1   **Elucidation of a unique sialic acid metabolism pathway in mucus-foraging**

2   ***Ruminococcus gnavus* unravels mechanisms of bacterial adaptation to the gut**

3   Andrew Bell[1], Jason Brunt[1,#], Emmanuelle Crost[1], Laura Vaux[1], Ridvan Nepravishta[2], C.

4   David Owen[3], Dimitrios Latousakis[1], An Xiao[4], Wanqing Li[4], Xi Chen[4], Martin A. Walsh[3], Jan

5   Claesen[5], Jesus Angulo[2], Gavin H. Thomas[6], and Nathalie Juge[1,*]

6   [1]The Gut Microbes and Health Institute Strategic Programme, Quadram Institute Bioscience,

7   Norwich Research Park, Norwich NR4 7UQ, UK

8   [2]School of Pharmacy, University of East Anglia, Norwich Research Park, Norwich, NR4 7TJ,

9   UK

10   [3]Diamond Light Source Ltd, Harwell Science and Innovation Campus, Didcot, OX11 0DE,

11   UK & Research Complex at Harwell, Harwell Science and Innovation Campus, Didcot, OX11

12   0FA, UK.

13   [4]Department of Chemistry, University of California-Davis, Davis, CA 95616, USA.

14   [5]Department of Cardiovascular and Metabolic Sciences, Lerner Research Institute,

15   Cleveland Clinic, Cleveland, OH 44195, USA

16   [6]Department of Biology, University of York, York, YO10 5DD, UK

17   # present address: Department of Chemical Engineering and Biotechnology, University of

18   Cambridge, Philippa Fawcett Drive, Cambridge, CB3 0AS

19   *corresponding author. nathalie.juge@quadram.ac.uk

20

21   **Abstract**

22   Sialic acid (Neu5Ac) is commonly found in terminal location of colonic mucins glycans where

23   it is a much-coveted nutrient for gut bacteria including *Ruminococcus gnavus*. *R. gnavus* is

24   part of the healthy gut microbiota in humans but shows a disproportionate representation in

25   diseases. There is therefore a need in understanding the molecular mechanisms

26   underpinning its adaptation to the gut. Previous *in vitro* work demonstrated that *R. gnavus*

27   mucin glycan-foraging strategy is strain-dependent and associated with the expression of an

28   intramolecular *trans*-sialidase releasing 2,7-anhydro-Neu5Ac instead of Neu5Ac from

29   mucins. Here, we have unravelled the metabolism pathway of 2,7-anhydro-Neu5Ac in *R.

30   gnavus* which is underpinned by the exquisite specificity of the sialic transporter for 2,7-

31   anhydro-Neu5Ac, and by the action of an oxidoreductase converting 2,7-anhydro-Neu5Ac

32   into Neu5Ac which then becomes substrate of a Neu5Ac-specific aldolase. Having

33   generated a *R. gnavus nan* cluster deletion mutant that lost the ability to grow on sialylated

34   substrates, we showed that in gnotobiotic mice colonised with *R. gnavus* wild-type and

35   mutant strains, the fitness of the *nan* mutant was significantly impaired with a reduced ability

36   to colonise the mucus layer. Overall, our study revealed a unique sialic acid pathway in

37 bacteria, with significant implications for the spatial adaptation of mucin-foraging gut
38 symbionts in health and disease.

39

43

Introduction

45 The gastrointestinal (GI) tract is heavily colonized with bacteria that play a vital role in human
46 health. The gut microbiota composition varies longitudinally along the GI tract but also
47 transversally from the mucosa to the lumen[1]. In the colon, the epithelium is covered with a
48 bi-layer of mucus, with the outer mucus layer providing a natural habitat for the commensal
49 bacteria whereas the stratified inner mucus layer restricts bacterial access to the epithelium[2].
50 Mucin proteins that form the mucus layer are highly glycosylated with a diverse and complex
51 array of *O*-glycan structures containing *N*-acetylgalactosamine, galactose and *N*-
52 acetylglucosamine (GlcNAc), and usually terminated by fucose, sialic acid (Neu5Ac)
53 residues, and sulfate[3,4]. The terminal mucin glycans have been proposed to serve as
54 metabolic substrates, providing a nutritional advantage to bacteria that have adapted to the
55 GI mucosal environment[5-7]. The proportion of these terminal glycan epitopes varies along
56 the GI tract with a decreasing gradient of fucose and an increasing gradient of sialic acid
57 from the ileum to the rectum in humans[8,9]. Therefore, sialic acid represents a much-coveted
58 source of nutrients for the gut bacteria inhabiting the mucus niche in the large intestine.

59 In bacteria, the genes involved in sialic acid metabolism are usually found clustered together
60 forming different *nan* gene clusters[10-12]. The canonical *nanATEK* cluster was first described
61 in *Escherichia coli* encompassing genes encoding the enzymes *N*-acetylneuraminate lyase
62 (NanA), epimerase (NanE), and kinase (NanK), necessary for the catabolism of sialic acid
63 into *N*-acetylglucosamine-6-P (GlcNAc-6-P) following its transport through the major
64 facilitator superfamily transporter NanT[13,14]. An alternative pathway for sialic acid metabolism
65 has been discovered later in *Bacteroides fragilis,* relying on the action of an MFS transporter
66 (NanT), an aldolase (NanL), a novel ManNAc-6-P epimerase (also named NanE), encoded
67 in the *nanLET* operon and a hexokinase (RokA), converting Neu5Ac into GlcNAc-6-P[15].
68 GlcNAc-6-P is then converted into fructose-6-P, which is a substrate in the glycolytic
69 pathway by genes encoding NagA (GlcNAc-6-P deacetylase) and NagB (glucosamine-6-P
70 deaminase)[16]. The majority of bacteria that harbour a *nan* cluster colonize mucus regions of

71 the human body[10-12]. To gain access to this substrate, bacteria are dependent on sialic acid

72 release and uptake. Several gut bacteria species, including strains of *Clostridia*, *Bacteroides*,

73 *Bifidobacterium longum*, *Vibrio cholerae*, *Ruminococcus gnavus* or *Akkermansia muciniphila*

74 express sialidases to release sialic acid from their terminal location in mucins[10].

75 Since sialic acid cleavage takes place outside of the cell, bacteria have evolved multiple

76 mechanisms to capture this important nutrient from their environment[12,17]. Such transport

77 mechanisms involve the aforementioned NanT MFS transporter used by *E. coli* and *B.*

78 *fragilis*, which in *E. coli* has been demonstrated biochemically to be a H^+-coupled

79 symporter[18] or secondary transporters from the sodium solute symport (SSS) family, present

80 in *C. difficile* and *S.* typhimurium[19,20]. High-affinity transport of sialic acid is mediated by

81 substrate-binding protein-dependent systems, including a tripartite ATP-independent

82 periplasmic (TRAP) transporter, SiaPQM, and ATP-binding cassette (ABC) transporters[21-25].

83 The sialic acid ABC transporters are classified into 3 types, SAT, SAT2 and SAT3[12,17]. To

84 date all these transporters have been shown to transport Neu5Ac, with some being able to

85 also move the related sialic acid Neu5Gc and KDN[26,27].

86 *R. gnavus* is an early coloniser of the infant gut[28] but persists in adults where it belongs to

87 the 57 species detected in more than 90% of human faecal samples[29]. *R. gnavus* belongs to

88 the Firmicutes division, Clostridia class and XIVa cluster, Lachnospiraceae family[30] and is

89 considered as a prevalent member of the 'normal' gut microbiota[29,31]. Further, *R. gnavus*

90 shows a disproportionate representation in a number of diseases such as inflammatory

91 bowel disease[32-40]. The ability of *R. gnavus* strains to utilise mucin glycans as a source

92 nutrient is associated with the expression of an intramolecular *trans*-sialidase (IT-sialidase)

93 that specifically cleaves off terminal α2–3-linked Neu5Ac from glycoproteins, releasing 2,7-

94 anhydro-Neu5Ac instead of Neu5Ac[41-45]. In *R. gnavus* ATCC 29149 and ATCC 35913

95 strains, the IT-sialidase (*Rg*NanH) is part of a *nan* cluster, which is induced when the cells

96 are grown in the presence of mucin and absent in non-mucin glycan-degrading strains such

97 as *R. gnavus* E1[43,45]. We enzymatically synthesised 2,7-anhydro-sialic acid derivatives[46],

98 that were used to confirm the ability of IT-sialidase expressing *R. gnavus* strains to grow on

99 2,7-anhydro-Neu5Ac as sole carbon source[43]. We proposed that the ability of *R. gnavus*

100 strains to produce and metabolise 2,7-anhydro-Neu5Ac, provide them with a competitive

101 nutritional advantage in mucus by scavenging sialic acid from mucins in a form that others

102 do not have access to[43,44].

103 In order to test this hypothesis and gain insights into *R. gnavus* 2,7-anhydro-Neu5Ac

104 metabolism pathway, we identified candidate genes of the *nan* cluster involved in 2,7-

105 anhydro-Neu5Ac transport and metabolism and characterised the proteins. Using

106 fluorescence spectroscopy, STD-NMR and ITC, we showed that the solute binding protein

107 (SBP) from *R. gnavus* ABC transporter was specific for 2,7-anhydro-Neu5Ac. Further

108 biochemical analyses uncovered an oxidoreductase activity allowing the conversion of 2,7-

109 anhydro-Neu5Ac into Neu5Ac and confirmed the specificity of the sialic acid aldolase for

110 Neu5Ac. Finally, we showed that the *nan* cluster was essential to support anaerobic growth

111 of the bacteria on sialoconjugates *in vitro* and for *in vivo* fitness using gnotobiotic mice

112 colonised with *R. gnavus* wild-type or *nan* mutant. These data demonstrate a unique sialic

113 acid metabolism pathway in bacteria, which provides *R. gnavus* with a competitive strategy

114 to colonise the mucus niche.

115

116 **Results**

117 **Identification of genes involved in 2,7-anhydro-Neu5Ac metabolism in *R. gnavus nan***

118 **cluster**

119 We first analysed the transcriptional activity of the *nan* cluster by qRT-PCR in *R. gnavus*

120 ATCC 29149 grown on 2,7-anhydro-Neu5Ac or α2–3-sialyllactose (3'SL) as the sole carbon

121 source. Expression of all genes constituting the *nan* cluster was induced upon bacterial

122 growth on 2,7-anhydro-Neu5Ac or 3'SL as compared to glucose whereas the expression of

123 the two genes flanking the cluster (RUMGNA_02702, RUMGNA_02690) remained

124 unchanged (**Figure 1**). The 3'SL and 2,7-anhydro-Neu5Ac induced the transcription of the

125 *nan* genes between 10 and 80-fold. Both substrates induced similar changes, which is not

126 unexpected as 2,7-anhydro-Neu5Ac is the sialic acid form produced by *R. gnavus* ATCC

127 29149 from 3'SL. These results indicate that the *R. gnavus nan* operon is dedicated to the

128 metabolism of 2,7-anhydro-Neu5Ac from host sialoglycans.

129 A sequence similarity network (SSN) analysis was then conducted to identify the proteins

130 encoded by the *nan* cluster, which are associated with the ability of the bacteria to

131 metabolise 2,7-anhydro-Neu5Ac over Neu5Ac. As expected, the IT-sialidase from *R. gnavus*

132 strains clustered together with proteins from *S. pneumoniae* strains whose genomes are

133 known to encode IT-sialidases (in addition to other sialidases)[47,48] (**Supplementary Figure**

134 **1a**). Other co-occurring bacterial species include *Rumminococcus torques, Lactobacillus*

135 *salivarius, Staphylococcus pseudintermedius, Streptococcus infantis* and *Streptococcus*

136 *mitis*. Bacterial species clustering for *Rg*NanH, also shared clusters for proteins encoding

137 RUMGNA_02698, the predicted solute binding protein (SBP) giving specificity to ABC

138 transporters, RUMGNA_02692 (sialic acid aldolase), the first protein of the canonical

139 Neu5Ac metabolism, and RUMGNA_02695, a putative oxidoreductase, suggesting that

140 these proteins may be associated with 2,7-anhydro-Neu5Ac metabolism (**Supplementary**

141 **Figure 1 and Supplementary Table 1**). In contrast, RUMGNA_02701 with homology to

142   sialic acid esterase proteins and RUMGNA_02700 with homology to the YhcH protein family
143   did not cluster with proteins from the same set of bacteria (**Supplementary Figure 1 and**
144   **Supplementary Table 1**). The candidate genes were then heterologous expressed, and the
145   recombinant proteins purified as described in Methods.

