagcacgggtgggacagcaacgactcgcattataa
BT3488_S70C*	BT3488_S70C_Fw	ttataatacttcccgttgctgtccggcacgtgcaagtc
	BT3488_S70C_Rv	gacttgcacgtgccggacagcaacgggaagtattataa
BT3489_S72C*	BT3489_S72C_Fw	ctatacagctcctatctgcacgcccactcgtgcc
	BT3489_S72C_Rv	ggcacgagtgggcgtgcagataggagctgtatag
BT3796_S88C*	BT3796_S88C_Fw	gccagcagtatctgctcccccacccgatg
	BT3796_S88C_Rv	catcgggtgggggagcagatactgctggc
BT3799_S81C*	BT3799_S81C_Fw	gcagttatcctgtttgctcacccgcccgtgc
	BT3799_S81C_Rv	gcacgggcgggtgagcaaacaggataactgc
BT4631_S85C*	BT4631_S85C_Fw1	catgctaccgcttctacctgcactccttcccgctatgctttg
	BT4631_S85C_Rv1	caaagcatagcgggaaggagtgcaggtagaagcggtagcatg
BT4683_S73C*	BT4683_S73C_Fw	gttcaaaaatacaggaagatgttgccctagccgtgctgc
	BT4683_S73C_Rv	gcagcacggctagggcaacatcttcctgtatttttgaac
BT4683_R74A	BT4683_R74A_Fw	ttcaaaaatacaggagcctgctgccctagccgtgctgct
	BT4683_R74A_Rv	acggctagggcagcaggctcctgtatttttgaacaggct
BT4683_H219A	BT4683_H219A_Fw	gctcattatgcccctgccttacccctccaagctcctgcc
	BT4683_H219A_Rv	agcttggaggggtaaggcaggggcataatgagctacgta
BT4683_E335A	BT4683_E335A_Fw	aacggtgcaaccaaagccggcggatacctgggccagctt
	BT4683_E335A_Rv	gcccaggtatccgccggctttggttgcaccgttatcgct
BT4683_W505A	BT4683_W505A_Fw	ctcgaaaacaaacctgccaccgaaagaattaactactat
	BT4683_W505A_Rv	gttaattctttcggtggcaggtttgttttcgagtggaaa

*Cysteine mutants recombinant expressed in *E. coli* and used in all enzymatic activity assay as “active proteins”

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Table 9

Primers designed to generate the in-frame gene deletions and complementations of *Bt* sulfatases

Mutant	Vector	Primer name	Primer sequence (5' to 3')
Δbt1596	pExchange-tdk	bt1596_F1_SalI_Fw	gcg**gtcgac** ggacacgacgtagtgatcttcctt
		bt1596_F1_Rv	ggcaagcaatcttcactctttttatagaagaatgg a taaattgttttcattcttgg
		bt1596_F2_Fw	ttctataaaaaagagtgaagattgcttgcc
		bt1596_F2_Xba_Rv	gcg**tctaga** ggagtccagctgtttcagatactc
Δbt1622	pExchange-tdk	bt1622_F1_SalI_Fw	ctc**gtcgac** gacggaactctccgtttc
		bt1622_F1_Rv	gttagttagcaaggaaggtcggatgttcttttt tttaatcccatt
		bt1622_F2_Fw	aatgggattaaaaaaaagaacatccgaccttcct tgctaactaac
		bt1622_F2_Xba_Rv	ctc**tctaga** gtggcaggtattcagctttc
Δbt1624	pExchange-tdk	bt1624_F1_SalI_Fw	ctc**gtcgac** agactgggaggctgtttc
		bt1624_F1_Rv	gctactctctaagcatttctttttgtcttttgat aattattagg
		bt1624_F2_Fw	cctaataattatcaaaagacaaaaagaaatgct tagagagtagc
		bt1624_F2_Xba_Rv	ctc**tctaga** cagaaagctgaatacctg
Δbt1628	pExchange-tdk	bt1628_F1_SalI_Fw	ctc**gtcgac** ccgatgatagaggtgaag
		bt1628_F1_Rv	ttgtcattgtattattcgggggaatgattaattt atttagattacatggc
		bt1628_F2_Fw	Gccatgtaatctaaataaattaatcattcccccga ataatacaatgacaa
		bt1628_F2_Xba_Rv	ctc**tctaga** aacctcctaccggagtg
*bt1636 _S77A**	pExchange-tdk	bt1636_S77A_F1_SalI_F w	ctc**gtcgac** ctgaccaaaccgttcgaag
		bt1636_S77A_F1_Rv	tgaagctcgggaaggagcggcgacgggacttcctg cgtat
		bt1636_S77A_F2_Fw	atacgcaggaagtcccgtcgccgctccttcccgag cttca
		bt1636_S77A_F2_Xba_ Rv	ctc**tctaga** gtcgctggagaaaatgacc
:bt1636#	pNBU2-bla-tetQb	bt1636_comp_NdeI_Fw	tgagc**catatg** aactacaaattattttctttagtaagcggtacgc
		bt1636_comp_XbaI_Rv	ctgac**tctaga** ttatttcttctccggtagtgtaa cggag
Δbt3051	pExchange-tdk	bt3051_F1_SalI_Fw	gcg**gtcgac** ccagactgtaatcttcatact
		bt3051_F1_Rv	cagttcagtttccgtaccagtcacatacgtatag atgttgcatttctt
		bt3051_F2_Fw	gtgactggtacggaaactgaactg
		bt3051_F2_Xba_Rv	gcg**tctaga** tccgtgaatcgttgtctacag
Δbt3057	pExchange-tdk	bt3057_F1_SalI_Fw	gcg**gtcgac** ggaggtgtcttcaccattaatcta
		bt3057_F1_Rv	gacagttgaaatagtagcacctattacatagta tttatgatttatatccta
		bt3057_F2_Fw	ataggtgctactatttcaactgtc
		bt3057_F2_Xba_Rv	gcg**tctaga** ggattggaatccagcaactctttg

