Elucidation of a sialic acid metabolism pathway in mucus-foraging Ruminococcus gnavus unravels mechanisms of bacterial adaptation to the gut

Andrew Bell 1, Jason Brunt 1,8, Emmanuelle Crost 1, Laura Vaux 1, Ridvan Nepravishta 2, C. David Owen 3,4, Dimitrios Latousakis 1, An Xiao 5, Wanqing Li 5, Xi Chen 5, Martin A. Walsh 3,4, Jan Claesen 6, Jesus Angulo 2, Gavin H. Thomas 7 and Nathalie Juge 1,*

Sialic acid (N-acetylneuraminic acid (Neu5Ac)) is commonly found in the terminal location of colonic mucin glycans where it is a much-coveted nutrient for gut bacteria, including Ruminococcus gnavus. R. gnavus is part of the healthy gut microbiota in humans, but it is disproportionately represented in diseases. There is therefore a need to understand the molecular mechanisms that underpin the adaptation of R. gnavus to the gut. Previous in vitro research has demonstrated that the mucin-glycan-foraging strategy of R. gnavus is strain dependent and is associated with the expression of an intramolecular trans-sialidase, which releases 2,7-anhydro-Neu5Ac, rather than Neu5Ac, from mucus. Here, we unravelled the metabolism pathway of 2,7-anhydro-Neu5Ac in R. gnavus that is underpinned by the specific epithelial mucus transporter for 2,7-anhydro-Neu5Ac and by the action of an oxido-reductase that converts 2,7-anhydro-Neu5Ac into Neu5Ac, which then becomes a substrate of a Neu5Ac-specific aldolase. Having generated an R. gnavus nan-cluster deletion mutant that lost the ability to grow on sialylated substrates, we showed that—in gnotobiotic mice colonized with R. gnavus wild-type (WT) and mutant strains—the fitness of the nan mutant was significantly impaired, with a reduced ability to colonize the mucus layer. Overall, we revealed a unique sialic acid pathway in bacteria that has important implications for the spatial adaptation of mucin-foraging gut symbionts in health and disease.

The gastrointestinal (GI) tract is heavily colonized with bacteria that play a vital role in human health. The composition of gut microbiota not only varies longitudinally along the GI tract, but also transversally from the mucosa to the lumen. In the colon, the epithelium is covered by a bilayer of mucus, with the outer mucus layer providing a natural habitat for the commensal bacteria, whereas the stratified inner mucus layer restricts bacterial access to the epithelium. Mucin proteins that form the mucus layer are highly glycosylated with a diverse and complex array of O-glycan structures containing N-acetylgalactosamine, galactose and N-acetylglucosamine (GlcNAc), and are usually terminated by fucose, sialic acid (Neu5Ac) residues and sulfate. The terminal mucin glycans have been proposed to serve as metabolic substrates, providing a nutritional advantage to bacteria that have adapted to the GI mucosal environment. The proportion of these terminal glycan epitopes varies along the GI tract with a decreasing gradient of fucose and an increasing gradient of sialic acid from the ileum to the rectum in humans. Sialic acid therefore represents a much-coveted source of nutrients for the gut bacteria that inhabit the mucus niche in the large intestine. In bacteria, the genes involved in the metabolism of sialic acid are usually found clustered together forming different nan gene clusters. The canonical nanATEK cluster was first described in Escherichia coli, encompassing genes that encode the enzymes N-acetylneuraminic lyase (NanA), epimerase (NanE) and kinase (NanK), which are necessary for the catabolism of sialic acid into GlcNAc-6-P following its transport through the major facilitator superfamily (MFS) transporter NanT. An alternative pathway for sialic acid metabolism has been described for Bacteroides fragilis that relies on the action of an MFS transporter (NanT), an aldolase (NanL) and a novel ManNAc-6-P epimerase (also known as NanE)—which are encoded in the nanLET operon—and a hexokinase (RokA); this pathway leads to the conversion of Neu5Ac into GlcNAc-6-P. GlcNAc-6-P is then converted into fructose-6-P, which is a substrate in the glycolytic pathway, by genes encoding NagA (GlcNAc-6-P deacetylase) and NagB (glucosamine-6-P deaminase). The majority of bacteria that harbour a nan cluster colonize mucus regions of the human body. To gain access to this substrate, bacteria are dependent on sialic acid release and uptake. Several gut bacteria species, including strains of Clostridia, Bacteroides, Bifidobacterium longum, Vibrio cholerae, R. gnavus and Akkermansia muciniphila, express sialidases to release sialic acid from their terminal location in mucins. As sialic acid cleavage occurs outside the cell, bacteria have evolved multiple mechanisms to capture this important nutrient from their environment. Such transport mechanisms involve the NanT MFS transporter, which is used by E. coli and B. fragilis and which—in E. coli—has been demonstrated biochemically to be an H+-coupled symporter or secondary transporter from the...
sodium solute symport family that is present in *C. difficile* and *Salmonella typhimurium*. High-affinity transport of sialic acid is mediated by substrate-binding protein-dependent systems, including a tripartite ATP-independent periplasmic (TRAP) transporter, SiaPQM and ATP-binding cassette (ABC) transporters. The sialic acid ABC transporters are classified into three types as follows: SAT, SAT2 and SAT3. To date, all of these transporters have been shown to transport Neu5Ac, with some also being able to transport the related sialic acid Neu5Gc and also being able to transport the related sialic acid Neu5Gc and Neu5Gc. The ability of *R. gnavus* to produce and metabolize 2,7-anhydro-Neu5Ac provides these strains with a competitive nutritional advantage in mucus by enabling the scavenging of sialic acid from mucins in a form that other bacteria do not have access to.