146

147   **Specificity of *R. gnavus* sialic acid transporter for 2,7-anhydro-Neu5Ac**
148   We first investigated the ligand specificity of the recombinant SBP (RUMGNA_02698),
149   *Rg*SBP, by measuring changes in the intrinsic protein fluorescence upon addition of potential
150   ligands. Addition of 10 μM or 20 μM 2,7-anhydro-Neu5Ac resulted in a significant shift at 350
151   nm, causing an ~16% quench in the fluorescence (**Figure 2a**). In marked contrast, addition
152   of Neu5Ac at 10 μM, 20 μM or 70 μM caused no change in the spectrum intensity,
153   suggesting an absence of binding (**Figure 2b**). Titration of 0.5 μM *Rg*SBP with 2,7-anhydro-
154   Neu5Ac resulted in a hyperbolic curve with a K_d of 1.349 μM (+/- 0.046) (**Figure 2c**). To
155   confirm the specificity of 2,7-anhydro-Neu5Ac over Neu5Ac we monitored sequential
156   changes in fluorescence following additions of 10 μM of the two ligands. When Neu5Ac was
157   added first, no change in fluorescence was observed and a quench was observed following
158   addition of 2,7-anhydro-Neu5Ac (**Figure 2d**). Conversely, when 2,7-anhydro-Neu5Ac was
159   added first the quench was observed and additions of 10 μM Neu5Ac caused no further
160   change in the intensity (**Figure 2d**), indicating that Neu5Ac is unable to displace 2,7-
161   anhydro-Neu5Ac, and further supporting the specificity of the interaction between *Rg*SBP
162   and 2,7-anhydro-Neu5Ac.

163   The affinity of the interaction between *Rg*SBP and sialic acid ligands was further assessed
164   by isothermal titration calorimetry (ITC). *Rg*SBP bound to 2,7-anhydro-Neu5Ac with a K_d of
165   2.42 ± 0.27 μM (**Figure 3a**) and no binding was observed when Neu5Ac was used as the
166   ligand (**Figure 3b**), in agreement with the findings from fluorescence spectroscopy. The
167   binding of 2,7-anhydro-Neu5Ac revealed a thermodynamic signature with both entropic (-
168   TΔS -7.05 ± 0.08 kcal mol^{-1}) and enthalpic (ΔH -0.93 ± 0.03 kcal mol^{-1}) components
169   contributing favourably to the binding process (ΔG -7.99 ± 0.05 kcal mol^{-1} **Figure 3a**).

170   To gain structural insights into the unique ligand specificity of *Rg*SBP, saturation transfer
171   difference nuclear magnetic resonance spectroscopy (STD NMR) studies were conducted
172   with *Rg*SBP in the presence of 2,7-anhydro-Neu5Ac or Neu5Ac. The transfer of
173   magnetization as saturation from the protein to the ligand was clearly observed for 2,7-
174   anhydro-Neu5Ac but not for Neu5Ac, confirming that *Rg*SBP preferentially selects 2,7-
175   anhydro-Neu5Ac (**Supplementary Figure 2**). STD NMR epitope binding revealed that
176   protons H3, H4 and H6 showed the highest STD (%) factors, indicating the close contacts

177   present at the interface of binding **(Figure 3c)**. On the other hand, protons H7, H8, H9 and

178   protons belonging to the CH_3 group showed lower STD (%) and are expected to be more

179   exposed to solvent. For the DEEP-STD NMR experiment, TEMPOL was used to gain

180   insights into *Rg*SBP binding pocket (**Supplementary Figure 3**). We found that protons H4,

181   H6, H7, H8, H9' were preferentially oriented toward aromatic residues while H3 and protons

182   belonging to the CH_3 group were oriented toward aliphatic residues (**Figure 3d**).

183   Together these data demonstrate that *Rg*SBP specifically binds to 2,7-anhydro-Neu5Ac but

184   not to Neu5Ac, in line with the growth profile of *R. gnavus* ATCC 29149 on these

185   substrates[43].

186

187   **Specificity of *R. gnavus* sialic acid aldolase for Neu5Ac**

188   The substrate specificity of recombinant sialic acid aldolase (RUMGNA_02692; *Rg*NanA),

189   was determined using a coupled activity assay where pyruvate released during the

190   conversion of Neu5Ac to ManNAc is converted to lactate by a lactate dehydrogenase and

191   the subsequent decrease in absorbance at 340 nm measured as NADH is converted to

192   NAD^+. *Rg*NanA and *Ec*NanA (*E. coli* Neu5Ac lyase/aldolase used as a control) showed

193   activity against Neu5Ac whilst neither enzyme showed activity against 2,7-anhydro-Neu5Ac

194   (**Figure 4a**). The product of the reaction with Neu5Ac was confirmed to be ManNAc by

195   HPLC (**Supplementary Figure 4**). *Rg*NanA showed a k_{cat} of 2.757 ± 0.033 s^{-1} and a K_M of

196   1.473 ± 0.098 mM (**Figure 4b**). These kinetic parameters are consistent with values from

197   other bacterial sialic acid aldolases characterised to date (**Supplementary Table 2**).

198   The *Rg*NanA crystal structure presents as a (β/α8) TIM barrel with an adjacent three-helix

199   bundle (for data collection and refinement statistics see **Supplementary Table 3**), a fold

200   shared with characterised Neu5Ac aldolases from *Staphylococcus aureus*, *E. coli,*

201   *Fusobacterium nucleatem*, *Pasteurella multocida,* and *Haemophilus influenzae*[49-54]

202   (**Supplementary Figure 5a**). The active site residues in *Ec*NanA, Ser47, Tyr110, and

203   Tyr137, identified to be catalytically important are conserved in *Rg*NanA[55] (**Figure 4c** and

204   **Supplementary Figure 5b**), supporting Neu5Ac specificity. The crystal structure of the

205   complex between an inactive mutant, *Rg*NanA K167A, and Neu5Ac showed Neu5Ac in the

206   open-chain ketone form, with the *N*-acetyl group oriented out of the active site (**Figure 4d**).

207   Neu5Ac forms extensive interactions with the enzyme active site (**Supplementary Table 4**).

208   The Tyr139 α-carbon was shifted 1.8 Å in the mutant compared to wild-type. This movement

209   is also present in the apo crystal structure, therefore presumably due to the absence of

210   Lys167 rather than the presence of Neu5Ac (**Supplementary Figure 5c**).

211

212   **Conversion of 2,7-anhydro-Neu5Ac to Neu5Ac by RUMGNA_02695**

213 RUMGNA_02695 is a putative oxidoreductase with a predicted Rossman fold. Therefore, the
214 activity of the recombinant protein was determined in the presence or absence of NAD$^+$,
215 NADH or FAD as potential cofactors. Reaction products were analysed by HPLC following
216 DMB labelling of the sialic acid[46]. When 2,7-anhydro-Neu5Ac (which cannot be labelled by
217 DMB) was used as a substrate, Neu5Ac was produced in the presence of NAD$^+$ or NADH,
218 but not in the presence of FAD or in the absence of a cofactor (**Figure 5a**). Mass
219 spectrometry analyses showed a ratio of 1:2 for 2,7-anhydro-Neu5Ac:Neu5Ac
220 (**Supplementary Figure 6a**), indicating that the reaction may be reversible. This was
221 confirmed enzymatically by assaying RUMGNA_02695 against Neu5Ac in the presence of
222 NAD$^+$ or NADH, producing a 1:2 for 2,7-anhydro-Neu5Ac:Neu5Ac. Additionally, these data
223 indicate that Neu5Ac is the favourable product (**Supplementary Figure 6b**). No net change
224 in NADH concentration was observed during the conversion reaction using 2,7-anhydro-
225 Neu5Ac or Neu5Ac as substrate, suggesting that the enzyme mechanism may involve
226 oxidation and reduction of NADH cofactor (**Supplementary Figure 6c**). The kinetic
227 parameters of the enzymatic reaction were therefore determined using the coupled reaction
228 described above in the presence of an excess of sialic acid aldolase and increasing
229 concentrations of 2,7-anhydro-Neu5Ac substrate (**Figure 5b**). Using these conditions, the
230 k_{cat} was calculated to be 0.0824 ± 0.0043 s^{-1} and the K_M 0.074 ± 0.014 mM. Taken together
231 these data indicate that RUMGNA_02695 is an oxidoreductase required for the conversion
232 of 2,7-anydro-Neu5Ac into Neu5Ac, which will then become a substrate for *Rg*NanA. We will
233 refer to RUMGNA_02695 as *Rg*NanOx in the rest of the study.

234

235 **Impact of *R. gnavus nan* cluster on *in vitro* growth and *in vivo* colonisation of mice**

236 The ClosTron transformation method[56] was successfully applied to *R. gnavus* ATCC 29149,
237 enabling the generation of *nan* deletion mutants with an erythromycin resistance gene
238 present in either the sense or antisense direction (relative to *Rg*NanH). The recombination
239 event was confirmed by PCR (**Supplementary Figure 7a**) and the expression of the full
240 cluster tested by qPCR (**Supplementary Figure 7b**). The expression of the genes flanking
241 the cluster, RUMGNA_02690 and RUMGNA_02702, showed levels comparable to the wild-
242 type strain, as also observed for the first three genes of the *nan* cluster, RUMGNA_02701-
243 02699, however, the *nan* cluster genes RUMGNA_02698-02691 showed significantly
244 reduced expression compared to the wild-type strain. *R. gnavus* ATCC 29149 wild-type
245 strain was able to utilise both 3'SL and 2,7-anhydro-Neu5Ac as a sole carbon source, while
246 no growth was detected using the *nan* deletion mutants on these substrates
247 (**Supplementary Figure 8**).

248 To assess the impact of the *nan* cluster on the fitness of *R. gnavus in vivo,* germ-free
249 C57BL/6J mice were gavaged with 1×10^8 CFU *R. gnavus* ATCC 29149 or *R. gnavus*
250 antisense *nan* deletion mutant or a mixture of wild-type and *nan* mutant strains at 1×10^8 CFU
251 each (**Figure 6**). During mono-colonisation experiments, both strains were detectable in the
252 faecal content at day 3, 7 and 14 post-gavage at mean levels of between 1×10^6 and 1×10^7
253 bacteria per mg of material (**Figure 6a**). Both strains were also detected in the caecal
254 content of mono-colonised mice sacrificed at day 14. The absence of the *nan* cluster did not
255 affect the mouse expression response, as shown by RNA seq (**Supplementary Figure 9**). In
256 competition experiments, the wild-type strain reached mean colonisation levels comparable
257 to the levels obtained during mono-colonisation, whereas the mutant strain was severely
258 outcompeted, reaching only 2×10^4 copies per mg at day 3, before decreasing further at day 7
259 and day 14 below the level of detection in the faecal and caecal contents (**Figure 6b**). The
260 impact of the *nan* deletion on the location of *R. gnavus* within the mucus layer was
261 determined in mono-colonised mice by measuring the distance of the *nan* mutant or wild-
262 type *R. gnavus* strains to the epithelial layer throughout the colon by fluorescent *in situ*
263 hybridization (FISH) staining using confocal microscopy. The data showed that the *nan*
264 mutant resided 19.70 μm from the epithelial layer, 5.06 μm further away than the wild-type
265 strain, 14.64 μm (**Figure 6c&d**).

