1

A single sulfatase is required to access colonic mucin by a gut bacterium

3

4

*[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

10

11

12

[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.*

25

*Correspondence to:

emartens@umich.edu

Alan.Cartmell@liverpool.ac.uk

ana.luis@medkem.gu.se

30

31

32

33

Summary

Humans have co-evolved with a dense community of microbial symbionts that inhabit the lower intestine. In the colon, secreted mucus creates a barrier that separates these microbes from the intestinal epithelium[1]. 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 degradation of the complex *O*-glycans found in mucins. In the distal colon, these glycans are heavily sulfated, but specific sulfatases that are active on colonic mucins have not been identified. Here we show that sulfatases are essential to the utilization of distal colonic mucin *O*-glycans by the human gut symbiont *Bacteroides thetaiotaomicron*. We characterized the activity of 12 different sulfatases produced by this species showing that they are collectively active on all of the known sulfate linkages in *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 a prominent group of gut bacteria, an important process for both normal microbial gut colonization[2] and diseases such as inflammatory bowel disease[3].

Main

The human gut microbiota (HGM) significantly impacts several aspects of intestinal health and disease, including inflammatory bowel disease (IBD)[4] and colorectal cancer (CRC)[5]. In the colon, secreted mucus creates a barrier that separates gut microbes from the intestinal epithelium[1] preventing close contact that can lead to inflammation and CRC if this barrier is either experimentally eliminated[6,7] or has reduced glycosylation[8-11]. A major component of colonic mucus is mucin 2 (MUC2), a glycoprotein containing up to 80% glycans by mass and more than 100 different glycan structures that are *O*-linked to serine or threonine residues[12]. Mucin glycosylation is variable along the gastrointestinal tract with an increase in sulfation in the colon, especially distal colon[13]. In mucins, *O*-linked sulfate may be attached to the 6-hydroxyl of *N*-acetyl-D-glucosamine (6S-GlcNAc) and terminal D-galactose (Gal) sugars at hydroxyl positions 3-, 4- or 6- (3S-, 4S- and 6S-Gal, respectively)[13-15] (**Fig. 1a**). Sulfation often occurs as terminal caps that typically

65 block enzymatic degradation of oligosaccharides. This requires members of the HGM to

66 express appropriate carbohydrate sulfatases to remove these modifications and access

67 *O*-glycan sugars, although enzymes that remove sugars with sulfate attached are an

68 alternate strategy and has been demonstrated in a human *Bifidobacterium*[16]. *Bacteroides*

69 *thetaiotaomicron* (*Bt*) is a dominant member of the human gut microbiota that utilizes *O*-

70 glycans as a sole nutrient source[17]. *Bt* requires active sulfatases for competitive

71 colonization of the wild-type mouse gut[2] and to induce inflammation in genetically

72 susceptible mice[3]. However, the specific sulfatases that mediate these effects remain

73 unknown. Indeed, despite the critical roles of sulfatases in many biological processes,

74 including diseases[18], a significant knowledge gap exists regarding their specificity and

75 mechanisms.

76 　　　　We hypothesized that *Bt* sulfatases play essential roles in *O*-glycan degradation.

77 To test this, we determined the specificities of 12 *Bt* sulfatases on pure glycan substrates

78 or colonic mucin *O*-glycans. We solved the structures for 3 of these enzymes, revealing

79 the basis of substrate specificity. Using molecular genetics, we assessed the

80 contributions of these sulfatases to *Bt* fitness *in vitro* and *in vivo,* unexpectedly revealing

81 that a single enzyme is essential for utilization of sulfated mucin *O*-glycans and plays a

82 large role in gut colonization.

83

84 **Utilization of colonic mucins by HGM species**

85 　　　　Several studies have identified HGM members that are able to utilize porcine

86 gastric mucin *O*-glycans (gMO) as a sole carbon source[17,19,20]. However, this substrate

87 does not adequately reflect the complexity of mucin *O*-glycans found in the colon,

88 especially those with increased sulfation that are lacking in gMO[13]. To identify HGM

89 species that utilize sulfated colonic mucins, we measured the growth of 16 *Bacteroides*

90 type strains, plus 3 *Phocaeicola* strains (previously classified as *Bacteroides*) and

91 *Akkermansia muciniphila*, on highly sulfated porcine colonic mucin oligosaccharides

92 (cMO) extracted from the distal colon. We identified six strains that utilize cMO (**Fig. 1b,**

93 **Extended Data Fig. 1a**). Interestingly, two known mucin-degraders, *A. muciniphila*[21] and

94 *P. massiliensis*[22]*,* grew on gMO but failed to utilize sulfated cMO, highlighting the

95 importance of employing colonic mucins as more relevant substrates. *Bt*, the bacterium

96 with the highest number of sulfatases (28)[2], was one of the strains with the best growth

97 on cMO (**Fig. 1b**), suggesting that sulfatases might play key roles in this ability. Therefore,

98 to understand the role of sulfatases in colonic mucin utilization by HGM bacteria, we

99 focused on biochemical and genetic characterization of the *Bt* enzymes. See

100 **Supplementary Discussion 1** for additional details of GI mucins and bacterial growth.

101

Substrate specificity of Bt sulfatases

103 Sulfatases are classified into four main families (S1 to S4) in the SulfAtlas

104 database according to sequence similarity, catalytic mechanism and fold[23]. Family S1 is

105 currently divided into 72 subfamilies (designated S1_X) and includes sulfatases which

106 operate via a hydrolytic mechanism and utilize a non-genetically coded formylglycine

107 amino acid as its catalytic nucleophile. In *Bt* and other anaerobes, formylglycine is

108 introduced co-translationally by the anaerobic sulfatase maturating enzyme (anSME)[2],

109 which converts a serine or cysteine to formylglycine[24]. The *Bt* genome encodes 28 S1

110 sulfatases classified into twelve different subfamilies (**Supplementary Table 1**). Four *Bt*

111 sulfatases have been characterized and all act on glycosaminoglycan polysaccharides in

112 the extracellular matrix[25-27]. Several of the uncharacterized S1 sulfatases are encoded

113 within polysaccharide utilization loci (PULs) that are known to be upregulated *in vivo* or

114 during growth on gMO[17] and encode other glycoside hydrolase enzymes potentially

115 involved in *O*-glycan degradation (**Extended Data Fig. 2**).

116 We tested 23 of the remaining 24 uncharacterized sulfatases for activity against a

117 panel of sulfated saccharides (**Supplementary Table 2**). These experiments identified

118 activities for 12 sulfatases, revealing that *Bt* has the capacity to remove sulfate from all of

119 the positions in *O*-glycans (**Fig. 1c, Extended Data Figs. 3, 4** and **Supplementary Table**

120 **3**). Among the active enzymes, 5 represent the first activities reported for their respective

121 subfamilies. Two S1_20 members (BT1636 and BT1622) were determined to target 3S-

122 Gal (**Extended Data Fig. 3a,b**), with BT1622 preferentially binding *N*-acetyl-D-

123 galactosamine (GalNAc) (**Extended Data Fig. 3b**) and cleaving 3S-GalNAc (**Extended**

124 **Data Fig. 3c**). This represents the first report of a bacterial sulfatase active on 3S-GalNAc,

125 indicating that this sulfation could exist as a yet unidentified modification of host glycans.

126 Two S1_16 enzymes (BT3796 and BT3057) cleave 4S-Gal/4S-GalNAc (**Extended Data**

127   **Fig. 4a**). One S1_46 enzyme (BT1918) cleaves 3S-GlcNAc, using the *N*-acetyl group as

128   an absolute specificity determinant (**Extended Data Fig. 4b**). Subsequently, we refer to

129   these enzymes by their gene/locus tag number with the corresponding activity in

130   superscript (*e.g.*, BT1636$^{3S\text{-}Gal}$).

131   We also identified sulfatases displaying new activities inside previously

132   characterized subfamilies. These include three S1_15 enzymes (BT1624$^{6S\text{-}Gal/GalNAc}$,

133   BT3109$^{6S\text{-}Gal/GalNAc}$ and BT4631$^{6S\text{-}Gal/GalNAc}$) that extend this family previously only known

134   to include 6S-GalNAc sulfatases (**Extended Data Fig. 3b and 4c**). Two members of S1_4

135   were active on either 3S-Gal (BT4683$^{3S\text{-}Gal}$) (**Extended Data Fig. 3a,b**) or 6S-Gal

136   (BT3487$^{6S\text{-}Gal}$) (**Extended Data Fig. 4c**). Finally, consistent with the activity previously

137   described for S1_11 members, two enzymes were 6S-GlcNAc sulfatases (BT1628$^{6S\text{-}}$

138   ^{GlcNAc} and BT3177$^{6S\text{-}GlcNAc}$) (**Extended Data Fig. 4d**).

139   The activities of some sulfatases were inhibited by the presence of reducing end

140   substitutions, such as fucose, or exhibited preferences for different glycosidic linkages

141   between sugars (**Extended Data Fig. 3a,b** and **Supplementary Table 3)**. This suggests

142   that *Bt* sulfatases are part of more complex multi-enzyme pathways involving glycoside

143   hydrolases and their expansion reflects adaptations to cleave sulfate in many different

144   glycan-specific contexts. See **Supplementary Discussion 2** for additional details of

145   sulfatase activity based on glycan context.

146

147   ***Bt sulfatase activity on cMO***

148   We next tested the activity of *Bt* sulfatases on a cMO fraction that we determined

149   to contain at least 131 different oligosaccharides (**Supplementary Table 4)**. Only 4 of the

150   6 sulfatases tested displayed activity on cMOs (**Fig. 2 and Supplementary Table 4)**.