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Δbt3109	pExchange-tdk	bt3109_F1_SalI_Fw	ctc**gtcgac** gaacggtaaacggtatc
		bt3109_F1_Rv	Ctttccagaaggggttatcgggcaatattata tagtttataatgatgc
		bt3109_F2_Fw	gcatcattataaactatataatattgcccgat aaccccttctggaaag
		bt3109_F2_Xba_Rv	ctc**tctaga** cagagtttcgggctta
Δbt3177	pExchange-tdk	bt3177_F1_SalI_Fw	ctc**gtcgac** gtggaacgggttctccg
		bt3177_F1_Rv	ttcgtcatgatagtcagttttattatttattta ttcattgatttaaaagaaatc
		bt3177_F2_Fw	gatttcttttaaatcaatgaataaataaataat aaaactgactatcatgacgaa
		bt3177_F2_Xba_Rv	ctc**tctaga** cgataggcgtaggttctc
Δbt3333	pExchange-tdk	bt3333_F1_SalI_Fw	ctc**gtcgac** cccgttgcaacaatctgg
		bt3333_F1_Rv	caaatctgctattctcattggttacataatgtt ggtcttttatggag
		bt3333_F2_Fw	ctccataaaagaccaacattatgtaaccaatg agaatagcagatttg
		bt3333_F2_Xba_Rv	ctc**tctaga** cgagcgactcgatatcg
Δbt3349	pExchange-tdk	bt3349_F1_SalI_Fw	ctc**gtcgac** atgtccggtgattatcg
		bt3349_F1_Rv	ccgtattagtatacggtatcacatgatataaag aataacagtta
		bt3349_F2_Fw	taactgttattctttatatcatgtgataccgta tactaatacgg
		bt3349_F2_Xba_Rv	ctc**tctaga** gcaaggggacatcata
Δbt3796	pExchange-tdk	bt3796_F1_SalI_Fw	ctc**gtcgac** ccaatacactcatactggg
		bt3796_F1_Rv	cccgactaaacaggtgaataaggattaattatt tagtgttatttcatta
		bt3796_F2_Fw	taatgaaataacactaaataattaatccttatt cacctgtttagtcggg
		bt3796_F2_Xba_Rv	ctc**tctaga** ttctcctggatggctaag
Δbt4631	pExchange-tdk	bt4631_F1_SalI_Fw	ctc**gtcgac** cattgattgacgaatataatcag
		bt4631_F1_Rv	gccatatattttgttgtttcaataaatggtgtt taaatcaaataaggttttgag
		bt4631_F2_Fw	ctcaaaaccttatttgatttaaacaccatttat tgaaacaacaaaatatatggc
		bt4631_F2_Xba_Rv	ctc**tctaga** gtacgtgtcccatcctg
Δbt4656	pExchange-tdk	bt4656_F1_SalI_Fw	ctc**gtcgac** gattacacaaatccgtgtttc
		bt4656_F1_Rv	gatatgattgttttgggttatatatccgcatggaaat ctgg
		bt4656_F2_Fw	ccagatttccatgcggatatataacccaaaacaatca tatc
		bt4656_F2_Xba_Rv	ctc**tctaga** tccggtaacattcagataaatg