266

267 **Discussion**

268 Sialic acid comprises a family of 9-carbon acidic sugar found predominantly on the cell-
269 surface glycans of humans and other animals. Neu5Ac, the most common form of sialic acid
270 in humans, is a major epitope of mucin glycans which can serve as a metabolic substrate to
271 the gut bacteria which have adapted to the mucosal environment[6,10]. *In vivo*, sialic acids may
272 be modified by *O*-acetylation, *O*-methylation [9,10,57-59]. Sialic acid metabolism is vital to the
273 ability of *R. gnavus* strains to utilise mucin as a nutrient source[43-45]. *R. gnavus* ATCC 29149
274 strain encodes an extended *nan* operon dedicated to the metabolism of 2,7-anhydro-Neu5Ac
275 from host sialoglycans.

276 Before being metabolised, a functional sialic acid transporter is essential for the uptake of
277 sialic acid derivatives into the bacterial cell. The *R. gnavus* ATCC 29149 *nan* cluster
278 contains a single ABC transporter, orthologous to the uncharacterised *Steptococcus*
279 *pneumoniae* SAT2 system (Sp_1690-2), including two permeases (RUMGNA_02696 and
280 02697) and *Rg*SBP (RUMNGA_02698). *R. gnavus* SAT2 transporter is expected to be
281 coupled with an MsiK-like ATPase encoded elsewhere in the genome, with
282 RUMGNA_03040 sharing 59% identity with the *S. pneumoniae* MsiK. Interestingly, in

283 contrast to *S. pneumoniae, R. gnavus* does not encode SAT or SAT3 transporters which are
284 known to recognise Neu5Ac with SAT3 being required for growth on Neu5Ac [60-62].

285 By studying *Rg*SBP subunit, we have discovered that SAT2 is a specific transporter for 2,7-
286 anhydro-Neu5Ac with a K_d of 2.42 ± 0.27 µM, which does not bind Neu5Ac. Using STD NMR
287 and DEEP-STD NMR, we characterized the orientation of the ligand in the binding site and
288 the contribution of aromatic and aliphatic residues in *Rg*SBP 2,7-anhydro-Neu5Ac binding
289 pocket. The lower affinity as compared to bacterial SAT (SatA) transporters specific for
290 Neu5Ac characterised to date, which bind in the nM range[63], might be consistent with the
291 'exclusive' access of the bacteria to the 2,7-anhydro-Neu5Ac substrate. Taken together
292 these findings indicate that the ability of *R. gnavus* strains to grow on 2,7-anhydro-Neu5Ac
293 (and not on Neu5Ac) can be explained by the exquisite specificity of *Rg*SBP
294 (RUMGNA_02698) *Rg*SBP is also orthologous (72% identity/86% similarity) with the SBP
295 from the putative sialic transporters in *Streptococcus sanguinis* SK36 (SSA_0076) and
296 *Streptococcus gordonii* str. Challis substr. CH1 (SGO_0122)[12]. It would therefore be of
297 interest to determine the specificity of *Streptococcus* SBPs towards 2,7-anhydro-Neu5Ac.

298 Once inside the cell, 2,7-anhydro-Neu5Ac needs to be converted back into Neu5Ac to
299 become a substrate for the sialic acid aldolase. *Rg*NanOx (RUMGNA_02695) was identified
300 as the oxidoreductase catalyzing the conversion of 2,7-anhydro-Neu5Ac into Neu5Ac,
301 following a mechanism of action which remains to be determined. Bioinformatic analysis
302 identified close homologous of this protein in a range of bacterial species, including YjhC
303 from *E. coli* (**Supplementary Figure 1, Supplementary Table 1**).   Neu5Ac is then
304 converted into ManNAc and pyruvate *via* the action of *Rg*NanA (RUMGNA_02692), a
305 Neu5Ac-specific aldolase with conserved structural features with NanA proteins from the *nan*
306 canonical pathway.

307 MultiGeneBlast analysis revealed that predicted homologs of the *R. gnavus nan* cluster are
308 shared by a limited number of species, including 37 homologous clusters in *S. pneumoniae*
309 isolates, *S. suis* A7, *Blautia hansenii* DSM 20583, *Blautia* sp. YL58 and *Intestinimonas*
310 *butyriciproducens* AF211 (**Supplementary Figure 10 & 11** and **Supplementary Table 5**).
311 The presence of this cluster in *S. pneumoniae* suggests that it can also transport this
312 unusual sialic acid into the cell. A major difference between NanB/NanH IT-sialidase and
313 NanC sialidase cluster types is the associated transporter class, a carbohydrate ABC
314 transporter for NanB/NanH as opposed to a sodium:solute symporter in NanC clusters[47],
315 which may indicate a difference in the form of sialic acid being transported. These
316 bioinformatics analyses support the specialization of the *R. gnavus nan* cluster.

317   We confirmed the importance of this metabolic pathway (**Supplementary Figure 12**) by

318   generating a *R. gnavus nan* deletion mutant that was tested *in vitro* and *in vivo* using germ-

319   free mice. In *in vivo* competition experiments, the fitness of the mutant was impaired as

320   compared to the wild-type strain with a reduced ability to colonise the mucus layer. The *nan*

321   cluster is therefore important to maintain the spatial distribution of *R. gnavus* strains in the

322   gut. The ability for *R. gnavus* strains harbouring a *nan* cluster to penetrate further down into

323   the mucus layer may contribute to protect the bacteria from the constant mucus turner-over.

324   This mechanism may serve as a determinant underlying *R. gnavus* success as one of the

325   most largely shared species among individuals[29,31].

326   Together these findings provide robust biochemical and *in vivo* evidence for the role of *R.*

327   *gnavus nan* cluster in the adaptation of this important gut symbiont to the mucosal

328   environment in the gut, providing defined molecular targets for biomarkers and therapeutic

329   strategies.

330

Methods

Materials

All chemicals were obtained from Sigma (St Louis, USA) unless otherwise stated. D-glucose (Glc), N-acetylneuraminic acid (Neu5Ac), were purchased from Sigma-Aldrich (St Louis, MO). 3′-sialyllactose (3′SL) was purchased from Carbosynth Limited (Campton, UK). 2,7-anhydro-Neu5Ac was prepared as previously described[46,64].

Bacterial strains and media

R. gnavus ATCC 29149 was routinely grown in an anaerobic cabinet (Don Whitley, Shipley, UK) in BHI-YH as previously described[45]. Growth on single carbon sources utilized anaerobic basal YCFA medium[65] supplemented with 11.1 mM of specific mono- or oligosaccharides (2,7-anhydro-Neu5Ac, 3'Sialyllactose (3'SL) or glucose). The bacteria were grown to late exponential phase for RNA extraction, the culture was performed in 14 ml tubes. Growth was determined spectrophotometrically by monitoring changes in optical density at 600 nm compared to the same medium without bacterium (Δ $OD_{595\,nm}$) hourly for 10 hours.

Quantitative real-time PCR (qRT-PCR)

Total RNA was extracted from 3 ml of mid- to late exponential phase cultures of *R. gnavus* ATCC 29149 in YCFA supplemented with one carbon source (Glc, 3′SL or 2,7-anhydro-Neu5Ac). Three biological replicates were performed for each carbon source. The RNA was stabilized prior to extraction by using RNAprotect Bacteria Reagent (Qiagen, Crawley, UK) according to the manufacturer's instructions. The RNA was then extracted after an enzymatic lysis followed by a mechanical disruption of the cells, using the RNeasy Mini Kit (Qiagen) according to manufacturer's instructions with an on-column DNAse treatment. The purity and quantity of the extracted RNA was assessed with NanoDrop 1000 UV-Vis Spectrophotometer (Thermo Fischer Scientific, Wilmington, DE) and with Qubit 2.0 (Invitrogen).

qPCR was carried out in an Applied Biosystems 7500 Real-Time PCR system (Life Technologies Ltd). One pair of primers was designed for each target gene using ProbeFinder version 2.45 (Roche Applied Science, Penzberg, Germany) to obtain an amplicon of around 60–80 bp long. The primers were between 18 and 23 nt-long, with a T_m of 59–60°C (Supplementary **Table S6**). Calibration curves were prepared in triplicate for each pair of primers using 2.5-fold serial dilutions of *R. gnavus* ATCC 29149 genomic DNA. The standard curves showed a linear relationship of log input DNA vs. the threshold cycle

367   (C_T), with acceptable values for the slopes and the regression coefficients (R^2). The

368   dissociation curves were also performed to check the specificity of the amplicons. Each

369   DNAse-treated RNA (1 µg) was converted into cDNA using QuantiTect® Reverse

370   Transcription kit (Qiagen) according to the manufacturer's instructions. DNAse-treated RNA

371   was also treated the same way but without addition of the reverse-transcriptase (RT−). Each

372   qPCR reaction (10 µl) was then carried out in triplicate with 1 µl of 1 ng/µl (cDNA or RT−)

373   and 0.2 µM of each primer, using the QuantiFast SYBR Green PCR kit (Qiagen) according

374   to the manufacturer's instructions (except for the combined annealing/extension step which

375   was extended to 35 s). Data obtained with cDNA were analyzed only when C_T values above

376   36 were obtained for the corresponding RT−. For each cDNA sample, the 3 C_T values

377   obtained for each gene were analyzed using the $2^{-\Delta\Delta CT}$ method using housekeeping *gyrB*

378   (RUMGNA_00867) gene as a reference gene and glucose as a reference condition. For

379   each gene in each condition, the final value of the relative level of transcription (expressed

380   as a fold change in gene transcription compared to glucose) is an average of 3 biological

381   replicates, 1-way Annova was used for statistical analysis, using Graph Pad Prism (V 5.03).

382

383   **Cloning, expression, mutagenesis and purification of recombinant proteins**

384   *R. gnavus* ATCC 29149 genomic DNA (gDNA) was purified from the cell pellet of a bacterial

385   overnight culture (1 ml) following centrifugation (5,000 g, 5 min) using the GeneJET

386   Genomic DNA Purification Kit (ThermoFisher, UK), according to the manufacturer's

387   instructions.

388   The full-length *Rg*SBP excluding the signal sequence (residues 1–29), the full length

389   *Rg*NanA and full length RUMGNA_02695 were amplified from *R. gnavus* ATCC 29149

390   gDNA, and cloned into the pEHISTEV[66] expression system, introducing a His-tag at the N

391   terminus using primers listed in Supplementary **Table S6**. DNA manipulation was carried out

392   in *E. coli* DH5α cells. Sequences were verified by DNA sequencing by Eurofins MWG

393   (Ebersberg, Germany) following plasmid preparation using the Monarch Plasmid Miniprep kit

394   (New England Biolabs). The *Rg*NanA active site mutant, K167A, was generated using the

395   QuikChange Lightning mutagenesis kit (Agilent) and primers listed in Supplementary **Table**

396   **S6**. *E. coli* BL21 (New England BioLabs) cells were transformed with the recombinant

397   plasmid harbouring the gene of interest according to manufacturer's instructions. Expression

398   was carried out in 800 ml 'Terrific Broth Base with Trace Elements' autoinduction media

399   (ForMedium, Dundee, UK) growing cells for 3 h at 37 °C and then at 16 °C for 48 h, with

400   shaking at 250 rpm. The cells were harvested by centrifugation at 10,000 *g* for 20 min. The

401   His-tagged proteins were purified by immobilized metal affinity chromatography (IMAC) and

402   further purified by gel filtration (Superdex 75 column) on an Akta system (GE Health Care

403 Life Sciences, Little Chalfont, UK). Protein purification was assessed by standard SDS–
404 polyacrylamide gel electrophoresis using NuPAGE Novex 4–12% Bis-Tris gels (Life
405 Technologies, Paisley, UK). Protein concentration was measured with NanoDrop 1000 UV-
406 Vis Spectrophotometer (Thermo Fischer Scientific, Wilmington, DE) and using the extinction
407 coefficient calculated by Protparam (ExPASy-Artimo, 2012) from the peptide sequence.