151   BT1628$^{6S\text{-}GlcNAc}$ and BT3177$^{6S\text{-}GlcNAc}$ removed 6-*O*-sulfate from all GlcNAc structures that

152   presented this modification at the non-reducing end of *O*-glycans, confirming an exo-

153   mode of action (**Fig. 2 and Extended Data Fig. 5a**). BT4683$^{3S\text{-}Gal}$ was active on a smaller

154   subset of structures (**Fig. 2a and Supplementary Table 4)**. We were only able to

155   determine the structure of one of those glycans, unexpectedly revealing that this enzyme

156   is endo-active on sialylated 3S-Gal (**Fig. 2b, Supplementary Fig. 1)**. After incubation

157   with BT1636$^{3S\text{-}Gal}$, we detected 14 new oligosaccharides (**Fig. 2a**) and an overall increase

158 of non-sulfated glycans (**Extended Data Fig. 5a)**. Compared to the non-enzyme treated

159 reference, 36 oligosaccharides could no longer be detected after incubation with

160 BT1636[3S-Gal] (**Fig. 2a and Supplementary Table 4).** We determined the structures of 8

161 of these glycans and all present a terminal 3S-Gal, but in the context of a variety of

162 different glycans built around core 1-4 structures (**Fig. 2c**). Repetition of the cMO

163 degradation experiment with a different batch of cMO supported the key enzymatic

164 findings described above (**Extended Data Fig. 5** and **Supplementary Table 5**). See

165 **Supplementary Discussion 3** for additional details of activity on cMO.

166

167 ### *Structures of 3S-Gal/GalNAc sulfatases*

168   To understand the molecular details of carbohydrate recognition by S1 sulfatases,

169 we determined the crystal structures, with ligand, of the three different 3S-Gal sulfatases

170 from two subfamilies. Consistent with previous structures[25,27], all 3 enzymes display a

171 $\alpha/\beta/\alpha$ topology with a C-terminal sub-domain, and the active site residues interacting with

172 the sulfate group are fully conserved (**Extended data Fig. 6a,b**). Consistent with *exo-*

173 activity observed for both S1_20 enzymes (BT1636[3S-Gal], BT1622[3S-Gal/GalNAc]) the

174 substrates are buried in a deep pocket and only *O1* is solvent exposed **(Fig. 3).** The

175 structure of BT1636[3S-Gal] and BT1622[3S-Gal/GalNAc] revealed that H177 and H176,

176 respectively, coordinate with *O4* of Gal (**Fig. 3**) and mutation to alanine had a major

177 impact on enzyme activity (**Supplementary Table 3**), suggesting that this histidine is the

178 major specificity determinant for galacto- configurations (presenting an axial *O4*) over

179 gluco- configurations (equatorial *O4*). This essential residue is highly conserved (92%)

180 within S1_20 sulfatases (**Extended data Fig. 7a**). BT1636[3S-Gal] also makes strong

181 interactions with *O2* of Gal via R353 and E334. In BT1622[3S-Gal/GalNAc], these residues are

182 replaced by C357 and N334, amino acids that present shorter side chains that

183 accommodate the *C2*-linked *N*-acetyl group found in GalNAc (**Fig. 3**). This is consistent

184 with the ability of BT1622[3S-Gal/GalNAc] to cleave 3S-GalNAc and preferentially bind GalNAc

185 over Gal, while BT1636[3S-Gal] only recognizes Gal (**Extended Data Fig. 3b,c**).

186   In the S1_4 enzyme BT4683[3S-Gal], the interaction with Gal is driven by the residues

187 R72 and E335, conserved in 19% and 62% of analysed sequences, respectively

188 (**Extended Data Fig. 7b**). These residues are spatially equivalent to R353 and E334 in

189 BT1636[3S-Gal] and form hydrogen bonds with *O2* of D-Gal (**Fig. 3**). Disruption of either of

190 these residues eliminates activity (**Supplementary Table 3**). In BT4683[3S-Gal], a sulfatase

191 that does not have any affinity for monosaccharides, the active site is located in an open

192 cleft (**Fig. 3**) that allows accommodation of additional substitutions on Gal (**Extended**

193 **Data Fig. 6c**). This finding is consistent with the apparent endo-activity found using cMO.

194 Together, these structures reveal the key specificity determinants in 3S-Gal/GalNAc

195 sulfatases, highlighting that they have evolved to target sulfate in different contexts in

196 mucins. This is especially true for BT1636[3S-Gal] which utilizes high affinity interactions with

197 both *O2* and *O4* to drive enhanced activity to remove terminal 3S-Gal linkages in diverse

198 cMO structures. See **Supplementary Discussion 4** for additional details of sulfatase

199 structural characterization and phylogeny.

200

201 ***Roles of sulfatases in Bt O-glycan utilization***

202   *Bt* is able to utilize cMO as a sole carbon source (**Fig. 1b**), but the key enzymes

203 involved remain unknown. Deletion of the gene encoding the only anaerobic sulfatase

204 maturating enzyme (*anSME*) eliminates activation of all 28 S1 sulfatases[2] and as

205 expected decreased the ability of *Bt* to grow on cMOs (**Fig. 4a**). Based on this, we

206 generated a series of strains with compounded gene deletions in which one or several

207 groups of sulfatases were eliminated based on their activity. Deletion of all 3S-

208 Gal/GalNAc sulfatases resulted in a growth phenotype similar to Δ*anSME* (**Extended**

209 **Data Fig. 8a**). Interestingly, we observed a similar growth defect when just BT1636[3S-Gal]

210 was deleted, but not the other 3S-Gal sulfatases (**Fig. 4a and Extended Data Fig. 8a**),

211 consistent with the prominent activity of the recombinant form of this enzyme on cMOs.

212 In contrast, a strain with compounded deletions of eight other sulfatases besides BT1636

213 displayed a growth phenotype similar to wild-type (**Fig. 4a**), indicating that these enzymes

214 are not essential. However, a Δ*10X sulf* mutant, which included the deletion of BT1636[3S-

215 Gal] and the two 4S-Gal/GalNAc sulfatases, showed a similar growth defect as Δ*anSME*

216 and Δ*bt1636*. Complementation of this strain with only *bt1636[3S-Gal]* restored growth on

217 cMO to levels similar to wild-type (**Fig. 4a**, **Extended Data Fig. 8a**). Cellular localization

218 experiments revealed that BT1636[3S-Gal], predicted to be a periplasmic protein

219 (**Supplementary Table 6**), is actually located at the cell surface (**Fig. 4b**).

220   To further investigate the role of BT1636[3S-Gal], we analysed the oligosaccharides
221   present in the culture supernatant of the wild-type and Δ*bt1636[3S-Gal]* strains after growth
222   on cMO. Consistent with a robust ability of *Bt* to degrade colonic *O*-glycans, no
223   oligosaccharides were detected in wild-type supernatant (**Supplementary Table 7 and**
224   **Extended Data Fig. 8b**). Compared to the cMO used as substrate, the supernatant of
225   Δ*bt1636[3S-Gal]* showed a 20-fold accumulation of terminal 3S-Gal capped glycans,
226   suggesting they could not be degraded (**Fig. 4c and Supplementary Table 7**). These
227   data, combined with the cell surface location of BT1636[3S-Gal], suggest that this sulfatase
228   is required early in *O*-glycan catabolism by cleaving 3S-Gal from *O*-glycans prior to
229   importing them into the periplasm, where these oligosaccharides will be degraded by
230   additional enzymes and serve as cues for PUL activation. Interestingly, all of the
231   *Bacteroides* species able to utilize cMO (**Fig. 1a**) have homologues of BT1636[3S-Gal],
232   suggesting that this activity could play a key role in mucin utilization by other HGM
233   members (**Supplementary Table 8 and Supplementary Fig. 2**).

234   Finally, to investigate the requirement for specific sulfatases *in vivo*, we utilized
235   gnotobiotic mice in which we competed individual mutants against the wild-type strain to
236   evaluate their colonization fitness. It has been reported that mouse colonic Muc2
237   prominently displays 6S-GlcNAc modifications[28]. However, three mutant strains that
238   either lacked both characterised 6S-GlcNAc sulfatases or these same two enzymes, plus
239   loss of another putative 6S-GlcNAc sulfatase and all three 6S-Gal/GalNAc sulfatases
240   competed equally with wild-type (**Extended Data Fig. 8c**), suggesting that neither of
241   these two sulfatase activities is essential *in vivo*. An overall mild defect was observed with
242   a mutant lacking all 3S-Gal/GalNAc sulfatases (**Fig. 4d**). The fitness defect was
243   exacerbated by eliminating 3S-Gal/GalNAc and 6S-GlcNAc sulfatases together (**Fig. 4d**),
244   suggesting that they synergise *in vivo*. Consistent with its prominent role *in vitro*, a mutant
245   lacking just BT1636[3S-Gal] displayed a strong defect when competed with the wild-type
246   strain (**Fig. 4d**), further suggesting that this enzyme plays an essential role in gut
247   colonization by allowing *Bt* to access 3S-Gal *O*-glycans (see **Supplementary**
248   **Discussion 5-7** for additional details regarding growth, prediction of cellular localization
249   of proteins and competition of mutants *in vivo*).

250

Conclusion

To degrade the complex *O*-glycans found in mucins some HGM bacteria have evolved complex arsenals of degradative enzymes which include sulfatases. Disarming all of the sulfatases in *Bt* via *anSME* deletion results in drastically reduced competitive colonization (**Extended Data Fig. 8c**)[2] and an inability to elicit colitis in an animal model of IBD[3]. While these findings support a critical role for active sulfatases in both normal colonization and inflammation, they do not provide insight into the complexity of catalytic events carried out by these enzymes. In this study, we reveal that *Bt* has a robust ability to grow on highly sulfated mucin oligosaccharides and that it possesses sulfatases capable of removing sulfate groups in all contexts in which they are known to occur in mucin, including novel specificities. Surprisingly, we found that a single key sulfatase is disproportionately important for growth on colonic mucin *O*-glycans. The critical role of BT1636$^{3S\text{-}Gal}$ supports the conclusion that keystone steps exist in the complex pathway of mucin degradation, although given that mucin glycan structures may vary between mammalian hosts these critical steps may eventually need to be validated in humans for therapeutic translation. Nevertheless, establishment of these critical steps is a prerequisite to blocking this complex enzyme pathway and potentially inhibiting mucin-degrading activities in bacteria that contribute to diseases such as IBD.