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

		bt4631_F1_SalI_Fw	ctc**gtcgac** cattgattgacgaatataatcag
Δ*bt4683*	pExchange-tdk	bt4631_F1_Rv	gccatatattttgttgtttcaataaatggtgtt taaatcaaataaggttttgag
		bt4631_F2_Fw	ctcaaaaccttatttgatttaaacaccatttat tgaaacaacaaaatatatggc
		bt4631_F2_Xba_Rv	ctc**tctaga** gtacgtgtcccatcctg

*catalytic residue was mutated to alanine to knockdown the sulfatase activity of this protein;
#complementation of *bt1636* deletion; Restriction sites highlighted in bold.

Supplementary Table 10

	BT1622$^{3S\text{-}Gal/GalNAc}$	BT1622$^{3S\text{-}Gal/GalNAc}$ with GalNAc	BT1636$^{3S\text{-}Gal}$ with LacNAc	BT4683$^{3S\text{-}Gal}$ with LacNAc
Beamline	I0-4	PROXIMA1	I0-4	I0-4
Date	08/07/18	02/09/20	07/08/18	15/02/19
Wavelength (Å)	0.98	0.98	0.98	0.98
Resolution (Å)	95.93-2.60(2.71-2.60)	70.12-2.30 (2.36-2.30)	51.7-1.50 (1.53-1.50)	66.5-1.63 (1.66-1.63)
Space group	$P4_32_12$	$P4_32_12$	$P2_12_12_1$	$P2_12_12_1$
Unit-cell parameters				
a, b, c (Å)	81.2, 81.2, 383.7	83.3, 83.25, 390.4	74.3, 87.9, 103.3	55.1, 87.8, 101.8
α , β γ, (°)	90, 90, 90	90, 90, 90	90, 90, 90	90, 90 ,90
No. of measured reflections	294628 (29026)	116409 (8620)	749320 (37283)	119073 (5859)
No. of independent reflections	40144 (4376)	62514 (4532)	107725 (5296)	62376 (3024)
Completeness (%)	97.8 (100)	100 (100)	99.1 (99.0)	99.9 (100)
Redundancy	7.3 (6.6)	1.9 (1.9)	7.0 (7.0)	1.9 (1.9)
<I>/<σ(I)>	5.5 (1.4)	9.7 (1.5)	11.2 (1.4)	7.8 (1.4)
CC(1/2)	0.987 (0.344)	0.999 (0.840)	0.997 (0.656)	0.999 (0.795)
Refinement statistics*				
R_{work}/R_{free} (%)	0.22/0.24	0.21/0.23	13/16	15/21
No. of protein, atoms	7570	7581	3921	3982
No. of solvent atoms	29	167	400	319
No. of ligand atoms	22	49	26	49
R.m.s. deviation from ideal values				
Bond angle (°)	1.3	1.3	1.6	1.6
Bond length (Å)	0.004	0.004	0.010	0.010
Average B factor (Å^2)				
Protein	61.5	62.3	23.7	23.7
Solvent	39.6	46.9	35.2	33.1
Ligand	54.84	53.1	54.4	45.8
Ramachandran plot$^+$, most favoured regions (%)	96.0	96.7	96.7	97.4
PDB code	7ANB	7ANA	7AN1	7ALL

Values in parenthesis are for the highest resolution shell. R_{free} was calculated using a set (5%) of randomly selected reflections that were excluded from refinement.

Supplementary Table 11

Melting temperatures of galactose targeting sulfatases with and without ligands.