408

Fluorescence spectroscopy

410 All protein fluorescence experiments used a FluoroMax 3 fluorescence spectrometer with
411 connecting water bath at 37°C. Because of the presence of 15 tyrosine residues, the protein
412 was excited at 297 nm with slit widths of 5 nm. Under these conditions, the protein has a
413 maximal emission at 331 nm. *Rg*SBP was used at a concentration of 0.2 µM in 50 mM Tris
414 pH 7.5 for all fluorescence experiments. Cumulative fluorescence changes from titration of
415 the protein with ligand were plotted in GraphPad and fitted to a single rectangular hyperbola.
416 The K_d values reported were averaged from three separate ligand titration experiments.

417

Isothermal titration calorimetry (ITC)

419 Isothermal titration calorimetry (ITC) experiments were performed using the PEAQ-ITC
420 system (Malvern, Malvern, UK) with a cell volume of 200 µl. Prior to titration, protein samples
421 were exhaustively dialysed into 50 mM Tris-HCl pH 7.5. The ligand was dissolved in the
422 dialysis buffer. The cell protein concentration was 100 µM and the syringe ligand
423 concentration was 2 mM. Controls with titrant (sugar) injected into the buffer only were
424 subtracted from the data. The analysis was performed using the Malvern software, using a
425 single-binding site model. Experiments were carried out in triplicate.

426

Sialic acid aldolase activity assays

428 Aldolase activity was measured by monitoring the decrease in absorbance at 340 nm
429 (A_{340nm}) as NADH is converted to NAD by lactate dehydrogenase in a coupled reaction
430 where pyruvate is released from sialic acid by the aldolase. Reactions were performed in a
431 100 µl volume with final concentrations of 150 µM NADH (Sigma, St Louis, USA), 0.5 U LDH
432 (Sigma, St Louis, USA), 10 mM sialic acid (Neu5Ac or 2,7-anhydro-Neu5Ac) and 1.5 µg
433 purified *Rg*NanA or *Ec*NanA (*E. coli* aldolase CAS: 9027-60-5, Carbosynth, UK) in 50 mM
434 Na-phosphate buffer (pH 7.0). The reactions were performed at 37 °C and monitored using
435 FLUOstar OPTIMA (BMG LABTECH). For kinetics experiments, the sialic acid concentration
436 was varied at 20, 10, 5, 4, 2, 1, 0.4, 0.2, 0.1 mM and the initial rate of reaction determined for
437 each concentration in triplicate before analysis was performed by fitting the data to a
438 Michaelis-Menten using Graph Pad Prism (V 5.03).

439 To monitor the production of ManNAc during the aldolase-catalyzed reactions, 2-AB labelling
440 was carried out on the products from the above reactions. Briefly, 50 ng GlcNAc was added
441 to 10 µl of each sample as an internal reference, before drying using a Concentrator Plus
442 (Eppendorf). 5 µl of labelling reagent was added and incubated at 65 °C for 3 h. The
443 labelling reagent was prepared by dissolving 50 mg 2-aminobenzamide in a solution
444 containing 300 µl acetic acid and 700 µl DMSO, before 60 mg sodium cyanoborohydride is
445 added. Following addition of H_2O to reach 100 µl total volume, the sample was transferred to
446 a HPLC vial and 10 µl loaded onto a HyperClone 3u ODS (C18) 120A 150x4.6 mm 3 µ
447 column. Mobile phases of 0.25% n-butylamine, 0.5% phosphoric acid, 0.1%
448 Tetrahydrofurane; 50% methanol; Acetonitrile and H_2O were used at a 0.7 ml/min flow rate.

449

450 **Bioinformatics analyses**

451 **Sequence Similarity Networks (SSN)** The InterPro families for *Rg*NanH (Glycoside
452 Hydrolase, family 34; IPR001860) and *Rg*NanA (N-acetylneuraminate lyase; IPR005264)
453 were identified using the UniProt database, this family identifier was used to extract protein
454 sequences using Enzyme Function Initiative (EFI) Enzyme Similarity tool[67]. For the other
455 proteins, the families found in the InterPro database were too large to be analysed, so the
456 sequence BLAST tool was used with a maximum of 2500 protein sequences extracted. From
457 this sequence similarity networks were generated and viewed in Cytoscape version 3.6[68].

458 **Cluster analysis** Homologous gene clusters were identified for the *R. gnavus* ATCC 29149
459 *nan* cluster[45] using MultiGeneBlast[69]. The BCT (Bacteria) GenBank subdivision was queried
460 with the sequence spanning locus tags RUMGMA_RS11835 – RUMGNA_RS11885 (from
461 scaffold AAYG02000020_1). The data was manually curated, excluding all clusters that do
462 not contain a predicted sialidase or are homologous to the functionally characterized *S.*
463 *pneumoniae* NanC cluster[47,70] and the clusters are summarized by organism and predicted
464 gene content in Supplementary **Table S5**.

465

466 **RUMGNA_02695 enzymatic activity assay**

467 To assay RUMGNA_02695 activity against 2,7-anhydro-Neu5Ac, the purified recombinant
468 protein was incubated in 100 µl reactions at 37 °C overnight with 1 mM 2,7-anhydro-
469 Neu5Ac, 50 mM sodium phosphate buffer pH 7.0 and 500 µM NADH, NAD, FAD or no
470 cofactor. The reactions were dried using a Concentrator Plus (Eppendorf) for 1 h. Samples
471 were then resuspended in 50 µl of water and 50 µl of reaction buffer (1.74 mg of 1,2-
472 Diamino-4,5-methylenedioxybenzene dihydrochloride (DMB, Carbosynth, UK), 324.6 µl
473 MilliQ water, 88.6 µl glacial acetic acid, 58.2 µl of β-Mercaptoethanol and 79.3 µl of sodium
474 hydrosulphite) and incubated for 2 h at 55 °C in the dark. The samples were then centrifuged

475 for 1 min and filtered using a 0.45 μm filter into a glass HPLC vial and directly analysed by
476 HPLC.

477 DMB-labelled samples were analysed by injecting 10 μl onto a Luna 5 μm C-18(2) LC
478 column 250x4.6 mm (Phenomenex) at 1 ml/min. Mobile phases methanol/acetonitrile/water
479 were used for separation of fluorescently labelled sialic acids[46]. The settings of the
480 fluorescence detector were 373 nm excitation and 448 nm emission. Samples were run
481 alongside a Neu5Ac standard.

482 To determine the kinetic parameters of RUMGNA_02695 enzymatic reaction, a coupled
483 reaction with lactate dehydrogenase and sialic acid aldolase was carried out as described
484 above but with 15 μg of *Rg*NanA and 10 μg RUMGNA_02695 in each reaction. For the
485 kinetics assays, 1, 0.4, 0.2, 0.1, 0.04, 0.02 and 0.01 mM 2,7-anhydro-Neu5Ac was used and
486 the initial rate of reaction determined for each concentration in triplicate before analysis was
487 performed by fitting the data to a Michaelis-Menten using Graph Pad Prism (V 5.03).

488 Electrospray ionisation spray mass spectrometry (ESI-MS) analysis was performed using the
489 Applied Biosystems 4000 Q-TRAP. The full 100 μl reaction was diluted with 500 ul of 50%
490 Acetonitrile and 0.1 % formic acid and samples analysed in negative ion mode using direct
491 injection.

492

493 **ClosTron mutagenesis**

494 *R. gnavus* mutants were generated using the ClosTron methodology[56], which inserts an
495 erythromycin resistance cassette into the gene of interest. The target site (270a) was
496 identified using the Perutka method[71]. The re-targeted introns were synthesised and ligated
497 into the pMTL007C-E2 vector by ATUM (MenloPark, USA). The plasmids were then
498 transformed into *E. coli* CA434 using the heat-shock 42°C for 45 seconds followed by 2 min
499 on ice before the recombinant clones were selected for chloramphenicol resistance (25
500 μg/ml). Recombinant *E. coli* cells were grown overnight aerobically in 10 ml LB, 1 ml of the
501 overnight culture was pelleted and washed with PBS. Continuing under anaerobic conditions
502 the *E. coli* cell pellet was resuspended in 200 μl of an *R. gnavus* overnight culture and the
503 cell suspension spotted onto a non-selective BHI-YH plate. Following incubation for 8 h at 37
504 °C the bacteria were washed from the plate using PBS and plated onto BHI-YH
505 supplemented with cycloserine (250 μg/ml) and thiamphenicol (15 μg/ml) and grown for 72 h
506 to select against *E. coli* and for transfer of the plasmid to *R. gnavus.* Individual colonies were
507 grown in non-selective BHI-YH broth overnight to allow expression of the plasmid and
508 genomic recombination. The culture was then plated onto a BHI-YH medium containing
509 cycloserine (250 μg/ml) and erythromycin (10 μg/ml) to select clones with successful

510 genomic recombination. PCR and sequencing were used to confirm recombination in the
511 gene of interest.

512 Expression of the *nan* cluster genes in the generated mutants was assessed as described
513 above using RNA samples from growth on YCFA supplemented with glucose.

514 The ability of the mutants to utilise sialic acids and sialoconjugates was assessed by
515 supplementing YCFA with 11.1 mM of 2,7-anhydro-Neu5Ac, 3'SL, glucose or Neu5Ac in
516 triplicate 200 µl cultures in 96-well microtiter plates. The $OD_{595\ nm}$ was measured hourly for
517 10 h in an infinite F50 plate reader (Tecan, UK) housed within an anaerobic cabinet
518 connected to Magellan V7.0 software.

519

520 **Saturation Transfer Difference (STD) NMR Spectroscopy.**

521 An amicon centrifuge filter unit with a 10 kDa MW cut-off was used to exchange the protein
522 in 25 mM d_{19}-2,2-bis(hydroxymethyl)-2,2′,2″-nitrilotriethanol pH* 7.4 (uncorrected for the
523 deuterium isotope effect on the pH glass electrode) D_2O buffer and 50 mM NaCl. 2,7-
524 anhydro-Neu5Ac and Neu5Ac were dissolved in 25 mM d19-2,2-bis(hydroxymethyl)-2,2′,2″-
525 nitrilotriethanol pH 7.4, 50 mM NaCl. Characterization of ligand binding by Saturation
526 Transfer Difference NMR Spectroscopy[72] was performed on a Bruker Avance 800.23 MHz at
527 298 K. The on- and off-resonance spectra were acquired using a train of 50 ms Gaussian
528 selective saturation pulses using a variable saturation time from 0.5 s to 4 s, for binding
529 epitope mapping determination while only 0.5 s of saturation time for each selected
530 frequency was used to perform the DEEP-STD NMR experiments[42]. The water signal was
531 suppressed by using the excitation sculpting technique[73], while the remaining protein
532 resonances were filtered using a T_2 filter of 40 ms. All the spectra were performed with a
533 spectral width of 10 KHz and 32768 data points using 256 or 512 scans. This time due to the
534 absence of a 3D structure it was impossible to derive the resonances for saturation of
535 aliphatic and aromatic residues found in the binding site as required by the DEEP-STD NMR
536 technique. Moreover, *Rg*SBP being a high molecular weight protein the NMR spectra
537 assignment is precluded. For this we adopted a search for druggable sites strategy using 4-
538 hydroxy-1-oxyl-2,2,6,6-tetramethylpiperidine (TEMPOL) as previously described[74].

539 1H-1H TOCSY spectra of the protein (500 µM) were acquired in the presence and in the
540 absence of TEMPOL (2.5 mM and 12.5 mM). The spectra were performed with a spectral
541 width of 10 kHz using a time domain of 2056 data points in the direct dimension and 32
542 scans. The indirect dimension was acquired using the non-uniform sampling (NUS)
543 technique acquiring a NUS amount of 50% of the original 256 increments resulting in 64
544 hypercomplex points. The spectra were processed with the Topspin 3.1 compressed sensing
545 (cs) routine. The final selected resonances were those identified by the TEMPOL PRE effect,

546   and not overlapping with ligand signals. The DEEP-STD NMR data obtained were used to

547   derive the average orientation of the ligand bound to *Rg*SBP by averaging the DEEP-STD

548   factors obtained from each saturated region. The DEEP-STD NMR and binding epitope

549   mapping analysis were performed using previously published procedures[42,74,75].