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 9**). The catalytic serine was mutated to cysteine since *Escherichia coli* is 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-thiogalactopyranoside; cells were cultured for another 16 h at 16°C and 180 rpm. Recombinant proteins were purified to >90% electrophoretic purity by immobilized metal

282   ion affinity chromatography using a cobalt-based matrix (Talon, Clontech) and eluted with
283   imidazole as described previously[25]. For the proteins selected for structural studies,
284   another step of size exclusion chromatography was performed using a Superdex 16/60
285   S200 column (GE Healthcare), with 10 mM HEPES, pH 7.5, and 150 mM NaCl as the
286   eluent, and they were judged to be ≥95% pure by SDS-PAGE. Protein concentrations
287   were determined by measuring absorbance at 280 nm using the respective molar
288   extinction coefficient. When necessary, proteins were then concentrated by centrifugaton
289   using a molecular mass cutoff of 30 kDa. Since only 12 of the 23 recombinantly-expressed
290   proteins showed activity, we cannot rule out that purification of enzymes from an
291   anaerobic bacterium in oxic conditions or poor installation of catalytic formylglycine
292   resulted in inactive enzymes. However, SDS-PAGE analysis suggested that most
293   enzymes were produced in relatively pure and soluble forms (**Supplementary Figure 5**).
294

295   **Site-Directed Mutagenesis**
296   Site-directed mutagenesis was conducted using the PCR-based QuikChange kit
297   (Stratagene) and conducted according to the manufacturer's instructions, using the
298   appropriate plasmid as the template and primers (**Supplementary Table 10**). All
299   mutations were confirmed by DNA sequencing.
300

301   **Sources of purified carbohydrates**
302   All carbohydrates were from Sigma, Carbosynth or Dextra Laboratories. All other
303   chemical reagents were purchased from Sigma. The 3S-GalNAc was chemically
304   synthesized as previously described[29].
305

306   **Mucin purification**
307   Gastric mucin oligosaccharides (gMO) were purified from commercial available
308   porcine gastric mucins (type III, Sigma) as previously described[17]. Colonic mucins
309   oligosaccharides (cMO) were purified from pig distal pig colons and rectum. Briefly, the
310   tissue was open and the fecal contents were carefully removed. The mucosa was
311   scrapped off and mucus was extracted by homogenizing the tissue in at least 5 times
312   volume of extraction buffer (6 M guanidine chloride, 5 mM EDTA, 10 mM NaH_2PO_4, pH

313   6.5) and slow stirring at 4⁰C for 16 h. The solution was spun down at 15,000 rpm and

314   10⁰C for 30 min and supernatant was discharged. The pellets were resuspended in

315   extraction buffer and the process was repeated until the supernatant was clear for at least

316   two extractions. After the extraction the mucins were solubilized by reducing the disulfide

317   bonds. The pellets were resuspended in fresh reduction buffer (6 M guanidine chloride,

318   0.1 M Tris, 5 mM EDTA, pH 8.0) containing 25 mM of 1,4-dithiothreitol and slowly stirred

319   at 37⁰C for 5 h. After this incubation, 62.5 mM of iodoacetamide were added and the

320   solution was stirred slowly in the dark at room temperature for 16 h. The solution was

321   centrifuged at 10,000 rpm at 4⁰C for 30 min and the supernatant containing the solubilized

322   mucins was extensively dialysed into water. Samples were dissolved into 100 mM Tris-

323   HCl pH 8.0 containing 1 mg/ml of trypsin and incubated slowly stirring at 37⁰C for 16 h.

324   The glycans were β-eliminated by adding 0.1 M NaOH and 1 M $NaBH_4$ and incubate the

325   solution at 65⁰C for 18 h. After cooling the solution to room temperature, the pH was

326   adjusted to 7.0 with concentrated HCl and extensively dialysed in water. The released

327   porcine colonic mucin glycans were recovered by lyophilization the solution until

328   completely dry and used in further experiments.

329

330   **HPLC and TLC sulfatase enzymatic assays**

331   The sulfatase activity screen against commercially available sulfated

332   oligosaccharides (**Supplementary Table 2**) was performed with 1 μM of recombinant

333   enzyme and 1 mM of substrate in 10 mM MES pH6.5 with 5 mM CaCl2 for 16h at 37⁰C

334   (pH optima were determined for each enzyme, **Supplementary Figure 6**, and this pH is

335   compatible with all enzymes). Sulfated *N*-acetyl-D-lactosamine and lacto-*N*-biose were

336   generated by incubating the respective sulfated Lewis antigens with 1 µM of α-1,3/1,4-

337   fucosidase BT1625[30] in the same conditions. Reactions were analysed by thin layer

338   chromatography (TLC). Briefly, 2 µL of each sample was spotted onto silica plates and

339   resolved in butanol:acetic acid:water (2:1:1) running buffer. The TLC plates were dried,

340   and the sugars were visualized using diphenylamine stain (1 ml of 37.5% HCl, 2 ml of

341   aniline, 10 ml of 85% H3PO3, 100 ml of ethyl acetate and 2 g diphenylamine) and heated

342   at 100°C for 20 min. When relevant, the enzymatic activity was confirmed by high-

343   performance anionic exchange chromatography (HPAEC) with pulsed amperometric

344   detection using standard methodology. The sugars (reaction substrate/products) were

345   bound to a Dionex CarboPac P100 column and eluted with an initial isocratic flow of 10

346   mM NaOH during 20 min then a gradient of 10-100 mM of NaOH for 20 min at a flow rate

347   of 1.0 ml min^{-1}. The reaction products were identified using the appropriated standards.

348   All experiments were performed in triplicate.

349

350   **Liquid Chromatograph-Electrospray Ionization Tandem Mass Spectrometry**

351   Enzymatic reactions of sulfatases in colonic mucin oligosaccharides and culture

352   supernatant were desalted with graphitized carbon[31]. Reactions with sulfated defined

353   saccharides were reduced and desalted. Briefly, reactions were dried in Speed vac,

354   reconstituted in 20 μL of 50 mM NaOH and 500 mM NaBH$_4$ and incubated at 50°C for 3

355   h. Reactions were cooled on ice, neutralized with 1 μL of glacial acetic acid and desalted

356   using a cation exchange column containing AG$^{®}$50W-X8 resin. All cleaned and desalted

357   reactions were reconstituted in water before analysis by liquid chromatograph-

358   electrospray ionization tandem mass spectrometry (LC-ESI/MS). The oligosaccharides

359   were separated on a column (10 cm × 250 μm) packed in-house with 5 µm porous

360   graphite particles (Hypercarb, Thermo-Hypersil, Runcorn, UK). The oligosaccharides

361   were injected on to the column and eluted with a 0-40 % acetonitrile gradient in 10 mM

362   ammonium bicarbonate over 46 min at a flow rate of 10 μl/min. A 40 cm × 50 µm i.d. fused

363   silica capillary was used as transfer line to the ion source. Samples were analyzed in

364   negative ion mode on a LTQ linear ion trap mass spectrometer (Thermo Electron, San

365   José, CA), with an IonMax standard ESI source equipped with a stainless-steel needle

366   kept at –3.5 kV. Compressed air was used as nebulizer gas. The heated capillary was

367   kept at 300°C, and the capillary voltage was –33 kV. Full scan (*m/z* 380-2,000, two

368   microscan, maximum 100 ms, target value of 30,000) was performed, followed by data

369   dependent MS^2 scans (two microscans, maximum 100 ms, target value of 10,000) with

370   normalized collision energy of 35%, isolation window of 2.5 units, activation q=0.25 and

371   activation time 30 ms). The threshold for MS^2 was set to 300 counts. Data acquisition and

372   processing were conducted with Xcalibur software (Version 2.0.7). Glycans were

373   identified from their MS/MS spectra by manual annotation and validated by available

374   structures stored in Unicarb-DB database (2020-01 version)[32]. *O*-Glycan structural

375 characterization was based on diagnostic fragment ions[33].The schematic glycosidic or

376 cross-ring cleavages were assigned according to the Domon and Costello

377 nomenclacture[34]. For comparison of glycan abundance between samples, the individual

378 glycan structures were quantified relative to the total content by integration of the

379 extracted ion chromatogram peak area using Progenesis QI. The area under the curve

380 (AUC) of each structure was normalized to the total AUC and expressed as a percentage.

381 The LC-MS/MS raw files and annotated structures are submitted to the Glycopost and

382 Unicarb-DB database, respectively.

383

Microfuidic-based enzymatic desulfation assays

385 Sulfated carbohydrates were labelled at their reducing end with BODIPY which

386 has a maximal emission absorbance of ~503 nm, which can be detected by the EZ

387 Reader via LED-induced fluorescence. Non-radioactive mobility shift carbohydrate

388 sulfation assays were optimised in solution with a 12-sipper chip coated with CR8 reagent

389 using a PerkinElmer EZ Reader II system using EDTA-based separation buffer. This

390 approach allows real-time kinetic evaluation of substrate de-sulfation[35]. Pressure and

391 voltage settings were adjusted manually (1.8 psi, upstream voltage:2250 V, downstream

392 voltage: 500 V) to afford optimal separation of the sulfated product and unsulfated

393 substrate with a sample (sip) time of 0.2 s, and total assay times appropriate for the

394 experiment. Individual de-sulfation assays were carried out at 28°C after assembly in a

395 384-well plate in a final volume of 80 µl in the presence of substrate concentrations

396 between 0.5 and 20 µM with 100 mM Bis-Tris-Propane, 150 mM NaCl, 0.02% (v/v) Brij-

397 35 and 5 mM $CaCl_2$. The degree of de-sulfation was calculated by peak integration using

398 EZ Reader software, which measures the sulfated carbohydrate : unsulfated

399 carbohydrate ratio at each individual time-point. The activity of sulfatase enzymes was

400 quantified in 'kinetic mode' by monitoring the amount of unsulfated glycan generated over

401 the assay time, relative to control assay with no enzyme; with sulfate loss limited to ~20%

402 to prevent loss of substrate and to ensure assay linearity. k_{cat}/K_M values, using the

403 equation $V_0=(k_{cat}/K_M)[E][S]$, were determined by linear regression analysis with GraphPad

404 Prism v8.3.0 software. Substrate concentrations were varied to ensure assay linearity,

405 and substrate concentrations present were significantly $<K_M$.