Substrate	concentration	Melting temperature (°C)						
		BT1622$^{3S-Gal/GalNAc}$	BT1636^{3S-Gal}	BT4683^{3S-Gal}	BT3333$^{6S-Gal/GalNAc}$	BT1624$^{6S-Gal/GalNAc}$	BT3109^{6S-Gal}	BT4631$^{6S-Gal/GalNAc}$
No ligand		44.87 ± 0.25	57.88 ± 0.32	47.56 ± 0.11	58.51 ± 0.04	56.12 ± 0.07	53.43 ± 0.02	54.71 ± 0.15
Gal	1 mM	44.59 ± 0.25	58.58 ± 0.41	47.61 ± 0.14	58.30 ± 0.23	56.29 ± 0.04	54.02 ± 0.09	54.89 ± 0.07
	10 mM	44.70 ± 0.23	60.75 ± 1.05	47.27 ± 0.53	58.43 ± 0.25	57.19 ± 0.24	55.01 ± 0.20	55.10 ± 0.14
	100 mM	45.44 ± 0.27	65.98 ± 3.72	48.08 ± 0.10	53.39 ± 0.05	58.92 ± 0.05	56.13 ± 0.11	56.02 ± 0.16
GalNac	1 mM	44.90 ± 0.44	58.32 ± 0.36	47.5 ± 0.26	58.96 ± 0.48	56.06 ± 0.30	53.24 ± 0.04	54.68 ± 0.04
	10 mM	45.28 ± 0.31	58.26 ± 0.35	47.84 ± 0.14	61.49 ± 0.07	57.79 ± 0.13	53.42 ± 0.08	54.96 ± 0.03
	100 mM	47.75 ± 0.81	57.80 ± 0.32	47.22 ± 0.10	64.63 ± 0.14	58.97 ± 0.11	53.73 ± 0.01	55.91 ± 0.07
No ligand		45.13 ± 0.09	55.09 ± 0.10	51.12 -± 0.19	-	-	-	-
3'S-LacNac	0.025 mM	45.10 ± 0.08	56.08 ± 0.13	52.4o ± 0.26	-	-	-	-
	0.25 mM	45.38 ± 0.09	58.44 ± 0.09	55.27 ± 0.29	-	-	-	-
	2.5 mM	45.38 ± 0.09	61.55 ± 0.07	58.67 ± 0.33	-	-	-	-
3'S-LNB	0.025 mM	45.38 ± 0.12	55.15 ± 0.06	51.54 ± 0.32	-	-	-	-
	0.25 mM	45.44 ± 0.09	55.95 ± 0.08	51.92 ± 0.20	-	-	-	-
	2.5 mM	45.98 ± 0.06	58.93 ± 0.09	54.56 ± 0.22	-	-	-	-
No ligand		47.13 ± 0.26	55.91 ± 0.17	51.12 -± 0.19	-	-	-	-
3'S-Lewis^{x}	0.025 mM	46.97 ± 0.26	55.52 ± 0.20	51.45 ± 0.40	-	-	-	-
	0.25 mM	47.16 ± 0.27	57.20 ± 0.16	52.71 ± 0.74	-	-	-	-
	2.5 mM	47.25 ± 0.29	60.03 ± 0.08	57.20 ± 0.19	-	-	-	-
No ligand		46.41 ± 0.03	55.07 ± 0.07	51.12 -± 0.19	-	-	-	-
3'S-Lewis^{a}	0.025 mM	46.37 ± 0.04	55.13 ± 0.07	51.20 ± 0.20	-	-	-	-
	0.25 mM	46.34 ± 0.05	55.73 ± 0.10	51.31 ± 0.16	-	-	-	-
	2.5 mM	45.99 ± 0.20	57.74 ± 0.15	52.29 ± 0.51	-	-	-	-

All reactions were perfomed in Bis-Tris-Propane, pH 7.0, supplemented with 150 mM NaCl. All reactions are technical triplicates with errors being standard errors of the mean. (Gal) Galactose, (GalNAc) *N*-acetyl-D-galactosamine, (LacNAc) *N*-acetyl-lactosamine; (LNB) lacto-*N*-biose.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Table 12