*R. gnavus* is not only an early colonizer of the infant gut but also persists in the adult gut, where it is one of the 57 species that are detected in more than 90% of human faecal samples. *R. gnavus* belongs to the Firmicutes division, Clostridia class and XIVa cluster, and the Lachnospiraceae family, and is considered to be a prevalent member of the normal gut microbiota. Furthermore, *R. gnavus* shows a disproportiorate representation in a number of diseases, such as inflammatory bowel disease. The ability of *R. gnavus* strains to utilize mucin glycans as a source nutrient is associated with the expression of an intramolecular transportase, SiaPQM and ATP-binding cassette (ABC) transporters. The sialic acid ABC transporters are classified into three types as follows: SAT, SAT2 and SAT3. To date, all of these transporters have been shown to transport Neu5Ac, with some also being able to transport the related sialic acid Neu5Gc and 2-keto-3-deoxy nononic acid. In *R. gnavus* ATCC 29149 and ATCC 35913 strains, the IT-sialidase (NanH) is part of a *nan* cluster, which is induced when the cells are grown in the presence of mucin and absent in strains that do not degrade mucin-glycans, such as *R. gnavus* E1. We enzymatically synthesized 2,7-anhydro-sialic acid derivatives, which we used to confirm the ability of IT-sialidase-expressing *R. gnavus* strains to grow on 2,7-anhydro-Neu5Ac as a sole source carbon. We proposed that the ability of *R. gnavus* strains to produce and metabolize 2,7-anhydro-Neu5Ac provides these strains with a competitive nutritional advantage in mucus by enabling the scavenging of sialic acid from mucins in a form that other bacteria do not have access to.

To test this hypothesis and gain insights into the 2,7-anhydro-Neu5Ac metabolism pathway in *R. gnavus*, we identified candidate genes of the *nan* cluster involved in 2,7-anhydro-Neu5Ac transport and metabolism and characterized the proteins. Using fluorescence spectroscopy, saturation transfer difference nuclear magnetic resonance spectroscopy (STD NMR) and isothermal titration calorimetry (ITC), we showed that the solute binding protein (SBP) from the *R. gnavus* ABC transporter was specific for 2,7-anhydro-Neu5Ac. Biochemical analyses also uncovered oxidoreductase activity that enables the conversion of 2,7-anhydro-Neu5Ac into Neu5Ac and confirmed the specificity of the sialic acid aldolase for Neu5Ac. Finally, we showed that the *nan* cluster is essential to support anaerobic growth of the bacteria on sialoconjugates in vitro and in vivo fitness of the bacteria using gnotobiotic mice colonized with WT or *nan*-mutant *R. gnavus*. These data demonstrate a unique sialic acid metabolism pathway in bacteria, which provides *R. gnavus* with a competitive strategy to colonize the mucus niche.