406

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[36] 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 v8.3.0, 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-Gal/GalNAc] and BT1636[3S-Gal]

437   were assessed against 3'S-LacNAc and 3'S-LNB 100 mM Hepes (pH 7.0) was employed
438   instead of BTP, although no difference in the T_m value of the proteins was observed.
439   Three independent assays were performed for each protein and protein ligand
440   combination **(Supplementary table 11)**. Separate DSF assays were conducted to
441   validate the structural integrity of selected inactive sulfatase mutants (**Supplementary**
442   **Figure 7**).

443

444   **Glycan labelling**

445   Sulfated saccharide samples were labelled according to a modification of the
446   method previously described reporting the formation of *N*-glycosyl amines for 4,6-O-
447   benzilidene protected D-gluopyranose monosaccharides with aromatic amines[37]. Briefly,
448   the lyophilised sugar (1 mg) was dissolved in 0.50 ml anhydrous methanol in a 1.5 ml
449   screw-top PTFE microcentrifuge tube. 0.1 mg, BODIPY-FL hydrazide (4,4–difluoro-5,7–
450   dimethyl-4-bora-3a,4a–diaza-s-indacene-3-propionic   acid   hydrazide,
451     $\lambda_{ex./em.\ 505/513}$, extinction coefficient 80,000 $M^{-1}\ cm^{-1}$) was added and the mixture vortexed
452   (1 min), then incubated in darkness at 65ºC for 24 h. The products were then cooled and
453   a portion purified by TLC on silica coated aluminium plates and developed with methanol
454   or 1:1 v/v ethyl acetate/methanol to provide R_f values suitable to allow separation of
455   unreacted label from labelled glycan product. The unreacted BODIPY-FL hydrazide label
456   (orange on the TLC plate) was identified by reference to a lane containing the starting
457   material, allowing differentiation from the putative labelled product (also orange). This
458   latter band was scraped from the plates and extracted in fresh methanol (2 x 0.5 ml), spun
459   for 3 min at 13,000 x *g* and the supernatant was recovered and dried (rotary evaporator)
460   to   recover   the   fluorescent-coloured product   (bright   green   when   dissolved
461   in aqueous solution), which was then employed in subsequent experiments.

462

463   **Crystallization of carbohydrate sulfatases**

464   After purification, all proteins were carried forward in the same eluent as used for
465   the size exclusion chromatography (see Recombinant Protein Production). Sparse matrix
466   screens were set up in 96-well sitting drop TTP Labtech plates (400-nL drops) at 20°C.
467   Initial hits crystals for all proteins were obtained between 20 and 35 mg/ml. For BT1622[3S-]

468   Gal/GalNAc and BT1636[3S-Gal] wildtype *Bt* variants were used, having a Ser at the catalytic

469   formylglycine position, whilst for for BT4683[3S-Gal] the S73C mutant was used. BT1622[3S-

470   Gal/GalNAc] with 20 mM LNB or 100 mM GalNAc crystallised in 20% PEG 3350 and 0.2 M

471   sodium citrate tribasic dihydrate. BT1636[3S-Gal] with 20 mM LacNAc, or 20 mM 3'S-Lewis-

472   a, crystallized in 40% MPD, 5 % PEG 8000, and 0.1 M sodium cacodylate pH 6.5. for

473   BT4683[3S-Gal] with 20 mM LacNAc crystallised in 20% PEG 3350, 0.2 M sodium iodide and

474   BTP pH 8.5. All crystals were cryoprotected with the addition of the ligand they were

475   crystallised with plus 20% PEG 400 and 20% glycerol was used as the cryoprotectant for

476   BT4683[3S-Gal] and BT1622[3S-Gal/GalNAc], respectively. No cryoprotectant was added to

477   BT1636[3S-Gal] crystals. Data were collected at Diamond Light Source (Oxford) on

478   beamlines I03, I04, I04-1 and I24 at 100 K, except for BT1636[3S-Gal] with 3'S-Lewis-a which

479   was collected 'in house' on a MetalJet D2 X-ray source at Newcastle University. The data

480   were integrated with XDS[38], or Xia2[39] 3di or 3dii and scaled with Aimless v0.7.7[40]. Five

481   percent of observations were randomly selected for the R_{free} set. The phase problem for

482   BT1636[3S-Gal] and BT4683[3S-Gal] was solved by molecular replacement using the automated

483   molecular replacement server Balbes[41]; for BT1636[3S-Gal] Balbes used the PDB 1FSU with

484   28 % identity, whilst for BT4683[3S-Gal] Balbes used the PDB 1HDH with 34 % identity. The

485   phase problem for BT1622[3S-Gal/GalNAc] was initially solved using Molrep v11.7.03[42] and

486   BT1636[3S-Gal] (44% identity) as the search model. This gave a partial solution, which could

487   not be fully solved due to twinning. An acceptable model of BT1622[3S-Gal/GalNAc] was

488   constructed to be used to better solve the phase problem and the molecular replacement

489   was re-performed. Models underwent recursive cycles of model building in Coot v0.9.5[43]

490   and refinement cycles in Refmac v5.8.0267[44]. Bespoke ligands were generated using

491   JLigand (in CCP4 v7.1.013)[45]. The protein models were validated using Coot v0.9.5[43] and

492   MolProbity v4.5.1[46], whilst carbohydrate structures were validated using privateer MKIII[47]

493   (**Supplementary Table 12**). Structural figures were made using Pymol v2.2.2, simulated

494   annealing composite omit maps were generated with Phenix v1.19.2[48]; all other programs

495   used were from the CCP4 v7.1.013[49] and CCP4i2 v7.1.013 suite[50]. The data processing

496   and refinement statistics are reported in **Supplementary Table 13**.

497

498   **Circular dichroism of purified proteins**

499   Circular dichroism (CD) spectra were recorded using a J-1100 Jasco CD
500 spectrometer, Spectral Manager II software and a 0.2 mm path length quartz cuvette
501 (Hellma, USA) scanning at 100 nm.min-1 with 1 nm resolution throughout the wavelength
502 range 190 - 260 nm. Reported spectra were the mean of 9 independent scans following
503 calibration against a buffer baseline. Collected data were analysed with Spectral Manager
504 II software prior to processing with GraphPad Prism v8.3.0. Secondary structural
505 prediction was calculated through the BeStSel analysis server[51]. For all samples, we used
506 10 μM of protein and the buffer was 20 mM sodium phosphate buffer at pH 7.0. See
507 **Supplemental Figure 6**.

508

509 **Anaerobic bacterial culture and genetic manipulation**

510   All strains were anaerobic grown at 37 °C in a chamber (10% H_2, 5% CO_2, and
511 85% N_2; Coy Manufacturing, Grass Lake, MI). *Bacteroides* type strains were culture in
512 either tryptone-yeast extract-glucose medium (TYG), brain heart infusion medium or
513 minimal medium (MM)[17] containing an appropriate carbon source. *Bacteroides*
514 *massilliensis* and *Akkermansia muciniphila* were cultured as described before[22,52]. *Bt*
515 strains containing specific gene deletions or inactivated versions of enzymes (BT1636[3S-Gal] S77A) were made by counterselectable allelic exchange as previously described[53].
517 Complemention of deletion strains was performed using pNBU2 vector as previously
518 described[53], containing a constitutive promotor used previously[54]. All primers used to
519 generate the mutants and complementation are listed in **Supplementary Table 14**.
520 Growth of the WT and mutants was measured on an automated plate reader (Biotek) by
521 increase in absorbance at 600 nm in 96-well plates containing 200 μl of minimal media
522 mixed with the respective filter-sterilised (monosaccharide and gMO) or autoclave-
523 sterilised cMO as described before[52]. To achieve consistent growth, all carbon sources
524 were used at 5 mg/ml with exception of gMO that was added in a final concentration of
525 10 mg/ml. Data was collected and analyzed in Gen5 v1.08 (Biotek) or Prism v8.3.0. All
526 growth curves presented are averages and s.e.m of three technical replicates.

527

528 **Immunolabelling of BT1636 in *Bt* cell surface**

529   To empirically test cellular location of BT1636 empirically using fluorescence
530   microscopy, *Bt* cells (Wild type (Δ*tdk*) and Δ*bt1636*[3S-Gal]) were grown to early exponential
531   phase (Abs$_{600nm}$ 0.25–0.35) in rich TYG medium. One ml of the cultures was collected,
532   centrifuged at 13,000 x *g*, and subsequently washed three times in MM with no carbon
533   source. *Bt* cells incubated with cMO for four hours and fixed in 4.5% formalin overnight at
534   4ºC with gentle rocking. Cells were stained with a polyclonal antibody raised in rabbit
535   against purified recombinant BT1636 (BT1636[Ab], Cocalico Biologicals) and detected with
536   an Alexa Fluor® 488-conjugated goat anti-rabbit IgG secondary antibody (Molecular
537   Probes). Images were taken with Zeiss Apotome using the same exposure time between
538   samples.
539

540   **Gnotobiotic Mouse Experiments**

541   All experiments involving animals, including euthanasia via carbon dioxide
542   asphyxiation, were approved by the University Committee on Use and Care of Animals
543   at the University of Michigan (NIH Office of Laboratory Animal Welfare number A3114-
544   01) and overseen by a veterinarian. Groups of 3 to 5, 6-8 week old male and female
545   germfree Swiss Webster mice were randomly assigned to each experiment. 7 days prior
546   gavage the animals diet was switched to a fiber-free diet (Envigo-Teklad TD 130343) that
547   was maintained through all the experiment. At day 0, mice were gavage with equal
548   amount of *Bt* WT strain and mutant and fecal samples were collected at day 2 and every
549   5 days until day 42. At the end-point of the experiment distal small intestine and cecal
550   contents were also collected. The bacteria gDNA extraction and quantification by qPCR
551   of the relative abundance of each strain on the various samples was carried out as
552   described previously[52].
553