S1_20 homologues of BT1636$^{3S\text{-}Gal}$

SulfAtlas entry	Bacterium species	Locus tag	Isolate
A0A069QTS3	*Prevotella loescheii* DSM 19665	HMPREF1991_00620	Oral cavity
A0A096BUS5	*Prevotella melaninogenica* DNF00666	HMPREF0661_08115	Human (vagina)
A0A098YN99	*Prevotella timonensis* S9-PR14	HMPREF9304_14080	Human (vagina)
A0A098YT81	*Prevotella timonensis* S9-PR14	HMPREF9304_01765	Human (vagina)
A0A099BU15	*Prevotella sp.* S7 MS 2	HMPREF0671_11000	Human (vagina)
A0A0K1NKV2	*Prevotella fusca* JCM 17724	ADJ77_05790	Host (subgingival)
A0A134BL06	*Prevotella sp.* DNF00663	HMPREF3034_01987	Human (vagina)
A0A1B1I5R7	*Prevotella scopos* JCM 17725	AXF22_00890	Host (oral cavity)
A0A1B1SBL2	*Muribaculum intestinale*	A4V02_11165	Mouse (gut)
A0A1C5U5I0	uncultured *Bacteroides sp.*	atsA_5 SAMEA3545269_02720	Human (feces)
A0A1C5UPC8	uncultured *Bacteroides sp.*	atsA_3 SAMEA3545311_02120	Human (feces)
A0A1C5YWH3	uncultured *Bacteroides sp*.	atsA_5 SAMEA3545309_02670	Human (feces)
A0A1F0TL61	*Prevotella sp.* HMSC073D09	HMPREF2955_03450	Human (lungs)
A0A1M5DJ19	*Bacteroides faecichinchillae*	SAMN05444349_12916	Human (feces)
A0A1M6EIV9	*Bacteroides stercorirosoris*	SAMN05444350_109121	Human (feces)
A7LSB2	*Bacteroides ovatus* ATCC 8483	BACOVA_00697	Human
B3C7H9	*Bacteroides intestinalis* DSM 17393	BACINT_00870	Human (feces)
B5D3C1	*Bacteroides plebeius* DSM 17135	BACPLE_03515	Human (feces)
B6VXI1	*Bacteroides dorei* DSM 17855	BACDOR_01990	Human (feces)
C9KV90	*Bacteroides finegoldii* DSM 17565	BACFIN_06228	Human (feces)
C9MNT2	*Prevotella veroralis* F0319	HMPREF0973_01270	Human (oral)
C9PTV0	*Prevotella sp.* oral taxon 472 str. F0295	HMPREF6745_0389	Human (oral)
D1VZ37	*Prevotella timonensis* CRIS 5C-B1	HMPREF9019_1482	Human (urogenital tract)
D3I2J1	*Prevotella melaninogenica* D18	HMPREF0660_00136	Human (oral)
D3IDU7	*Prevotella sp*. oral taxon 299 str. F0039	HMPREF0669_01601	Human (oral)
D3IKS6	*Prevotella sp.* oral taxon 317 str. F0108	HMPREF0670_02095	Oral
D9RR92	*Prevotella melaninogenica* ATCC 25845	HMPREF0659_A5130	Human (sputum)
E6MSG7	*Prevotella salivae* DSM 15606	HMPREF9420_2435	Human (oral)
E6SQJ1	*Bacteroides helcogenes* ATCC 35417	Bache_0950	Pig (Abscess)
F3PXR8	*Bacteroides fluxus* YIT 12057	HMPREF9446_03566	Human (gut)
F8N7E6	*Prevotella multisaccharivorax* DSM 17128	Premu_0918	Human (oral)
G1VBK7	*Prevotella sp*. C561	HMPREF0666_00790	Human (Upper respiratory tract)
G1WC24	*Prevotella oulorum* F0390	HMPREF9431_01375	Human
G5GCE3	*Alloprevotella rava* F0323	HMPREF9332_01220	Human (oral)

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

G6AUB5	*Prevotella stercorea* DSM 18206	HMPREF0673_00199	Human (feces)
I0TBD8	*Prevotella sp.* oral taxon 306 str. F0472	HMPREF9969_1074	Human
I4A3E0	*Ornithobacterium rhinotracheale* ATCC 51463	Ornrh_2345	Turkey (respiratory tract)
I8UTS1	*Bacteroides caccae* CL03T12C61	HMPREF1061_04278	Human (feces)
I9URB5	*Bacteroides xylanisolvens* CL03T12C04	HMPREF1074_03349	Human (feces)
K0XJ44	*Barnesiella intestinihominis* YIT 11860	HMPREF9448_01724	Human (feces)
K9E4G4	*Bacteroides oleiciplenus* YIT 12058	HMPREF9447_02064	Human (feces)
L1MPW2	*Prevotella sp.* oral taxon 473 str. F0040	HMPREF9999_00479	Human
Q5L7L3	*Bacteroides fragilis* ATCC 25285	BF9343_4156	Human (appendix abscess)
Q64MS8	*Bacteroides fragilis* (strain YCH46)	BF4472	Human (feces)
Q8A789	***Bacteroides thetaiotaomicron* ATCC 29148**	**BT_1636**	Human (feces)
R5CEF1	*Bacteroides sp*. CAG:598	BN727_01244	Human (feces)
R5FN99	*Prevotella sp.* CAG:924	BN812_01762	Human
R5GUJ3	*Prevotella sp.* CAG:755	BN773_00514	Human
R6A723	*Prevotella sp.* CAG:5226	BN693_01879	Human
R6AXI7	*Prevotella sp.* CAG:604	BN731_00867	Human
R6C161	*Prevotella copri* CAG:164	BN510_00715	Human (feces)
R6CZP5	*Bacteroides coprocola* CAG:162	BN509_01894	Human
R6VFS1	*Bacteroides faecis* CAG:32	BN607_00824	Human
R6VQE6	*Alistipes sp.* CAG:268	BN576_01026	Human (feces)
R7GWH5	*Prevotella stercorea* CAG:629	BN741_01110	Human
R7KKS2	*Bacteroides thetaiotaomicron* CAG:40	BN644_01632	Human
R7LBR5	*Prevotella sp.* CAG:891	BN805_00880	Human
R7PAA8	*Prevotella sp.* CAG:617	BN736_00581	Human
R7PCA3	*Prevotella sp.* CAG:617	BN736_01655	Human
R9H324	*Bacteroides vulgatus* dnLKV7	C800_03863	Mouse (feces)
R9IKJ4	*Bacteroides sartorii*	C802_00436 E5339_11605	Mouse (gut)
S0F422	*Bacteroides coprophilus* DSM 18228	BACCOPRO_00171	Human (feces)
U2JII2	*Prevotella sp.* F0091	HMPREF9148_00350	Human (oral)
U2KTU7	*Prevotella salivae* F0493	HMPREF9145_0178	Human (oral)
U2NM46	*Prevotella baroniae* F0067	HMPREF9135_1146	Human (?)
U6R7U0	*Bacteroides massiliensis* B84634	HMPREF1534_03795	Human (blood)
W0EPN9	*Barnesiella viscericola* DSM 18177	BARVI_00180	Chicken (cecum)
W4P634	*Bacteroides pyogenes* JCM 6292	JCM6292_731	Pig
X6Q5W1	*Prevotella sp*. ICM33	HMPREF1505_2359	