550

Crystal structure determination

552   Sitting drop vapour diffusion crystallisation experiments of *Rg*NanA wt were set up at a

553   concentration of 20 mg/ml and monitored using the VMXi beamline at Diamond Light

554   Source[76]. The described *Rg*NanA wild-type crystal structure was acquired from a crystal

555   grown in the Morpheus screen (Molecular Dimensions), 0.2 M 1,6-hexandiol, 0.2 M 1-

556   butanol, 0.2 M 1,2-propanediol, 0.2 M 2-propanol, 0.2 M 1,4-butanediol, 0.2 M 1,3-

557   propanediol, 0.1 M Hepes/MOPS pH 6.5, 20% ethylene glycol, 10% PEG 8000. The

558   diffraction experiment was performed on beamline I24 beamline at Diamond Light Source

559   Ltd at 100K using a wavelength of 0.9686 Å. The data were processed with Xia2 making use

560   of aimless, dials, and pointless (for data collection and refinement statistics see

561   Supplementary **Table S3**). The structure was phased using MrBump through CCP4 online

562   and Molrep[77-79], by CdNaI from *C. difficile* (PDB 4woq) prepared using Chainsaw.

563   Refinement was carried out using Refmac, Buster, and PDB redo[80-84]. Coot and ArpWarp

564   were used for model building and Molprobity for structure validation[85]. It was not possible to

565   crystallise *Rg*NanA wt in the presence of Neu5Ac as it caused protein precipitation and

566   *Rg*NanA crystals dissolved in Neu5Ac soaking experiments, as also observed previously

567   with *P. multocida* Neu5Ac aldolase[53], possibly due to conformational changes during

568   substrate binding or catalysis. Experiments with *R*gNanA K167A mutant were set up at 25

569   mg/ml. Diffracting crystals grew in 0.1 M Tris/BICINE pH 8.5, 20% ethylene glycol, 100 mM

570   MgCl$_2$, 10% PEG 8000 and diffraction experiments performed on beamline I04 at Diamond

571   Light Source using a wavelength of 0.9795 Å. The crystal structure was phased with

572   PHASER using the *Rg*NanA wild-type crystal structure[86]. A 60 second 5 mM Neu5Ac soak

573   prior to freezing generated the *Rg*NanA K167A Neu5Ac complex. Due to data anisotropy, we

574   processed the data in autoPROC[86,87] with the STARANISO option[88] and used these data for

575   refinement with Buster using the previously obtained models of *Rg*NanA wild-type and

576   K167A Neu5Ac complex.

577

***In vivo* colonisation and analyses**

579   The impact of the *nan* deletion mutation on *R. gnavus* fitness was assessed by its ability to

580   colonise germ-free C57BL/6J mice. Groups containing four 7-9 week old germ-free mice

581   (two male, two female) were gavaged with 1x10^8 CFU of *R. gnavus* ATCC 29149 wild-type

582   or antisense *nan* mutant in 100 µl PBS, individually or in combination. Sample size was
583   selected following the 3 R's principles of reduction, replacement, refinement, whilst ensuring
584   data collected allowed for statistical analysis, randomization was not possible due to the
585   constraints of germ-free isolators, scientists were blinded for the FISH analysis. Care and
586   treatment of animals was in accordance guidelines from and approval by the University of
587   East Anglia Disease Modelling Unit and all animal experiments were conducted in strict
588   accordance with the Home Office Animals (Scientific Procedures) Act 1986. Faecal samples
589   were collected from each mouse at 3,7 and 14 days post gavage, and caecal content taken
590   at day 14. DNA was extracted from these samples using the MP Biomedicals Fast
591   DNA™ SPIN kit for Soil DNA extraction with the following modifications. The samples were
592   resuspended in 978 µl of sodium phosphate buffer before being incubated at 4 °C for one
593   hour following addition of 122 µl MT Buffer. The samples were then transferred to the lysing
594   tubes and homogenised in a FastPrep® Instrument (MP Biomedicals) 3 times for 40 s at a
595   speed setting of 6.0 with 5 min on ice between each bead-beating step. The protocol was
596   then followed as recommended by the supplier.

597   Colonisation was quantified using qPCR carried out in an Applied Biosystems 7500 Real-
598   Time PCR system (Life Technologies Ltd). In competition experiments, primers based on the
599   insertion in the *Rg*NanH gene were used to distinguish between wild-type and *nan* mutant.
600   One pair of primers was designed to specifically target *R. gnavus* wild-type strain by
601   spanning the area of insertion into the *nan* cluster and one pair of primers was designed to
602   specifically amplify the inserted DNA, therefore targeting the *nan* mutant (Supplementary
603   **Table S6**). The primers were between 18 and 23 nt-long, with a T_m of 59–60°C. Standard
604   curves were prepared in triplicate for both primer pairs using a 10-fold serial dilution of DNA
605   corresponding to 1×10^7 copies of *Rg*NanH/2ul to 1×10^2 copies/2ul diluted in 5 µg/ml Herring
606   sperm DNA. The standard curves showed a linear relationship of log input DNA vs. the
607   threshold cycle (C_T), with acceptable values for the slopes and the regression coefficients
608   (R^2). The dissociation curves were also performed to check the specificity of the amplicons.
609   Each qPCR reaction (10 µl) was then carried out in triplicate with 2 µl of 1 ng/µl DNA (diluted
610   in 5 µg/ml Herring sperm DNA) and 0.2 µM of each primer, using the QuantiFast SYBR
611   Green PCR kit (Qiagen) according to the manufacturer's instructions (except that the
612   combined annealing/extension step was extended to 35 s instead of 30 s). Data obtained
613   were analysed using the prepared standard curves.

614

615   **RNAseq analysis**

616   For RNAseq analysis, the colonic tissues from mono-colonised mice were gently washed
617   and stored in RNAlater at -80ºC until extraction. RNA extraction was performed using the

618 RNeasy mini kit (QIAGEN) following the manufacturer's instructions for purification of total
619 RNA from animal tissues, including the on-column DNase digestion. Homogenisation was
620 achieved with acid washed glass beads using the FastPrep®-24 (MP Biomedicals, Solon,
621 USA) by 3 intermittent runs of 30 s at 6 m/s speed every 5 min, at room temperature. Elution
622 was performed as recommended with 50 µl RNAse-free water. The quality and
623 concentration of the RNA samples was assessed using NanoDrop 2000 Spectrophotometer
624 Nanodrop, the Qubit RNA HS assay on Qubit® 2.0 fluorometer (Life Technologies) and
625 Agilent RNA 600 Nano kit on Agilent 2100 Bioanalyzer (Agilent Technologies, Stockport,
626 UK).

627 RNAseq was carried out by Novogene (HK) (Hong Kong). Briefly, mRNA was enriched using
628 oligo(dT) beads, fragmented randomly in fragmentation buffer, followed by cDNA synthesis
629 using random hexamers and reverse transcriptase. After first-strand synthesis, a custom
630 second-strand synthesis buffer (Illumina) was added with dNTPs, RNase H and *Escherichia*
631 *coli* polymerase I to generate the second strand by nick-translation. The final cDNA library
632 was obtained after a round of purification, terminal repair, A-tailing, ligation of sequencing
633 adapters, size selection and PCR enrichment. Library concentration was first quantified
634 using a Qubit® 2.0 fluorometer (Life Technologies), and then diluted to 1 ng/µl before
635 checking insert size on an Agilent 2100 and quantifying to greater accuracy by qPCR (library
636 activity >2 nM). Sequencing of the library was carried out on Illumina Hiseq platform and
637 125/150 bp paired-end reads were generated.

638 FASTQ files containing base calls and quality information for all reads that passed quality
639 filtering were generated. Reads were mapped to the mouse reference genome using
640 TopHat2[89]. The mismatch parameter was set to two, and other parameters were set to
641 default. Appropriate parameters were also set, such as the longest intron length. Filtered
642 reads were used to analyze the mapping status of RNA-seq data to the reference genome.
643 The HTSeq software was used to analyze the gene expression levels, using the union
644 mode[90]. In order for the gene expression levels estimated from different genes and
645 experiments to be comparable, the FPKM (Fragments Per Kilobase of transcript sequence
646 per Millions base pairs sequenced) was used to take into account the effects of both
647 sequencing depth and gene length. The differential gene expression analysis was carried
648 out using the DESeq package[90] and the readcounts from gene expression level analysis as
649 input data. An adjusted p value (padj) cut-off of 0.05 was used to determine differential
650 expressed transcripts.

651

652 **Fluorescent in situ hybridization (FISH) staining**

653 For FISH analysis, the colonic tissue was fixed in methacarn (60% dry methanol, 30%
654 chloroform and 10% acetic acid), processed and embedded in paraffin as previously
655 described[2]. Tissue sections were prepared at 8-10 μm. Paraffin sections were dewaxed and
656 washed in 95% ethanol. The tissue sections were incubated with 100 μl of Alexa Fluor 555-
657 conjugated Erec482 probe (5' – GCTTCTTAGTCARGTACCG -3') at a concentration of 10
658 ng/μl, in hybridisation buffer (20 mM Tris-HCl, pH 7.4, 0.9M NaCl, 0.1% SDS) at 50°C
659 overnight. The sections were then incubated in a 50°C prewarmed wash buffer (20m M Tris-
660 HCl, pH 7.4, 0.9 M NaCl) for 20 min. All subsequent steps were performed at 4°C. The
661 sections were washed with PBS, the blocked with TNB buffer (0.5% w/v blocking reagent in
662 100 mM Tris-HCl, pH 7.5, 150 mM NaCl) supplemented with 5% goat serum. To detect
663 mucin, the sections were then counterstained with a Muc2 antibody (sc-15334) at 1:100
664 dilution in TNB buffer overnight. The sections were washed in PBS, then goat anti-rabbit
665 antibodies (diluted 1:500) were used for immunodetection. The sections were counterstained
666 with Sytox blue (S11348, ThermoFisher) diluted 1:1000 in PBS and mounted in Prolong gold
667 anti-fade mounting medium. The slides were imaged using a Leica TCS SP2 confocal
668 microscope with a x63 objective. The distance between the leading front of bacteria and the
669 base of the mucus layer was measured with FIJI[91]. A total of 70 images from 8 mice were
670 analysed, and scientists were blinded for the analysis due the subjectivity of determining the
671 leading front of bacteria and base of the mucus layer. The association between genotype
672 and distances was estimated by a linear mixed model, including fixed effects of genotype
673 and area and random effects of mouse and each individual image. There was substantial
674 spatial correlation between adjacent observations and so an AR(1) correlation structure was
675 added. The resulting model had no residual autocorrelation as judged by visual inspection of
676 autocorrelation function. The nmle package version 3.1-137 using R version 3.5.3 was used
677 to estimate the model.
678

679 **Data Availability**
680 Genome and protein sequences are available from NCBI and referenced within the text or
681 supplementary information. Accession numbers of all genomes used for multigene
682 alignments, are available in Supplementary Table 5. Raw FASTQ files for the RNA-seq
683 libraries were deposited to the NCBI Sequence Read Archive (SRA), and have been
684 assigned BioProject accession PRJNA559470. The crystal structures described in this
685 paper have been deposited in the protein data bank (ODB) with the following identifiers
686 6RAB (WT), 6RB7 (K167A), and 6RD1 (K167A Neu5Ac complex). All other data are
687 available upon request from the corresponding author.