554   **Phylogenetic analysis**

555   To maximise sequence coverage, and avoid repetition, we selected 800 and 920
556   representative sequences of subfamily S1_20 (composed of 1356 sequences) and S1_4
557   (composed of 1895 sequences), respectively. To generate the phylogenetic tree of
558   sulfatases from type strains (**Supplementary Table 1**) we selected all the sulfatases
559   reported in SulfAtlas database (328 sequences). The sequences were aligned by MAFFT

560   v.7[55] using L-INS-i algorithm. The multiple sequence alignment was visualized by Jalview

561   software v.11.0[56] and non-aligned regions were removed. For phylogeny, we used 404,

562   364 and 294 positions for the S1_4, S1_20 and type strains, respectively. Phylogeny was

563   made using RAxML v. 8.2.4[57]. The phylogenetic tree was build with the Maximum

564   Likelihood method[58] and the LG matrix as evolutive model[59] using a discrete Gamma

565   distribution to model evolutionary rate differences among sites (4 categories). The rate

566   variation model allowed for some sites to be evolutionarily invariable. The reliability of the

567   trees was tested by bootstrap analysis using 1,000 (S1_4 and S1_20) and 100 (type-

568   strains phylogeny) resamplings of the dataset[60]. All the final global phylogenetic trees

569   were obtained with MEGA v.7[61]. Fifteen S1_0 sequences from the sulfAtlas database

570   were used as an outgroup in S1_4 and S20 phylogeny. The S1_0 sequence

571   (Phosphonate monoester hydrolase / phosphodiesterase) of *Rhizobium leguminosrum*

572   *viciae* 3841 was used as outgroup in the phylogenetic tree of the type strains. Homologs

573   to BT1636[3S-Gal] and BT16223S-Gal are listed in **Supplementary Tables 15** and **16**,

574   respectively. Input fasta and output (.nwk) alignment files are provides as **Source Data**

575   **2**.

576

577 **Quantification and statistical analysis**

578   For *in vivo* competitions, when three or more fecal samples were collected,

579   Student's t tests were performed for each time point in GraphPad Prism v8.3.0 with a

580   paired, two-tailed distribution. When necessary, the statistical analysis for remaining

581   samples in stated in the respective figure legend.

582

583 **Data availability statement**

584   Source Data for all experiments, along with corresponding statistical test values,

585   where appropriate, are provided within the paper and in Supplementary Information. The

586   crystal structure dataset generates have been deposit in the in the Protein Data Bank

587   (PDB) under the following accession numbers: 7ANB, 7ANA, 7AN1, 7OQD, and 7ALL.

588   The MS raw files were deposit in the GlycoPOST database under the following ID:

589   GPST000150 and GPST000196. Glycan structural annotations were deposited to the

590   Unicarb database as accession (https://unicarb-dr.glycosmos.org/references/462). There

are no restrictions on data or biological resource availability. Data and biological resources can be obtained by contacting the corresponding authors.

Code availability statement

No new codes were developed or compiled in this study

Competing interests statement

The authors declare no competing interests.

Acknowledgements

Author contributions

ASL, AC, and ECM designed experiments and wrote the manuscript.

622   ASL and AC cloned, expressed, purified sulfatases and performed the enzymatic assays.

623   AC, DPB, JAL and PAE carried out and analysed kinetic and binding experiments

624   EY,MR, SO performed chemical syntheses

625   AC and AB performed structural biology experiments.

626   CJ, ASL GCH and NGK performed and interpreted analytical glycobiology experiments.

627   ASL, GP, RWPG, SG, SS and NAP performed bacterial growth experiments and

628   analysed in vivo competition data.

629   MC, GM and TB performed sulfatase phylogenetic analyses.

630   All authors read and approved the manuscript.

631

632   **Figure legends**

633

634   **Figure 1. Bacterial growth on colonic mucin and *Bt* sulfatase activities. a,** Example

635   mucin *O*-glycan structures with relevant terminal epitopes highlighted (dashed boxes).

636   Sugars are shown according to the Symbol Nomenclature for Glycan system[62]. **b**, Growth

637   of *Bacteroides*/*Phocaeicola* strains and A*kkermansia muciniphila* on colonic mucin *O*-

638   glycans (cMO) and number of respective encoded S1 sulfatases. The bars represent the

639   average of two independent experiments with different batches of cMO (total biological

640   replicates n = 3 to 6, error bars denote s.e.m.). Bacterial species able to utilize gastric

641   mucin glycans are highlighted in blue. Maximum absorbance is the maximum A_{600nm} for

642   each culture minus the initial absorbance at time 0. The inset to panel **b** shows example

643   growth curves for *B. fragilis* (*Bf*), *B. thetaiotaomicron* (*Bt*), *P. vulgatus* (*Pv*) and *P.*

644   *massilliensis* (*Pm*). **c**, Phylogeny of *Bt* sulfatases showing the 28 S1 sulfatases and their

645   respective substrates where known. Enzymes are color coded by subfamiy with those

646   characterized in this study in bold. * indicates previously characterized activity and arrows

647   point the sulfate preferentially targeted by the respective enzyme. Sulfatases on a shared

648   branch that share more than 86% and 39-58% of sequence identity are highlighted in blue

649   and green background, respectively.

650

651   **Figure 2. Activity of *Bt* sulfatases on colonic mucin *O*-glycans. a,** Representation of

652   *O*-glycans detected by mass spectrometry in cMO (control) and after sulfatase treatment

653 from the lower (top) to the higher (bottom) mass range. **b,** Relative abundance and
654 putative structures for the specific m/z shown in panel a. **c,** Schematic representation of
655 the putative structures that were not detected after treatment with BT1636^{3S-Gal}.

656

657 **Figure 3. Crystal structures of 3S-Gal/GalNAc sulfatases. a,** Stick representation of
658 the residues interacting with targeted sugars, including the putative catalytic residues (in
659 dark red), the calcium ion (grey sphere) and subsites S, 0 and +1 highlighted in red.
660 BT1636^{3S-Gal} and BT4683^{3S-Gal} in complex with LacNAc (D-Gal-β1,4-D-GlcNAc) and
661 BT1622$^{3S-Gal/GalNAc}$ in complex with GalNAc. Note that in BT1636 the S77 occurs in two
662 conformations. Since this in an inactive enzyme, we do not believe these conformations
663 play a role in activity. **b,** Surface representation of the active site pocket. The equivalent
664 Gal/GalNAc specificity residues in BT1636^{3S-Gal} and BT4683^{3S-Gal} are highlighted in red
665 and blue. The open active site of BT4683^{3S-Gal} is highlighted in purple. In all structures the
666 amino acids and ligands are represented as stick. Note that the GlcNAc has been
667 removed in panel b for clarity.

668

669 **Figure 4. BT1636^{3S-Gal} activity is required for the utilization of cMO and competitive**
670 **fitness *in vivo*. a,** Growth of *Bt* wild-type Δtdk (WT), different sulfatase gene-deletion
671 mutants (named "$\Delta btXXXX$") and strains complemented with *bt1636^{3S-Gal}* on cMO and
672 gMO. Note that gMO were used at 10mg/ml final concentration, while cMO were used at
673 5mg/ml due to background turbidity. This reduced concentration and the higher amount
674 of sulfate in cMO account for the lower growth. Line represents the average of biological
675 replicates (n = 3) and error bars denote s.e.m. **b,** Immunofluorescent and differential
676 interference contrast (DIC) microscopy of *Bt* WT and sulfatase mutant staining with
677 polyclonal antibody (Ab) against BT1636^{3S-Gal} (green) and DNA staining with DAPI (blue).
678 **c,** Relative abundance of different *O*-glycans detected by mass spectrometry in
679 $\Delta bt1636^{3S-Gal}$ culture supernatant or cMO in minimal media without bacteria (control), after
680 96h in anaerobic conditions. The mass and associated structures of the 3 most abundant
681 glycans in both samples are shown. **d,** *in vivo* competitions in gnotobiotic mice (n = 7-9)
682 fed fiber-free diet and inoculated with WT and mutants. The fecal relative abundance of

683   each strain was determined at regular intervals until day 42. The relative abundance of

684   time 0 represents the abundance in the gavaged inoculum. At the experimental endpoint

685   the relative abundance was also determined in small intestine and cecum (represented

686   in the histogram bars at the bottom). The error bars denote s.e.m.

687

688   **Extended Data 1. Growth of *Bacteroides* and *Phocaeicola* type strains and**

689   **A*kkermansia muciniphila* in different mucin *O*-glycans. a**, Graphs showing the

690   growth of strains that are able to utilize colonic or gastric *O*-glycans. Growths were

691   performed in minimal media containing the indicated carbon source. **b**, Growth

692   experiments performed identically to panel **a**, but with two species, *P. massiliensis* and

693   *A. muciniphila*, that grow on gMO but not cMO. A control experiment was performed with

694   *A. muciniphila* grown on cMO plus added GlcNAC to verify that cMO does not contain

695   material that is inhibitory to this species (biological replicates n =3 for both panels, error

696   bars denote the s.e.m. for each time point). Note that gMO were used at 10mg/ml final

697   concentration, while cMO were used at 5mg/ml due to background turbidity. This reduced

698   concentration and the higher amount of sulfate in cMO account for the lower growth on

699   this substrate. cMO, colonic mucin *O*-glycans; gMO, gastric mucin *O*-glycans, GlcNAc,

700   *N*-acetyl-D-glucosamine.

701

702   **Extended Data 2. Schematic representation of polysaccharide utilization loci**

703   **(PULs) encoding sulfatases (sulf).** Genes are colour coded according to the predicted

704   function of the respective proteins. Glycoside hydrolases (GH) in known families are

705   indicated by GHXX or GH*, where XX and * indicates the respective family number or

706   non-classified, respectively.

707

708   **Extended Data 3. Activity and affinity of sulfatases to targeted substrates. a,**

709   Recombinant enzymes (1 μM) were incubated with 1 mM of substrate in 10 mM MES

710   pH6.5 with 5 mM CaCl$_2$ for 16h at 37 °C. Sulfated disaccharides were generated by

711   adding 1 μM of a characterized α1,3/1,4-fucosidase (BT1625) in the enzymatic reaction.