In all proteins the residues implicated in BT1636$^{3S\text{-}Gal}$ substrate specificity (E100, H179, E334, R353) are fully conserved

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Table 13

S1_20 homologues of BT1622$^{3S\text{-}Gal/GalNAc}$

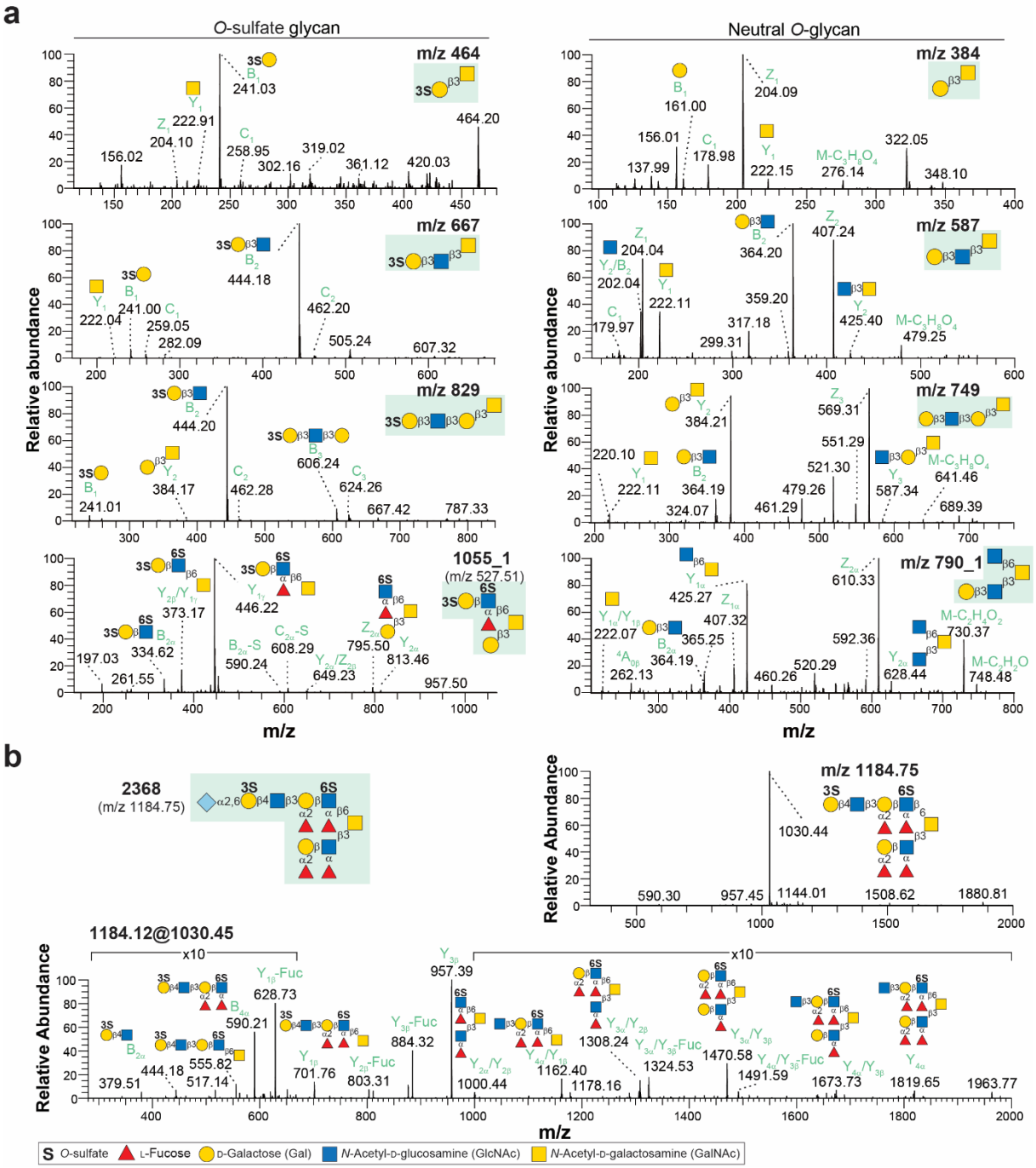
SulfAtlas entry	Bacterium species	Locus tag	Isolate
A0A0F5J3Y2	Parabacteroides sp. HGS0025	HMPREF1212_00321	
A0A0F5JB90	Parabacteroides gordonii DSM 23371	HMPREF1536_02590	Human (blood)
A0A0J6CEB7	Parabacteroides goldsteinii 910340	ACM15_21935	Human (blood)
A0A0J6CPL0	Parabacteroides goldsteinii 910340	ACM15_03535	Human (blood)
A0A0F5JG52	Parabacteroides goldsteinii DSM 19448	HMPREF1535_02433	Human (blood)
S0GPL4	Parabacteroides goldsteinii dnLKV18	C803_04370	Mouse (feces)
K6A8P6	Parabacteroides goldsteinii CL02T12C30	HMPREF1076_00302	Human
A0A0I9TML0	Bacteroides fragilis AM18-6	DW228_02015	Human (feces)
Q5LAT4	Bacteroides fragilis ATCC 25285	BF9343_3008	Human (appendix abscess)
Q64R82	Bacteroides fragilis YCH46	BF3255	Human (feces)
U6RFP1	Bacteroides massiliensis B84634	HMPREF1534_01949	Human (blood)
U6RDB9	Bacteroides massiliensis B84634	HMPREF1534_02027	Human (blood)
Q8A7A3	**Bacteroides thetaiotaomicron VPI-5482**	**BT_1622**	Human (feces)
E6SR73	Bacteroides helcogenes strain ATCC 35417	Bache_0047	Pig (Abscess)
R5C3W0	Bacteroides sp. CAG:598	BN727_02041	Human (feces)
R7KUB8	Bacteroides thetaiotaomicron CAG:40	BN644_03376	Human