688 Computer code for statistical analysis is available on request from the corresponding
689 author.
690

705

706 **Contributions**
707 NJ conceived the study and wrote the manuscript with contribution from co-authors. AB
708 carried out bioinformatics analyses, transcriptomics, heterologous expression, site-directed
709 mutagenesis, enzymatic assays, analytical product characterisation (HPLC, MS), protein-
710 ligand interaction experiments (ITC and fluorescence spectroscopy) and *R. gnavus*
711 mutagenesis. JB supervised the Clostron mutagenesis. GHT supervised the fluorescence
712 spectroscopy experiments. DL developed the HPLC and MS analysis protocols. CDO carried
713 out the X-ray crystallography under MAW's supervision. LV carried out the immuno-
714 histochemistry experiments, EC contributed to the mouse study and RNASeq analyses. AB,
715 EC, LV, DL worked under NJ's supervision. RN carried out the NMR experiments under JA's
716 supervision, AX and WL synthesized the 2,7-anydro-Neu5Ac used in this study under XC's
717 supervision. JC carried out the cluster bioinformatics analyses. All authors reviewed and
718 corrected the final manuscript.
719

720 **Corresponding authors**
721 Correspondence to Nathalie Juge.
722

723 **Competing interests**
724 The authors declare no competing interests.

References

725

726

727 1   Donaldson, G. P., Lee, S. M. & Mazmanian, S. K. Gut biogeography of the bacterial
728   microbiota. *Nature Reviews Microbiology* **14**, 20-32, doi:10.1038/nrmicro3552
729   (2016).

730 2   Johansson, M. E. V., Larsson, J. M. H. & Hansson, G. C. The two mucus layers of
731   colon are organized by the MUC2 mucin, whereas the outer layer is a legislator of
732   host-microbial interactions. *Proceedings of the National Academy of Sciences of the*
733   *United States of America* **108**, 4659-4665, doi:10.1073/pnas.1006451107 (2011).

734 3   Etienne-Mesmin, L. *et al.* Experimental models to study intestinal microbes-mucus
735   interactions in health and disease. *FEMS Microbiol. Rev.*, doi:10.1093/femsre/fuz013
736   (2019).

737 4   Jensen, P. H., Kolarich, D. & Packer, N. H. Mucin-type O-glycosylation--putting the
738   pieces together. *FEBS J.* **277**, 81-94, doi:10.1111/j.1742-4658.2009.07429.x (2010).

739 5   Ndeh, D. & Gilbert, H. J. Biochemistry of complex glycan depolymerisation by the
740   human gut microbiota. *FEMS Microbiol. Rev.* **42**, 146-164,
741   doi:10.1093/femsre/fuy002 (2018).

742 6   Tailford, L. E., Crost, E. H., Kavanaugh, D. & Juge, N. Mucin glycan foraging in the
743   human gut microbiome. *Frontiers in genetics* **6**, 81-81, doi:10.3389/fgene.2015.00081
744   (2015).

745 7   Martens, E. C., Chiang, H. C. & Gordon, J. I. Mucosal Glycan Foraging Enhances
746   Fitness and Transmission of a Saccharolytic Human Gut Bacterial Symbiont. *Cell*
747   *Host & Microbe* **4**, 447-457, doi:10.1016/j.chom.2008.09.007 (2008).

748 8   Robbe, C., Capon, C., Coddeville, B. & Michalski, J. C. Structural diversity and
749   specific distribution of 0-glycans in normal human mucins along the intestinal tract.
750   *Biochem. J.* **384**, 307-316, doi:10.1042/bj20040605 (2004).

751 9   Robbe, C. *et al.* Evidence of regio-specific glycosylation in human intestinal mucins -
752   Presence of an acidic gradient along the intestinal tract. *J. Biol. Chem.* **278**, 46337-
753   46348, doi:10.1074/jbc.M302529200 (2003).

754 10   Juge, N., Tailford, L. & Owen, C. D. Sialidases from gut bacteria: a mini-review.
755   *Biochem. Soc. Trans.* **44**, 166-175, doi:10.1042/bst20150226 (2016).

756 11   Lewis, A. L. & Lewis, W. G. Host sialoglycans and bacterial sialidases: a mucosal
757   perspective. *Cell. Microbiol.* **14**, 1174-1182, doi:10.1111/j.1462-5822.2012.01807.x
758   (2012).

759 12   Almagro-Moreno, S. & Boyd, E. F. Insights into the evolution of sialic acid
760   catabolism among bacteria. *BMC Evol. Biol.* **9**, 118, doi:10.1186/1471-2148-9-118
761   (2009).

762 13   Plumbridge, J. & Vimr, E. Convergent pathways for utilization of the amino sugars
763   N-acetylglucosamine, N-acetylmannosamine, and N-acetylneuraminic acid by
764   Escherichia coli. *J. Bacteriol.* **181**, 47-54 (1999).

765 14   Martinez, J., Steenbergen, S. & Vimr, E. Derived structure of the putative sialic acid
766   transporter from Escherichia coli predicts a novel sugar permease domain. *J.*
767   *Bacteriol.* **177**, 6005-6010 (1995).

768 15   Brigham, C. *et al.* Sialic Acid (N-Acetyl Neuraminic Acid) Utilization by Bacteroides
769   fragilis Requires a Novel N-Acetyl Mannosamine Epimerase. *J. Bacteriol.* **191**, 3629-
770   3638, doi:10.1128/jb.00811-08 (2009).

771 16   Vimr, E. R., Kalivoda, K. A., Deszo, E. L. & Steenbergen, S. M. Diversity of
772   microbial sialic acid metabolism. *Microbiol. Mol. Biol. Rev.* **68**, 132-153,
773   doi:10.1128/mmbr.68.1.132-153.2004 (2004).

774   17   Thomas, G. H. Sialic acid acquisition in bacteria - one substrate, many transporters.
775   *Biochem. Soc. Trans.* **44**, 760-765, doi:10.1042/bst20160056 (2016).

776   18   Mulligan, C. *et al.* The substrate-binding protein imposes directionality on an
777   electrochemical sodium gradient-driven TRAP transporter. *Proceedings of the*
778   *National Academy of Sciences* **106**, 1778, doi:10.1073/pnas.0809979106 (2009).

779   19   Wahlgren, W. Y. *et al.* Substrate-bound outward-open structure of a Na(+)-coupled
780   sialic acid symporter reveals a new Na(+) site. *Nat Commun* **9**, 1753,
781   doi:10.1038/s41467-018-04045-7 (2018).

782   20   Severi, E., Hosie, A. H., Hawkhead, J. A. & Thomas, G. H. Characterization of a
783   novel sialic acid transporter of the sodium solute symporter (SSS) family and in vivo
784   comparison with known bacterial sialic acid transporters. *FEMS Microbiol. Lett.* **304**,
785   47-54, doi:10.1111/j.1574-6968.2009.01881.x (2010).

786   21   Gangi Setty, T., Cho, C., Govindappa, S., Apicella, M. A. & Ramaswamy, S.
787   Bacterial periplasmic sialic acid-binding proteins exhibit a conserved binding site.
788   *Acta crystallographica. Section D, Biological crystallography* **70**, 1801-1811,
789   doi:10.1107/s139900471400830x (2014).

790   22   Mulligan, C., Leech, A. P., Kelly, D. J. & Thomas, G. H. The membrane proteins
791   SiaQ and SiaM form an essential stoichiometric complex in the sialic acid tripartite
792   ATP-independent periplasmic (TRAP) transporter SiaPQM (VC1777-1779) from
793   Vibrio cholerae. *J. Biol. Chem.* **287**, 3598-3608, doi:10.1074/jbc.M111.281030
794   (2012).

795   23   Muller, A. *et al.* Conservation of structure and mechanism in primary and secondary
796   transporters exemplified by SiaP, a sialic acid binding virulence factor from
797   Haemophilus influenzae. *J. Biol. Chem.* **281**, 22212-22222,
798   doi:10.1074/jbc.M603463200 (2006).

799   24   Severi, E. *et al.* Sialic acid transport in Haemophilus influenzae is essential for
800   lipopolysaccharide sialylation and serum resistance and is dependent on a novel
801   tripartite ATP-independent periplasmic transporter. *Mol. Microbiol.* **58**, 1173-1185,
802   doi:10.1111/j.1365-2958.2005.04901.x (2005).

803   25   Post, D. M., Mungur, R., Gibson, B. W. & Munson, R. S., Jr. Identification of a novel
804   sialic acid transporter in Haemophilus ducreyi. *Infect. Immun.* **73**, 6727-6735,
805   doi:10.1128/IAI.73.10.6727-6735.2005 (2005).

806   26   North, R. A. *et al.* The Sodium Sialic Acid Symporter From Staphylococcus aureus
807   Has Altered Substrate Specificity. *Frontiers in Chemistry* **6**, 233 (2018).

808   27   Hopkins, A. P., Hawkhead, J. A. & Thomas, G. H. Transport and catabolism of the
809   sialic acids N-glycolylneuraminic acid and 3-keto-3-deoxy-D-glycero-D-
810   galactononomic acid by Escherichia coli K-12. *FEMS Microbiol. Lett.* **347**, 14-22,
811   doi:10.1111/1574-6968.12213 (2013).

812   28   Sagheddu, V., Patrone, V., Miragoli, F., Puglisi, E. & Morelli, L. Infant Early Gut
813   Colonization by Lachnospiraceae: High Frequency of Ruminococcus gnavus.
814   *Frontiers in pediatrics* **4**, 57, doi:10.3389/fped.2016.00057 (2016).

815   29   Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic
816   sequencing. *Nature* **464**, 59-65, doi:10.1038/nature08821 (2010).

817   30   Ludwig, W., Schleifer, K.-H. & Whitman, W. B. in *Bergey's Manual® of Systematic*
818   *Bacteriology: Volume Three The Firmicutes* (eds Paul De Vos *et al.*) 1-13 (Springer
819   New York, 2009).

820   31   Kraal, L., Abubucker, S., Kota, K., Fischbach, M. A. & Mitreva, M. The Prevalence
821   of Species and Strains in the Human Microbiome: A Resource for Experimental
822   Efforts. *PLOS ONE* **9**, e97279, doi:10.1371/journal.pone.0097279 (2014).

823 32 Olbjorn, C. *et al.* Fecal microbiota profiles in treatment-naive pediatric inflammatory
824   bowel disease - associations with disease phenotype, treatment, and outcome. *Clinical*
825   *and experimental gastroenterology* **12**, 37-49, doi:10.2147/ceg.S186235 (2019).
826 33 Sokol, H. *et al.* Specificities of the intestinal microbiota in patients with inflammatory
827   bowel disease and Clostridium difficile infection. *Gut Microbes* **9**, 55-60,
828   doi:10.1080/19490976.2017.1361092 (2018).
829 34 Nishino, K. *et al.* Analysis of endoscopic brush samples identified mucosa-associated
830   dysbiosis in inflammatory bowel disease. *Journal of gastroenterology* **53**, 95-106,
831   doi:10.1007/s00535-017-1384-4 (2018).
832 35 Machiels, K. *et al.* Specific members of the predominant gut microbiota predict
833   pouchitis following colectomy and IPAA in UC. *Gut* **66**, 79-88, doi:10.1136/gutjnl-
834   2015-309398 (2017).
835 36 Hall, A. B. *et al.* A novel Ruminococcus gnavus clade enriched in inflammatory
836   bowel disease patients. *Genome Medicine* **9**, 103, doi:10.1186/s13073-017-0490-5
837   (2017).
838 37 Fuentes, S. *et al.* Microbial shifts and signatures of long-term remission in ulcerative
839   colitis after faecal microbiota transplantation. *The ISME journal* **11**, 1877-1889,
840   doi:10.1038/ismej.2017.44 (2017).
841 38 Joossens, M. *et al.* Dysbiosis of the faecal microbiota in patients with Crohn's disease
842   and their unaffected relatives. *Gut* **60**, 631-637, doi:10.1136/gut.2010.223263 (2011).
843 39 Willing, B. P. *et al.* A pyrosequencing study in twins shows that gastrointestinal
844   microbial profiles vary with inflammatory bowel disease phenotypes.
845   *Gastroenterology* **139**, 1844-1854.e1841, doi:10.1053/j.gastro.2010.08.049 (2010).
846 40 Png, C. W. *et al.* Mucolytic bacteria with increased prevalence in IBD mucosa
847   augment in vitro utilization of mucin by other bacteria. *The American journal of*
848   *gastroenterology* **105**, 2420-2428, doi:10.1038/ajg.2010.281 (2010).
849 41 Owen, C. D. *et al.* Unravelling the specificity and mechanism of sialic acid
850   recognition by the gut symbiont Ruminococcus gnavus. *Nature Communications* **8**,
851   2196, doi:10.1038/s41467-017-02109-8 (2017).
852 42 Monaco, S., Tailford, L. E., Juge, N. & Angulo, J. Differential Epitope Mapping by
853   STD NMR Spectroscopy To Reveal the Nature of Protein-Ligand Contacts.
854   *Angewandte Chemie (International ed. in English)* **56**, 15289-15293,
855   doi:10.1002/anie.201707682 (2017).
856 43 Crost, E. H. *et al.* The mucin-degradation strategy of Ruminococcus gnavus: The
857   importance of intramolecular trans-sialidases. *Gut Microbes* **7**, 302-312,
858   doi:10.1080/19490976.2016.1186334 (2016).
859 44 Tailford, L. E. *et al.* Discovery of intramolecular trans-sialidases in human gut
860   microbiota suggests novel mechanisms of mucosal adaptation. *Nat Commun* **6**, 7624,
861   doi:10.1038/ncomms8624 (2015).
862 45 Crost, E. H. *et al.* Utilisation of mucin glycans by the human gut symbiont
863   Ruminococcus gnavus is strain-dependent. *PLoS One* **8**, e76341,
864   doi:10.1371/journal.pone.0076341 (2013).
865 46 Monestier, M. *et al.* Membrane-enclosed multienzyme (MEME) synthesis of 2,7-
866   anhydro-sialic acid derivatives. *Carbohydr. Res.* **451**, 110-117,
867   doi:https://doi.org/10.1016/j.carres.2017.08.008 (2017).
868 47 Xu, G. *et al.* Three Streptococcus pneumoniae sialidases: three different products. *J.*
869   *Am. Chem. Soc.* **133**, 1718-1721, doi:10.1021/ja110733q (2011).
870 48 Xu, G. *et al.* Crystal structure of the NanB sialidase from Streptococcus pneumoniae.
871   *J. Mol. Biol.* **384**, 436-449, doi:10.1016/j.jmb.2008.09.032 (2008).