712   Control reactions without sulfatases were carried in the same conditions. Samples were

713   analysed by mass spectrometry and the intensity of the substrate and reaction products

714   was used for comparison of the relative abundance of these sugars after incubation with

715   the respective enzymes. **b,** Affinity studies looking at the effect of ligand binding on the

716   melting temperature of 3S and 6S-Gal sulfatases. All reactions were performed in 100

717   mM BTP, pH 7.0 with 150 mM NaCl. For sample melting temperatures see

718   **Supplementary Table 11**. **c,** Activity of 3S-Gal/GalNAc sulfatases (10 μM) against 3S-

719   GalNAc (10 mM). Reactions were performed in 10 mM Hepes, pH 7.0, with 150 mM NaCl

720   and 5 mM CaCl$_2$. The data shown are one representative from the biological replicates

721   conducted (n = 3).

722

723   **Extended Data 4. Enzymatic screen of *Bt* sulfatases using sulfated**

724   **monosaccharides.** Recombinant enzymes (1 μM) were incubated with 1 mM of

725   substrate in 10 mM MES pH6.5 with 5 mM CaCl$_2$ for 16 h at 37ºC. Reactions were

726   analyzed by thin layer chromatography (left side) or HPAEC with pulsed amperometric

727   detection (right side). Control reactions without sulfatases were carried out in the same

728   conditions. The standards in TLC and HPAEC-PAD are labelled on the left side and top,

729   respectively. The different panel represent activities found for sulfatases targeting: **(a)** 4S-

730   Gal/GalNAc; **(b)** 3S-GlcNAc **(c)** 6S-Gal/GalNAc; **(d)** 6S-GlcNAc. The data shown are

731   representative from biological replicates (n = 3).

732

733   **Extended Data 5. Activity of *Bt* sulfatases against colonic mucin *O*-glycans (cMO)**

734   **analysed by mass spectrometry. a,** Relative abundance of structures detected in

735   different samples organized by sulfate-linkage (top panel) or presence of one or several

736   sugar substitutions such as sulfate, sialic acid and fucose (bottom panel). The colour-

737   coded bars represent the relative abundance and the total number of the structures

738   containing the specific linkage/substitution; **b,** Representation of *O*-glycans detected by

739   mass spectrometry in cMO batch 2 (control) and after sulfatase treatment from the lower

740   (top) to the higher (bottom) mass range; **c**, Relative abundance and putative structures

741   for the specific m/z shown in panel b. The putative structure for the different mass is

742   shown on the right side of the graphic. The reactions were performed with 1 μM of enzyme

743   and 0.5% cMO in 10 mM MES pH 6.5 with 5 mM CaCl$_2$ for 16 h at 37ºC. The complete

744   dataset is provided in **Supplementary Table 4 and 5** for cMO batch 1 and 2, respectively.

745

Extended Data 6. Schematic representation of 3S-Gal/GalNAc sulfatases. a,(i) Cartoon representation colour ramped from blue (α/β/α N-terminal domain) to red (β-sheet C-terminal domain); (ii) the final $2mF_{obs}$-DF_{calc} maps contoured at 1σ for GalNAc in BT1622[3S-Gal/GalNAc] (Top) LacNAc in BT1636[3S-Gal] (middle) and BT4683[3S-Gal] (bottom); (iii) represents the simulated annealed composite omit $2mF_{obs}$-DF_{calc} maps contoured at 1σ and (iv) represents the mF_{obs}-DF_{calc} maps, prior to building of the ligand contoured at 3σ;. **b,**(i) Overlay of the active site S residues of BT1636[3S-Gal] (green) BT1622[3S-Gal/GalNAc] (blue) and BT4683[3S-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. The 3S-Gal substrate is from the BT1636[3S-Gal] 3'S-Lewis-a complex, and BT1622[3S-Gal/GalNAc] and BT4683[3S-Gal] structures have been overlaid, (ii) the final $2mF_{obs}$-DF_{calc} maps of the observed 3'S-Lewis-a substrate contoured at 1σ, (iii) represents the simulated annealed composite omit $2mF_{obs}$-DF_{calc} maps contoured at 1σ, and (iv) represents the mF_{obs}-DF_{calc} maps of the observed 3'S-Lewis-a substrate, prior to building of the ligand, contoured at 3σ; **c,** Docking of putative structures of *O*-glycans targeted by BT4683[3S-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 shown as sticks (middle panel) and a schematic is 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.

766

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. 3 and 4**. For 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_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

776 recognition by BT1636 (acc-code Q8A789, coloured red). A grey X in italics specifically

777 designates that the residue E100 is absent in that sequence, and no obvious orthologous

778 residue can be found from the alignment. **b,** Radial representation of the phylogenetic

779 tree constructed with representative sequences of the sulfatase S1_4 subfamily. The

780 colour code is given as a pattern of presence or absence of the residues R72, E335 and

781 W505, which are crucial in substrate recognition by BT4683 (acc-code Q89YP8, coloured

782 red). A grey X in italics specifically designates that the residue W505 is absent in that

783 sequence, and no obvious orthologous residue can be found from the alignment.

784

785 **Extended Data 8. Sulfatase activity is required for growth in cMO and *in vivo***

786 **fitness. a,** Growth curves of *Bt* wild-type Δtdk (WT), different sulfatase mutants ($\Delta btXXX$)

787 and complemented strains on glucose, colonic or gastric mucin *O*-glycans (cMO and

788 gMO, respectively). The curves represent the average of biological replicates (n = 3) and

789 the error bars denote s.e.m. **b,** Relative abundance of oligosaccharides detected by mass

790 spectrometry in culture supernatant of WT and $\Delta bt1636^{3S\text{-}Gal}$ after growth in cMO for 96h

791 at anaerobic conditions. The control corresponds to cMO incubated in the same

792 conditions without bacterium. The colours represent the relative abundance of structures

793 grouped according to the presence of epitopes (sulfate, fucose and sialic acid) and the

794 numbers represent the total number of structures that contain the respective substitution.

795 **c,** Colonization of gnotobiotic mice fed a fiber-free diet by *Bt* WT and mutants lacking the

796 full ($\Delta anSME$, no S1 sulfatases active) or specific sulfatase activity ($\Delta 6S\text{-}GlcNAc$ and

797 $\Delta 6S\text{-}GlcNAc+\Delta 6S\text{-}Gal/GalNAc$). The fecal relative abundance of each strain was

798 determined at regular intervals until day 42. The relative abundance of time 0 represents

799 the abundance in the gavaged inoculum. At the experimental endpoint the relative

800 abundance was also determined in small intestine and cecum. The graphs represent the

801 average of n=3-7 and the error bars denote the s.e.m. The relative abundance in each

802 individual animal is represented in a lighter colour in each of the respective graphics.

803

804 **Source Data**

805

806   **Source Data 1.** Excel file (multi-tab) containing all data arrays, statistics and uncropped

807   images in order of presentation.

808

809   **Source Data 2**. Input fasta and output Newick alignment files are provided for the large

810   phylogenetic trees shown in Supplemental Figures 2, 3 and 4.

811

812   **Main References**

813

814 1   Johansson, M. E. *et al.* The inner of the two Muc2 mucin-dependent mucus layers in
815   colon is devoid of bacteria. *Proceedings of the National Academy of Sciences of the*
816   *United States of America* **105**, 15064-15069, doi:10.1073/pnas.0803124105 (2008).

817 2   Benjdia, A., Martens, E. C., Gordon, J. I. & Berteau, O. Sulfatases and a radical S-
818   adenosyl-L-methionine (AdoMet) enzyme are key for mucosal foraging and fitness of the
819   prominent human gut symbiont, Bacteroides thetaiotaomicron. *J Biol Chem* **286**, 25973-
820   25982, doi:10.1074/jbc.M111.228841 (2011).

821 3   Hickey, C. A. *et al.* Colitogenic Bacteroides thetaiotaomicron Antigens Access Host
822   Immune Cells in a Sulfatase-Dependent Manner via Outer Membrane Vesicles. *Cell Host*
823   *Microbe* **17**, 672-680, doi:10.1016/j.chom.2015.04.002 (2015).

824 4   Packey, C. D. & Sartor, R. B. Commensal bacteria, traditional and opportunistic
825   pathogens, dysbiosis and bacterial killing in inflammatory bowel diseases. *Current*
826   *opinion in infectious diseases* **22**, 292-301 (2009).

827 5   Sears, C. L. & Garrett, W. S. Microbes, Microbiota, and Colon Cancer. *Cell Host*
828   *Microbe* **15**, 317-328, doi:10.1016/j.chom.2014.02.007 (2014).

829 6   Van der Sluis, M. *et al.* Muc2-deficient mice spontaneously develop colitis, indicating
830   that Muc2 is critical for colonic protection. *Gastroenterology* **131**, 117-129, doi:Doi
831   10.1053/J.Gastro.2006.04.020 (2006).

832 7   Velcich, A. *et al.* Colorectal cancer in mice genetically deficient in the mucin Muc2.
833   *Science* **295**, 1726-1729, doi:10.1126/science.1069094 (2002).

834 8   Bergstrom, K. *et al.* Core 1- and 3-derived O-glycans collectively maintain the colonic
835   mucus barrier and protect against spontaneous colitis in mice. *Mucosal immunology*,
836   doi:10.1038/mi.2016.45 (2016).

837 9   Larsson, J. M. *et al.* Altered O-glycosylation profile of MUC2 mucin occurs in active
838   ulcerative colitis and is associated with increased inflammation. *Inflamm Bowel Dis* **17**,
839   2299-2307, doi:10.1002/ibd.21625 (2011).

840 10   Fu, J. *et al.* Loss of intestinal core 1-derived O-glycans causes spontaneous colitis in
841   mice. *J Clin Invest* **121**, 1657-1666, doi:10.1172/JCI45538 (2011).

842 11   Kudelka, M. R. *et al.* Cosmc is an X-linked inflammatory bowel disease risk gene that
843   spatially regulates gut microbiota and contributes to sex-specific risk. *Proceedings of the*
844   *National Academy of Sciences of the United States of America* **113**, 14787-14792,
845   doi:10.1073/pnas.1612158114 (2016).