R7KIB9	Bacteroides thetaiotaomicron CAG:40	BN644_01646	Human
W4PE04	Bacteroides pyogenes DSM 20611	JCM6294_837	Pig (feces)
A0A174DUP3	Alistipes finegoldii 2789STDY5608890	ERS852447_01898 (AtsA)	Human (feces)
R6Z6X0	Alistipes sp. CAG:29	BN590_00806	Human (gut)
A0A1H4D201	Alistipes timonensis JC136	SAMN05444145_105110	Human (feces)
R6YJX7	Alistipes sp. CAG:435	BN655_01401	Human (gut)
A0A0F5IQ73	Parabacteroides gordonii DSM 23371	HMPREF1536_05149	Human (blood)
I8UX74	Bacteroides caccae CL03T12C61	HMPREF1061_03710	Human (feces)
S0F4Z6	Bacteroides coprophilus DSM 18228	BACCOPRO_00801	Human (feces)
R5CCQ2	Bacteroides sp. CAG:598	BN727_00697	Human (feces)
A0A0F5J6A8	Parabacteroides sp. HGS0025	HMPREF1212_01172	

In all proteins the residues implicated in BT1622$^{3S\text{-}Gal/GalNAc}$ substrate specificity (E98, H176, Q172 and N334) are fully conserved

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Figure 1

Characterization of negatively charged O-glycans from porcine colonic mucins using LC-MS/MS



a, MS/MS spectra of selected *O*-sulfate and neutral non-sulfated O-glycans glycan (left and right hand side panels, respectively). **b,** MS^2 (m/z 1184.75, $[M-2H]^{2-}$) and MS^3 (m/z 1184.12@1030.45) spectra with a composition of NeuGc1Hex3HexNAc4deHex4Sul2.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Figure 2

Phylogenetic analysis of S1_4 subfamily showing the conservation of BT4683$^{3S\text{-}Gal}$ specificity residues