872   49   Kumar, J. P., Rao, H., Nayak, V. & Ramaswamy, S. Crystal structures and kinetics of
873   N-acetylneuraminate lyase from Fusobacterium nucleatum. *Acta Crystallographica
874   Section F* **74**, 725-732, doi:10.1107/S2053230X18012992 (2018).
875   50   Campeotto, I. *et al.* Pathological macromolecular crystallographic data affected by
876   twinning, partial-disorder and exhibiting multiple lattices for testing of data
877   processing and refinement tools. *Scientific Reports* **8**, 14876, doi:10.1038/s41598-
878   018-32962-6 (2018).
879   51   North, R. A. *et al.* Structure and inhibition of N-acetylneuraminate lyase from
880   methicillin-resistant Staphylococcus aureus. *FEBS Lett.* **590**, 4414-4428,
881   doi:10.1002/1873-3468.12462 (2016).
882   52   Timms, N. *et al.* Structural insights into the recovery of aldolase activity in N-
883   acetylneuraminic acid lyase by replacement of the catalytically active lysine with
884   gamma-thialysine by using a chemical mutagenesis strategy. *ChemBioChem* **14**, 474-
885   481, doi:10.1002/cbic.201200714 (2013).
886   53   Huynh, N. *et al.* Structural Basis for Substrate Specificity and Mechanism of N-
887   Acetyl-D-neuraminic Acid Lyase from Pasteurella multocida. *Biochemistry* **52**, 8570-
888   8579, doi:10.1021/bi4011754 (2013).
889   54   Barbosa, J. A. *et al.* Active site modulation in the N-acetylneuraminate lyase sub-
890   family as revealed by the structure of the inhibitor-complexed Haemophilus
891   influenzae enzyme. *J. Mol. Biol.* **303**, 405-421, doi:10.1006/jmbi.2000.4138 (2000).
892   55   Daniels, A. D. *et al.* Reaction mechanism of N-acetylneuraminic acid lyase revealed
893   by a combination of crystallography, QM/MM simulation, and mutagenesis. *ACS
894   Chem Biol* **9**, 1025-1032, doi:10.1021/cb500067z (2014).
895   56   Heap, J. T. *et al.* The ClosTron: Mutagenesis in Clostridium refined and streamlined.
896   *J. Microbiol. Methods* **80**, 49-55, doi:10.1016/j.mimet.2009.10.018 (2010).
897   57   Mandal, C., Schwartz-Albiez, R. & Vlasak, R. in *Sialoglyco Chemistry and Biology I:
898   Biosynthesis, Structural Diversity and Sialoglycopathologies* Vol. 366 *Topics in
899   Current Chemistry* (eds R. GerardySchahn, P. Delannoy, & M. VonItzstein) 1-30
900   (Springer Int Publishing Ag, 2015).
901   58   Vimr, E. R. Unified theory of bacterial sialometabolism: how and why bacteria
902   metabolize host sialic acids. *ISRN Microbiol* **2013**, 816713, doi:10.1155/2013/816713
903   (2013).
904   59   Robbe-Masselot, C., Maes, E., Rousset, M., Michalski, J. C. & Capon, C.
905   Glycosylation of human fetal mucins: a similar repertoire of O-glycans along the
906   intestinal tract. *Glycoconjugate journal* **26**, 397-413, doi:10.1007/s10719-008-9186-9
907   (2009).
908   60   Pezzicoli, A., Ruggiero, P., Amerighi, F., Telford, J. L. & Soriani, M. Exogenous
909   Sialic Acid Transport Contributes to Group B Streptococcus Infection of Mucosal
910   Surfaces. *J. Infect. Dis.* **206**, 924-931, doi:10.1093/infdis/jis451 (2012).
911   61   Bidossi, A. *et al.* A Functional Genomics Approach to Establish the Complement of
912   Carbohydrate Transporters in Streptococcus pneumoniae. *Plos One* **7**,
913   doi:10.1371/journal.pone.0033320 (2012).
914   62   Marion, C., Burnaugh, A. M., Woodiga, S. A. & King, S. J. Sialic Acid Transport
915   Contributes to Pneumococcal Colonization. *Infect. Immun.* **79**, 1262-1269,
916   doi:10.1128/iai.00832-10 (2011).
917   63   Gangi Setty, T. *et al.* Molecular characterization of the interaction of sialic acid with
918   the periplasmic binding protein from Haemophilus ducreyi. *J. Biol. Chem.* **293**,
919   20073-20084, doi:10.1074/jbc.RA118.005151 (2018).
920
921   Methods References

922

64   Xiao, A. *et al.* Streptococcus pneumoniae Sialidase SpNanB-Catalyzed One-Pot
        Multienzyme (OPME) Synthesis of 2,7-Anhydro-Sialic Acids as Selective Sialidase
        Inhibitors. *J. Org. Chem.*, doi:10.1021/acs.joc.8b01519 (2018).

65   Duncan, S. H., Hold, G. L., Harmsen, H. J., Stewart, C. S. & Flint, H. J. Growth
        requirements and fermentation products of Fusobacterium prausnitzii, and a proposal
        to reclassify it as Faecalibacterium prausnitzii gen. nov., comb. nov. *Int. J. Syst. Evol.
        Microbiol.* **52**, 2141-2146, doi:10.1099/00207713-52-6-2141 (2002).

66   Liu, H. & Naismith, J. H. A simple and efficient expression and purification system
        using two newly constructed vectors. *Protein Expr Purif* **63**, 102-111,
        doi:10.1016/j.pep.2008.09.008 (2009).

67   Gerlt, J. A. *et al.* Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A
        web tool for generating protein sequence similarity networks. *Biochim. Biophys. Acta*
        1854, 1019-1037, doi:10.1016/j.bbapap.2015.04.015 (2015).

68   Shannon, P. *et al.* Cytoscape: a software environment for integrated models of
        biomolecular interaction networks. *Genome Res.* **13**, 2498-2504,
        doi:10.1101/gr.1239303 (2003).

69   Medema, M. H., Takano, E. & Breitling, R. Detecting Sequence Homology at the
        Gene Cluster Level with MultiGeneBlast. *Mol. Biol. Evol.* **30**, 1218-1223,
        doi:10.1093/molbev/mst025 (2013).

70   Owen, C. D. *et al.* Streptococcus pneumoniae NanC: STRUCTURAL INSIGHTS
        INTO THE SPECIFICITY AND MECHANISM OF A SIALIDASE THAT
        PRODUCES A SIALIDASE INHIBITOR. *J. Biol. Chem.* **290**, 27736-27748,
        doi:10.1074/jbc.M115.673632 (2015).

71   Perutka, J., Wang, W., Goerlitz, D. & Lambowitz, A. M. Use of computer-designed
        group II introns to disrupt Escherichia coli DExH/D-box protein and DNA helicase
        genes. *J. Mol. Biol.* **336**, 421-439 (2004).

72   Mayer, M. & Meyer, B. Characterization of Ligand Binding by Saturation Transfer
        Difference NMR Spectroscopy. *Angewandte Chemie International Edition* **38**, 1784-
        1788, doi:10.1002/(SICI)1521-3773(19990614)38:12<1784::AID-
        ANIE1784>3.0.CO;2-Q (1999).

73   Hwang, T. L. & Shaka, A. J. Water Suppression That Works. Excitation Sculpting
        Using Arbitrary Wave-Forms and Pulsed-Field Gradients. *Journal of Magnetic
        Resonance, Series A* **112**, 275-279, doi:https://doi.org/10.1006/jmra.1995.1047
        (1995).

74   Nepravishta, R., Walpole, S., Tailford, L., Juge, N. & Angulo, J. Deriving Ligand
        Orientation in Weak Protein-Ligand Complexes by DEEP-STD NMR Spectroscopy
        in the Absence of Protein Chemical-Shift Assignment. *ChemBioChem* **20**, 340-344,
        doi:10.1002/cbic.201800568 (2019).

75   Mayer, M. & James, T. L. NMR-based characterization of phenothiazines as a RNA
        binding scaffold. *J. Am. Chem. Soc.* **126**, 4453-4460, doi:10.1021/ja0398870 (2004).

76   Sanchez-Weatherby, J. *et al.* VMXi: a fully automated, fully remote, high-flux in situ
        macromolecular crystallography beamline. *Journal of synchrotron radiation* **26**, 291-
        301, doi:10.1107/s1600577518015114 (2019).

77   Krissinel, E., Uski, V., Lebedev, A., Winn, M. & Ballard, C. Distributed computing
        for macromolecular crystallography. *Acta crystallographica. Section D, Structural
        biology* **74**, 143-151, doi:10.1107/s2059798317014565 (2018).

78   Vagin, A. & Teplyakov, A. Molecular replacement with MOLREP. *Acta
        crystallographica. Section D, Biological crystallography* **66**, 22-25,
        doi:10.1107/s0907444909042589 (2010).