846 12   Larsson, J. M., Karlsson, H., Sjovall, H. & Hansson, G. C. A complex, but uniform O-
847   glycosylation of the human MUC2 mucin from colonic biopsies analyzed by
848   nanoLC/MSn. *Glycobiology* **19**, 756-766, doi:10.1093/glycob/cwp048 (2009).

849   13   Holmen Larsson, J. M., Thomsson, K. A., Rodriguez-Pineiro, A. M., Karlsson, H. &
850   Hansson, G. C. Studies of mucus in mouse stomach, small intestine, and colon. III.
851   Gastrointestinal Muc5ac and Muc2 mucin O-glycan patterns reveal a regiospecific
852   distribution. *American journal of physiology. Gastrointestinal and liver physiology* **305**,
853   G357-363, doi:10.1152/ajpgi.00048.2013 (2013).
854   14   Thomsson, K. A. *et al.* Detailed O-glycomics of the Muc2 mucin from colon of wild-
855   type, core 1- and core 3-transferase-deficient mice highlights differences compared with
856   human MUC2. *Glycobiology* **22**, 1128-1139, doi:10.1093/glycob/cws083 (2012).
857   15   Robbe, C. *et al.* Evidence of regio-specific glycosylation in human intestinal mucins -
858   Presence of an acidic gradient along the intestinal tract. *Journal of Biological Chemistry*
859   **278**, 46337-46348, doi:DOI 10.1074/jbc.M302529200 (2003).
860   16   Katoh, T. *et al.* Identification and characterization of a sulfoglycosidase from
861   Bifidobacterium bifidum implicated in mucin glycan utilization. *Bioscience,*
862   *biotechnology, and biochemistry* **81**, 2018-2027, doi:10.1080/09168451.2017.1361810
863   (2017).
864   17   Martens, E. C., Chiang, H. C. & Gordon, J. I. Mucosal Glycan Foraging Enhances Fitness
865   and Transmission of a Saccharolytic Human Gut Bacterial Symbiont. *Cell Host Microbe*
866   **4**, 447-457, doi:DOI 10.1016/j.chom.2008.09.007 (2008).
867   18   Diez-Roux, G. & Ballabio, A. Sulfatases and human disease. *Annu Rev Genomics Hum*
868   *Genet* **6**, 355-379, doi:10.1146/annurev.genom.6.080604.162334 (2005).
869   19   Wlodarska, M. *et al.* Indoleacrylic Acid Produced by Commensal Peptostreptococcus
870   Species Suppresses Inflammation. *Cell Host Microbe* **22**, 25-37 e26,
871   doi:10.1016/j.chom.2017.06.007 (2017).
872   20   Tramontano, M. *et al.* Nutritional preferences of human gut bacteria reveal their
873   metabolic idiosyncrasies. *Nat Microbiol* **3**, 514-522, doi:10.1038/s41564-018-0123-9
874   (2018).
875   21   Derrien, M., Vaughan, E. E., Plugge, C. M. & de Vos, W. M. Akkermansia muciniphila
876   gen. nov., sp. nov., a human intestinal mucin-degrading bacterium. *Int J Syst Evol*
877   *Microbiol* **54**, 1469-1476 (2004).
878   22   Pudlo, N. A. *et al.* Symbiotic Human Gut Bacteria with Variable Metabolic Priorities for
879   Host Mucosal Glycans. *mBio* **6**, e01282-01215, doi:10.1128/mBio.01282-15 (2015).
880   23   Barbeyron, T. *et al.* Matching the Diversity of Sulfated Biomolecules: Creation of a
881   Classification Database for Sulfatases Reflecting Their Substrate Specificity. *PLoS ONE*
882   **11**, e0164846, doi:10.1371/journal.pone.0164846 (2016).
883   24   Benjdia, A. *et al.* Anaerobic sulfatase-maturating enzymes, first dual substrate radical S-
884   adenosylmethionine enzymes. *J Biol Chem* **283**, 17815-17826,
885   doi:10.1074/jbc.M710074200 (2008).
886   25   Cartmell, A. *et al.* How members of the human gut microbiota overcome the sulfation
887   problem posed by glycosaminoglycans. *Proceedings of the National Academy of Sciences*
888   *of the United States of America* **114**, 7037-7042, doi:10.1073/pnas.1704367114 (2017).
889   26   Ndeh, D. *et al.* Metabolism of multiple glycosaminoglycans by Bacteroides
890   thetaiotaomicron is orchestrated by a versatile core genetic locus. *Nature communications*
891   **11**, 646, doi:10.1038/s41467-020-14509-4 (2020).
892   27   Ulmer, J. E. *et al.* Characterization of glycosaminoglycan (GAG) sulfatases from the
893   human gut symbiont Bacteroides thetaiotaomicron reveals the first GAG-specific

894 bacterial endosulfatase. *J Biol Chem* **289**, 24289-24303, doi:10.1074/jbc.M114.573303
895 (2014).
896 28 Tobisawa, Y., Imai, Y., Fukuda, M. & Kawashima, H. Sulfation of colonic mucins by N-
897 acetylglucosamine 6-O-sulfotransferase-2 and its protective function in experimental
898 colitis in mice. *J Biol Chem* **285**, 6750-6760, doi:10.1074/jbc.M109.067082 (2010).
899

900 **Methods References**

901
902
903 29 Egan, M., Jiang, H., O'Connell Motherway, M., Oscarson, S. & van Sinderen, D.
904 Glycosulfatase-Encoding Gene Cluster in Bifidobacterium breve UCC2003. *Appl*
905 *Environ Microbiol* **82**, 6611-6623, doi:10.1128/AEM.02022-16 (2016).
906 30 Briliute, J. *et al.* Complex N-glycan breakdown by gut Bacteroides involves an extensive
907 enzymatic apparatus encoded by multiple co-regulated genetic loci. *Nat Microbiol* **4**,
908 1571-1581, doi:10.1038/s41564-019-0466-x (2019).
909 31 Packer, N. H., Lawson, M. A., Jardine, D. R. & Redmond, J. W. A general approach to
910 desalting oligosaccharides released from glycoproteins. *Glycoconjugate journal* **15**, 737-
911 747, doi:10.1023/a:1006983125913 (1998).
912 32 Hayes, C. A. *et al.* UniCarb-DB: a database resource for glycomic discovery.
913 *Bioinformatics* **27**, 1343-1344, doi:10.1093/bioinformatics/btr137 (2011).
914 33 Everest-Dass, A. V., Abrahams, J. L., Kolarich, D., Packer, N. H. & Campbell, M. P.
915 Structural feature ions for distinguishing N- and O-linked glycan isomers by LC-ESI-IT
916 MS/MS. *J Am Soc Mass Spectrom* **24**, 895-906, doi:10.1007/s13361-013-0610-4 (2013).
917 34 Domon, B. & Costello, C. E. Structure elucidation of glycosphingolipids and
918 gangliosides using high-performance tandem mass spectrometry. *Biochemistry* **27**, 1534-
919 1543, doi:10.1021/bi00405a021 (1988).
920 35 Byrne, D. P. *et al.* New tools for carbohydrate sulfation analysis: heparan sulfate 2-O-
921 sulfotransferase (HS2ST) is a target for small-molecule protein kinase inhibitors.
922 *Biochem J* **475**, 2417-2433, doi:10.1042/BCJ20180265 (2018).
923 36 Byrne, D. P. *et al.* cAMP-dependent protein kinase (PKA) complexes probed by
924 complementary differential scanning fluorimetry and ion mobility-mass spectrometry.
925 *Biochem J* **473**, 3159-3175, doi:10.1042/BCJ20160648 (2016).
926 37 Das, T. M., Rao, C. P. & Kolehmainen, E. Synthesis and characterisation of N-glycosyl
927 amines from the reaction between 4,6-O-benzylidene-D-glucopyranose and substituted
928 aromatic amines and also between 2-(o-aminophenyl)benzimidazole and pentoses or
929 hexoses. *Carbohydrate research* **334**, 261-269, doi:10.1016/s0008-6215(01)00202-6
930 (2001).
931 38 Kabsch, W. Xds. *Acta Crystallogr D Biol Crystallogr* **66**, 125-132,
932 doi:10.1107/S0907444909047337 (2010).
933 39 Winter, S. E., Lopez, C. A. & Baumler, A. J. The dynamics of gut-associated microbial
934 communities during inflammation. *EMBO Rep* **14**, 319-327, doi:10.1038/embor.2013.27
935 (2013).
936 40 Evans, P. R. & Murshudov, G. N. How good are my data and what is the resolution? *Acta*
937 *Crystallogr D Biol Crystallogr* **69**, 1204-1214, doi:10.1107/S0907444913000061 (2013).