Provided as separate PDF file due to size

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Figure 3

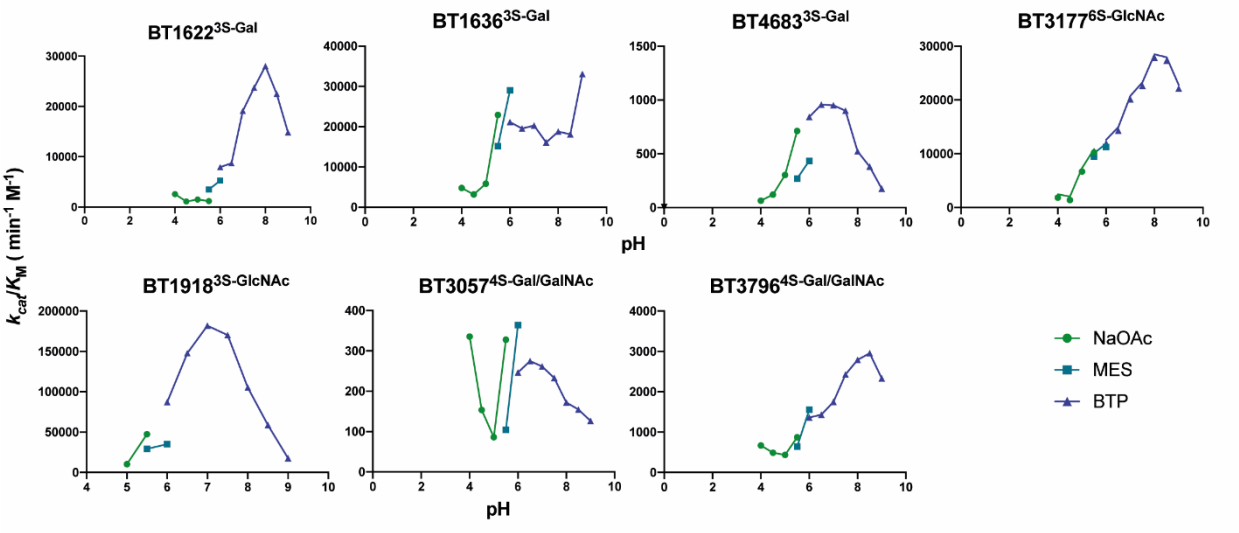
Phylogenetic analysis of S1_20 subfamily showing the conservation of BT1636$^{3S\text{-}Gal}$ specificity residues

Provided as separate PDF file due to size

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Figure 4

pH profile of different sulfatases



Graphics showing the pH optimum for BT1636$^{3S\text{-}Gal}$ and BT3057$^{4S\text{-}Gal/GalNAc}$ is 6; BT4683$^{3S\text{-}Gal}$ had a pH optimum of 6.5; BT19183S-GlcNAc had a pH optimum of 7.0; BT1622$^{S\text{-}Gal}$ and BT3177$^{6S\text{-}GlcNAc}$ had a pH optimum of 8, whilst BT3796$^{4S\text{-}Gal/GalNAc}$ had a pH optimum of 8.5. All reactions were performed using 100 mM of the appropriate buffer supplemented with 150 mM NaCl and 5 mM $CaCl_2$. A substrate concentration of 1 μM was used and an enzyme concentration of between 0.1 – 2 μM was deployed depending on the enzyme.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Suplementary Fig. 2. Phylogenetic tree of representative sulfatases from subfamily S1_4. The annotations concern the presence or absence of conservation of the following residues R72, E335 and W505, which are crucial in substrate recognition by BT4683 (acc-code Q89YP8; i.e. R72 means an equivalent arginine is present, and for example R72I means that the R-equivalent position is replaced by an isoleucine); the branches have the same colour if a pattern is in common; the colour codes are the same and as defined in Extended Fig 7. Red filled circles designate sequences of S1_4 sulfatases from *B. thetaiotaomicron*.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.20.392076; this version posted November 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Suplementary Fig. 2. Phylogenetic tree of representative sulfatases from subfamily S1_20 The annotations concern the presence or absence of conservation of the following residues E100, Q173 H177, E334, R353, which are crucial in substrate recognition by BT1636 (acc-code Q8A789; i.e. E100 means an equivalent glutamate is present, and for example E334A means that the E-equivalent position is replaced by an alanine); in a certain number of sequences no obvious ortologous residue can be deduced from the alignment and this is indicated (red text); the branches have the same colour if a pattern is in common; the colour codes are the same and as defined in Extended Fig. 7 . Red filled circles designate sequences of S1_20 sulfatases from *B. thetaiotaomicron*.