972 79 Keegan, R. M. & Winn, M. D. MrBUMP: an automated pipeline for molecular
973   replacement. *Acta crystallographica. Section D, Biological crystallography* **64**, 119-
974   124, doi:10.1107/s0907444907037195 (2008).
975 80 van Beusekom, B., Joosten, K., Hekkelman, M. L., Joosten, R. P. & Perrakis, A.
976   Homology-based loop modelling yields more complete crystallographic protein
977   structures. *bioRxiv*, 329219, doi:10.1101/329219 (2018).
978 81 Emsley, P. Tools for ligand validation in Coot. *Acta crystallographica. Section D,*
979   *Structural biology* **73**, 203-210, doi:10.1107/s2059798317003382 (2017).
980 82 Smart, O. S. *et al.* Exploiting structure similarity in refinement: automated NCS and
981   target-structure restraints in BUSTER. *Acta crystallographica. Section D, Biological*
982   *crystallography* **68**, 368-380, doi:10.1107/s0907444911056058 (2012).
983 83 Langer, G., Cohen, S. X., Lamzin, V. S. & Perrakis, A. Automated macromolecular
984   model building for X-ray crystallography using ARP/wARP version 7. *Nature*
985   *protocols* **3**, 1171-1179, doi:10.1038/nprot.2008.91 (2008).
986 84 Winn, M. D., Murshudov, G. N. & Papiz, M. Z. in *Methods Enzymol.* Vol. 374   300-
987   321 (Academic Press, 2003).
988 85 Williams, C. J. *et al.* MolProbity: More and better reference data for improved all-
989   atom structure validation. *Protein Sci.* **27**, 293-315, doi:10.1002/pro.3330 (2018).
990 86 McCoy, A. J. *et al.* Gyre and gimble: a maximum-likelihood replacement for
991   Patterson correlation refinement. *Acta crystallographica. Section D, Structural*
992   *biology* **74**, 279-289, doi:10.1107/s2059798318001353 (2018).
993 87 Vonrhein, C. *et al.* Data processing and analysis with the autoPROC toolbox. *Acta*
994   *crystallographica. Section D, Biological crystallography* **67**, 293-302,
995   doi:10.1107/s0907444911007773 (2011).
996 88 Tickle, I. J., Flensburg, C., Keller, P., Paciorek, W., Sharff, A., & Vonrhein, C.,
997   Bricogne, G. STARANISO. *Cambridge, United kingdom: Global Phasing Ltd.*
998   (2018).
999 89 Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of
1000   insertions, deletions and gene fusions. *Genome Biology* **14**, R36, doi:10.1186/gb-
1001   2013-14-4-r36 (2013).
1002 90 Anders, S. & Huber, W. Differential expression analysis for sequence count data.
1003   *Genome Biology* **11**, R106, doi:10.1186/gb-2010-11-10-r106 (2010).
1004 91 Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat.*
1005   *Methods* **9**, 676, doi:10.1038/nmeth.2019 (2012).

1006 **Figure legends**

1007 **Figure 1** *R. gnavus* **ATCC 29149** *nan* **operon a)** Diagram depicting the genomic
1008 organisation of the *nan* operon RUMGNA_02701 (putative sialic acid esterase; tan)
1009 RUMGNA_02700 (putative YhcH family protein; dark blue), RUMGNA_02699 (predicted
1010 transcriptional regulator; purple), RUMGNA_02698 – 02696 (putative sialic acid ABC
1011 transporter of the SAT2 family; green), RUMGNA_02695 (putative oxidoreductase, pink),
1012 RUMGNA_02694 (RgNanH (Intramolecular *trans* sialidase), gold), RUMGNA_02693 (NanE
1013 (epimerase), blue), RUMGNA_02692 (NanA (aldolase), dark green), RUMGNA_02691
1014 (kinase) (NanK, red). **b)** qPCR analysis showing fold changes in expression of *nan* genes
1015 when *R. gnavus* was grown with 3'SL or 2,7-anhydro-Neu5Ac compared to glucose using
1016 ΔΔCt calculation. Error represent standard deviation and are based on three biological

1017 replicates analysed in triplicate. Statistical significance was determined using a 1-way
1018 ANOVA with a Dunnett's multiple comparison test. NS – no significant change in expression
1019 ($p > 0.05$), * - p 0.05 – 0.01, ** - p 0.01 – 0.001, *** - $p < 0.001$.

1020

1021 **Figure 2 Steady-state fluorescence analysis of ligand binding to *Rg*SBP**. Fluorescence
1022 emission spectrum of 0.5 µM *Rg*SBP excited at 297 nm in the presence or absence of **a)**
1023 2,7-anhydro-Neu5Ac or **b)** Neu5Ac. The data shown are representative of triplicate readings.
1024 **c)** Titration of 0.5 µM *Rg*SBP with 2,7-anhydro-Neu5Ac. The data represents the mean of
1025 triplicate readings. **d)** Displacement of Neu5Ac with 2,7-anhydro-Neu5Ac, six sequential
1026 additions of 10 µM Neu5Ac to 0.5 µM *Rg*SBP followed by one addition of 10 µM 2,7-
1027 anhydro-Neu5Ac, and displacement of 2,7-anhydro-Neu5Ac with Neu5Ac, one addition of 10
1028 µM 2,7-anhydro-Neu5Ac followed by 6 subsequent additions of 10 µM Neu5Ac. The data
1029 shown are representative of triplicate experiments, the signal peaks are artefacts attributed
1030 to external light during sample addition.

1031 **Figure 3 Biophysical analysis of ligand binding to *Rg*SBP.** ITC Isotherms of *Rg*SBP
1032 binding to **a**) 2,7-anhydro-Neu5Ac or **b**) Neu5Ac, showing both DP – differential power and
1033 ΔH – enthalpy change. The data shown are representative of triplicate experiments. **c)**
1034 Saturation Transfer Difference (STD) NMR binding epitope mapping of 2,7-anhydro-Neu5Ac
1035 interacting with *Rg*SBP. The initial slopes STD_0 (%) were normalized against the highest
1036 STD_0, assigned as 100%. The obtained factors were then classified as weak (0-60 %),
1037 intermediate (60-80 %), and strong (80-100%) and used to identify the close contacts found
1038 at the interface of binding, data is representative of triplicate readings **d)** Average Differential
1039 Epitope Mapping (DEEP) STD factors for 2,7-anhydro-Neu5Ac obtained saturating *Rg*SBP
1040 in spectral regions 0.6, 0.78, 1.44 ppm for aliphatic and 7.5, 7.23, 7.27 ppm for aromatic
1041 residues. Each differential mapping epitope obtained using different saturation frequencies
1042 are combined and the average DEEP STD is calculated resulting in five points for each
1043 frequency and a total of fifteen points for each proton receiving saturation. The data reported
1044 are the mean ± SEM of a sample of data of fifteen points for each proton receiving
1045 saturation.

1046 **Figure 4 *R. gnavus* sialic acid aldolase enzymatic reaction. a)** Change of A_{340nm} over
1047 time using *R. gnavus* sialic acid aldolase (*Rg*NanA) with Neu5Ac (pink) or 2,7-anhydro-
1048 Neu5Ac (orange), or *E. coli* sialic acid aldolase (*Ec*NanA) with Neu5Ac (black) or 2,7-
1049 anhydro-Neu5Ac (green) reactions coupled to lactate dehydrogenase, error bars represent
1050 standard error from 3 independent experiments. **b)** Michaelis-Menten plot of *Rg*NanA rate of
1051 reaction with increasing concentration of Neu5Ac, error bars represent standard error. The
1052 rate of reaction at each concentration (µM NADH) was determined in triplicate by measuring

1053 A_{340nm} change using a standard curve. The mean value is plotted with standard error of the

1054   meaning shown with error bars **c)** Cartoon representation of the wild type *Rg*NanA crystal

1055   structure showing the (β/α8) TIM barrel organisation and Lys167 as yellow sticks. **d)** The

1056   *Rg*NanA K167A active site is shown in orange with bound Neu5Ac in the open-chain ketone

1057   form shown in cyan. The green mesh represents the Neu5Ac F_o-F_c difference map at the 3σ

1058   level (for a stereo image of Neu5Ac and F_o-F_c difference map see **Supplementary Figure**

1059   **5d**). Hydrogen bonding interactions are depicted using black dashed lines. In addition, the

1060   unbound *Rg*NanA wt active site is shown in grey.

1061   **Figure 5 RUMGNA_02695 catalyses the conversion of 2,7-anhydro-Neu5Ac to Neu5Ac.**

1062   **a**) High Performance Liquid Chromatography (HPLC) expand acronyms number replicates

1063   with same outcome analysis of DMB labelled RUMGNA_02695 reactions with 2,7-anhydro-

1064   Neu5Ac using different co-factors. NAD (black), NADH (pink), FAD (blue), no co-factor

1065   (brown), and a Neu5Ac standard (green), data is representative of five independent

1066   experiments. **b**) Michaelis-Menten plot of the rate of reaction for RUMGNA_02695 with

1067   increasing concentration of 2,7-anhydro-Neu5Ac. The rate of reaction (μM NADH) at each

1068   concentration was determined in triplicate by measuring A_{340nm} change and using a

1069   standard curve, the mean value is plotted with standard error of the meaning shown with

1070   error bars

1071   **Figure 6 Colonisation of germ-free C57BL/6J mice with *R. gnavus* ATCC 29149 wild-**

1072   **type or *nan* mutant strains.** Mice were monocolonised with a and b sample size (n) and

1073   define centre measure mean(**a**) *R. gnavus* wild-type (black; n = 4) or *nan* mutant (red; n = 4)

1074   strains individually or (**b**) in competition (n = 4). Mice were orally gavaged with 1×10^8 of each

1075   strain, faecal samples were analysed at 3,7 and 14 days after inoculation and caecal

1076   samples at 14 days after inoculation using qPCR, centre line denotes the mean. **(c)**

1077   Fluorescent in situ hybridisation (FISH) and immunostaining of the colon from *R. gnavus*

1078   monocolonised C57BL/6 mice. *R. gnavus* ATCC 29149 and *R. gnavus nan* mutant are

1079   shown in red. The mucus layer is shown in green and an outline of the mucus is shown in

1080   the first panels. Cell nuclei were counterstained with Sytox blue, shown in blue. Scale bar:

1081   20 μm. Image is representative of 70 total images **(d)** Quantification of the distance between

1082   the leading front of bacteria and the base of the mucus layer. A total of 70 images of stained

1083   colon from 8 *R. gnavus* monocolonised mice were analysed. The asterisks (***) show the

1084   significance (P=0.0135, by linear mixed model analysis, including fixed effects of genotype

1085   and area and random effects of mouse and each individual image. There was substantial

1086   spatial correlation between adjacent observations and so an AR(1) correlation structure was

1087   added. The resulting model had no residual autocorrelation as judged by visual inspection

1088   of autocorrelation function. The nmle package version 3.1-137 using R version 3.5.3 was

1089   used to estimate the model), centre point indicates the mean, box limits, upper and lower

1090   quartiles; whiskers, minimum and maximum.

1091

a

RUMGNA_02701 · RUMGNA_02700 · RUMGNA_02699 · RUMGNA_02698 · RUMGNA_02697 · RUMGNA_02696 · RUMGNA_02695 · RUMGNA_02694 · RUMGNA_02693 · RUMGNA_02692 · RUMGNA_02691

nanH · *nanE* · *nanA* · *nanK*

b

Fold change in transcription (VS glucose)

□ 2,7-anhydro-Neu5Ac
▨ 3'SL

a

b

c

d

SBP

SBP + 10 uM
2,7-anhydro-Neu5Ac

SBP + 20 uM
2,7-anhydro-Neu5Ac

SBP

SBP + 10 uM Neu5Ac

SBP + 20 uM Neu5Ac

SBP + 70 uM Neu5Ac

Displacement of Neu5Ac with 2,7-anhydro-Neu5Ac

Displacement of 2,7-anhydro-Neu5Ac with Neu5Ac

a

b

c

d

Rg NanA + Neu5Ac
Ec NanA + Neu5Ac
Rg NanA + 2,7-anhydro-Neu5Ac
Ec NanA + 2,7-anhydro-Neu5Ac

a

Neu5Ac standard

RUMGNA_02695 + 2,7-anhydro-Neu5Ac

RUMGNA_02695 + 2,7-anhydro-Neu5Ac + FAD

RUMGNA_02695 + 2,7-anhydro-Neu5Ac + NADH

RUMGNA_02695 + 2,7-anhydro-Neu5Ac + NAD

b

a

Day 3   Day 7   Day 14   Caecal content

Colonisation Level
(Log₁₀ *Rg*NanH copies / mg)

Limit of detection

b

Colonisation Level
(Log₁₀ *Rg*NanH copies / mg)

Limit of detection

c

| FISH probe | Muc2 | Merged |

R. gnavus

R. gnavus Nan mutant

d

Distance (µm)

***

Rg WT   Rg Nan⁻