938   41   Long, F., Vagin, A. A., Young, P. & Murshudov, G. N. BALBES: a molecular-
939   replacement pipeline. *Acta Crystallogr D Biol Crystallogr* **64**, 125-132,
940   doi:10.1107/S0907444907050172 (2008).
941   42   Vagin, A. & Teplyakov, A. Molecular replacement with MOLREP. *Acta Crystallogr D*
942   *Biol Crystallogr* **66**, 22-25, doi:10.1107/S0907444909042589 (2010).
943   43   Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot.
944   *Acta Crystallogr D Biol Crystallogr* **66**, 486-501, doi:10.1107/S0907444910007493
945   (2010).
946   44   Murshudov, G. N. *et al.* REFMAC5 for the refinement of macromolecular crystal
947   structures. *Acta Crystallogr D Biol Crystallogr* **67**, 355-367,
948   doi:10.1107/S0907444911001314 (2011).
949   45   Lebedev, A. A. *et al.* JLigand: a graphical tool for the CCP4 template-restraint library.
950   *Acta Crystallogr D Biol Crystallogr* **68**, 431-440, doi:10.1107/S090744491200251X
951   (2012).
952   46   Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular
953   crystallography. *Acta Crystallogr D Biol Crystallogr* **66**, 12-21,
954   doi:10.1107/S0907444909042073 (2010).
955   47   Agirre, J. *et al.* Privateer: software for the conformational validation of carbohydrate
956   structures. *Nat Struct Mol Biol* **22**, 833-834, doi:10.1038/nsmb.3115 (2015).
957   48   Terwilliger, T. C. *et al.* Iterative-build OMIT maps: map improvement by iterative model
958   building and refinement without model bias. *Acta Crystallogr D Biol Crystallogr* **64**,
959   515-524, doi:10.1107/S0907444908004319 (2008).
960   49   Collaborative computational project, n. The CCP4 suite: Programs for protein
961   crystallography. *Acta Crystallographica* **D50**, 760-763 (1994).
962   50   Potterton, L. *et al.* CCP4i2: the new graphical user interface to the CCP4 program suite.
963   *Acta Crystallogr D Struct Biol* **74**, 68-84, doi:10.1107/S2059798317016035 (2018).
964   51   Micsonai, A. *et al.* BeStSel: a web server for accurate protein secondary structure
965   prediction and fold recognition from the circular dichroism spectra. *Nucleic Acids Res* **46**,
966   W315-W322, doi:10.1093/nar/gky497 (2018).
967   52   Desai, M. S. *et al.* A Dietary Fiber-Deprived Gut Microbiota Degrades the Colonic
968   Mucus Barrier and Enhances Pathogen Susceptibility. *Cell* **167**, 1339-1353 e1321,
969   doi:10.1016/j.cell.2016.10.043 (2016).
970   53   Koropatkin, N. M., Martens, E. C., Gordon, J. I. & Smith, T. J. Starch catabolism by a
971   prominent human gut symbiont is directed by the recognition of amylose helices.
972   *Structure* **16**, 1105-1115, doi:10.1016/j.str.2008.03.017 (2008).
973   54   Degnan, P. H., Barry, N. A., Mok, K. C., Taga, M. E. & Goodman, A. L. Human gut
974   microbes use multiple transporters to distinguish vitamin B(1)(2) analogs and compete in
975   the gut. *Cell Host Microbe* **15**, 47-57, doi:10.1016/j.chom.2013.12.007 (2014).
976   55   Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid
977   multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**,
978   3059-3066, doi:10.1093/nar/gkf436 (2002).
979   56   Clamp, M., Cuff, J., Searle, S. M. & Barton, G. J. The Jalview Java alignment editor.
980   *Bioinformatics* **20**, 426-427, doi:10.1093/bioinformatics/btg430 (2004).
981   57   Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
982   large phylogenies. *Bioinformatics* **30**, 1312-1313, doi:10.1093/bioinformatics/btu033
983   (2014).

984   58   Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach.
985   *Journal of molecular evolution* **17**, 368-376, doi:10.1007/BF01734359 (1981).
986   59   Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol Biol*
987   *Evol* **25**, 1307-1320, doi:10.1093/molbev/msn067 (2008).
988   60   Felsenstein, J. Confidence Limits on Phylogenies: An Approach Using the Bootstrap.
989   *Evolution* **39**, 783-791, doi:10.1111/j.1558-5646.1985.tb00420.x (1985).
990   61   Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics
991   Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol* **33**, 1870-1874,
992   doi:10.1093/molbev/msw054 (2016).
993   62   Neelamegham, S. *et al.* Updates to the Symbol Nomenclature for Glycans guidelines.
994   *Glycobiology* **29**, 620-624, doi:10.1093/glycob/cwz045 (2019).
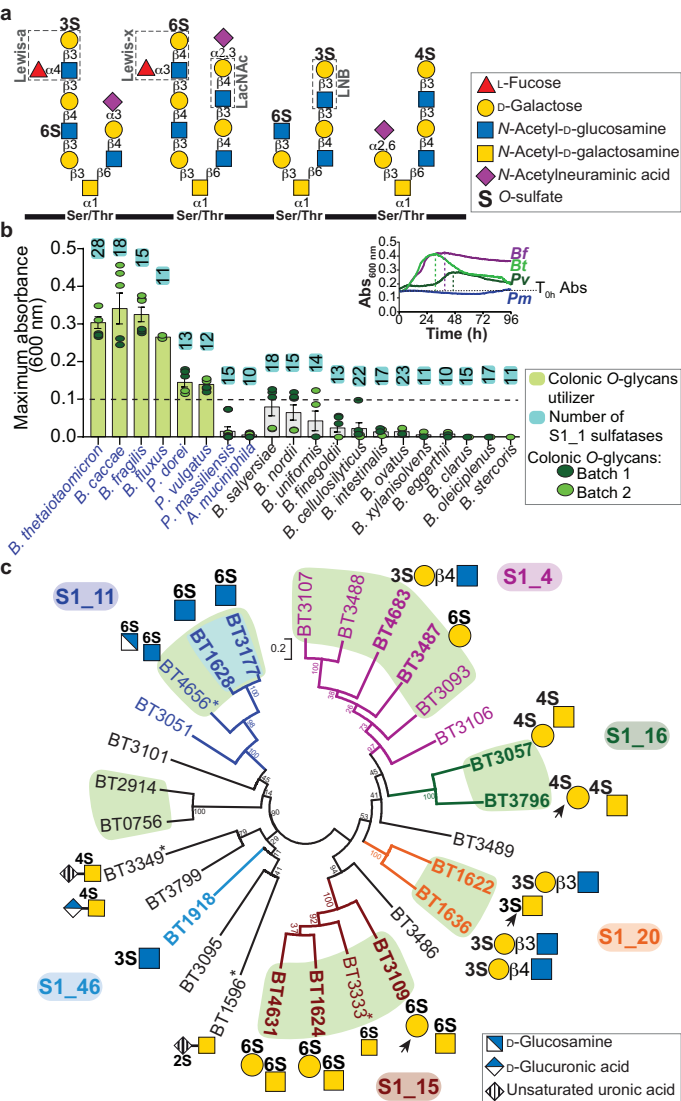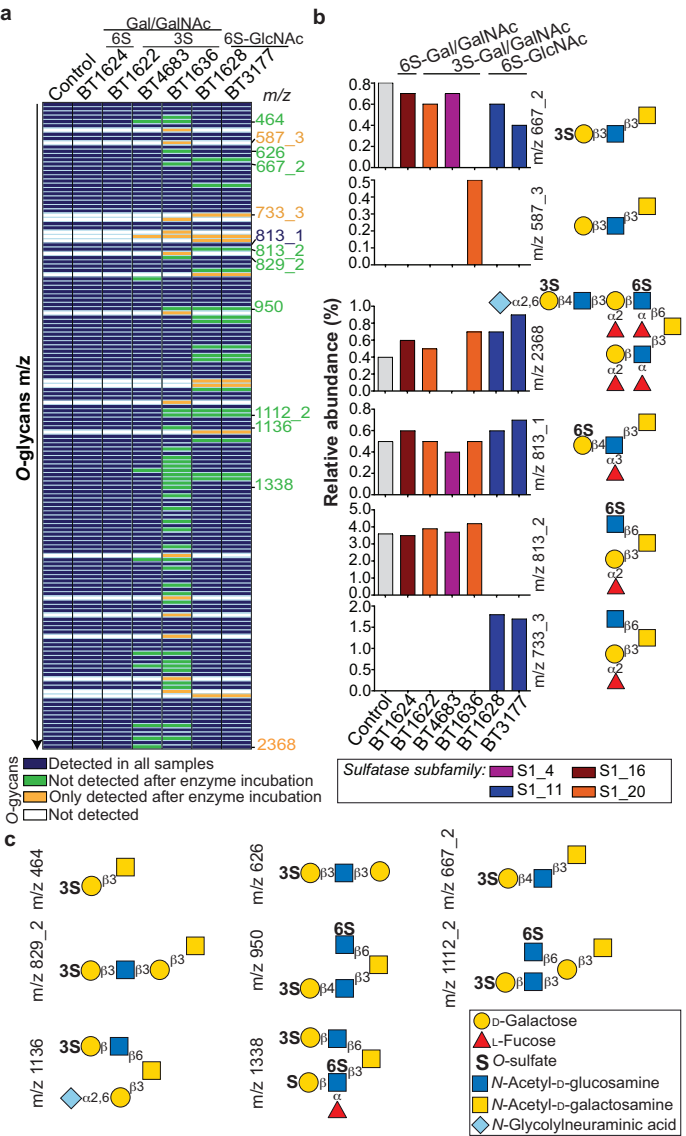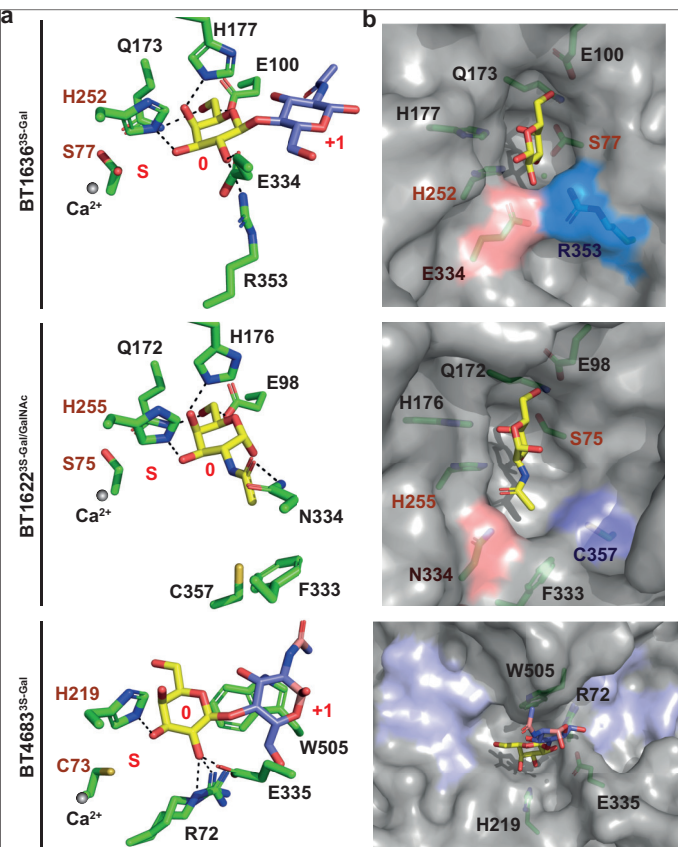995

Fig 1.

Fig 2.

Fig 3.

Fig 4.