### Results

**Identification of genes involved in 2,7-anhydro-Neu5Ac metabolism in *R. gnavus nan* cluster.** First, we analysed the transcriptional activity of the *nan* cluster using quantitative PCR.
with reverse transcription (RT–qPCR) in R. gnavus ATCC 29149 grown on 2,7-anhydro-Neu5Ac or α2–3-sialyllactose (3’S-L) as the sole carbon source. Expression of all of the genes that constitute the nan cluster was induced after bacterial growth on 2,7-anhydro-Neu5Ac or 3’S-L compared with growth on glucose, whereas the expression of the two genes that flank the cluster (RUMGNA_02702, RUMGNA_02690) remained unchanged (Fig. 1). The 3’S-L and 2,7-anhydro-Neu5Ac induced the transcription of the nan genes between tenfold and eightyfold. Both of the substrates induced similar changes, which is not unexpected as 2,7-anhydro-Neu5Ac is the sialic acid form produced by R. gnavus ATCC 29149 from 3’S-L. These results indicate that the R. gnavus nan operon is dedicated to the metabolism of 2,7-anhydro-Neu5Ac from host sialoglycans.

A sequence similarity network (SSN) analysis was then performed to identify the proteins encoded by the nan cluster, which are associated with the ability of the bacteria to metabolize 2,7-anhydro-Neu5Ac over Neu5Ac. As expected, the IT-sialidase from R. gnavus strains clustered together with proteins from Streptococcus pneumoniae strains, the genomes of which are known to encode IT-sialidases (as well as other sialidases)17,18 (Supplementary Fig. 1a). Other co-occurring bacterial species include Ruminococcus torques, Lactobacillus salivarius, Staphylococcus pseudintermedius, Streptococcus infantis and Streptococcus mitis. Bacterial species clustering for RgNanH also shared clusters for proteins encoding RUMGNA_02698, the predicted SBP that gives specificity to ABC transporters, RUMGNA_02692 (sialic acid aldolase), the first protein of the canonical Neu5Ac metabolism pathway, and RUMGNA_02695, a putative oxidoreductase, suggesting that these proteins may be associated with 2,7-anhydro-Neu5Ac metabolism (Supplementary Fig. 1, Supplementary Table 1). By contrast, RUMGNA_02701, with homology to sialic acid esterase proteins, and RUMGNA_02700, with homology to the YhcH protein family, did not cluster with proteins from the same set of bacteria (Supplementary Fig. 1, Supplementary Table 1). The candidate genes were then heterologously expressed, and the recombinant proteins were purified as described in the Methods.

**Specificity of R. gnavus sialic acid transporter for 2,7-anhydro-Neu5Ac.** First, we investigated the ligand specificity of recombinant SBP (RUMGNA_02698), RsBP, by measuring changes in the intrinsic protein fluorescence after the addition of potential ligands. The addition of 10 μM or 20 μM 2,7-anhydro-Neu5Ac resulted in a significant shift at 350 nm, causing an approximately 16% quench in the fluorescence (Fig. 2a). By contrast, the addition of Neu5Ac at concentrations of 10 μM, 20 μM or 70 μM caused no change in the spectrum intensity, suggesting an absence of binding (Fig. 2b). Titration of 0.5 μM RsBP with 2,7-anhydro-Neu5Ac resulted in a hyperbolic curve with a dissociation constant ($K_d$) of 1.349 ± 0.046 μM (Fig. 2c). To confirm the specificity of 2,7-anhydro-Neu5Ac compared with Neu5Ac, we monitored sequential changes in fluorescence following the addition of 10 μM of the two ligands. When Neu5Ac was added first, no change in fluorescence was observed and a quench was observed when 2,7-anhydro-Neu5Ac was added (Fig. 2d). By contrast, the quench was observed when 2,7-anhydro-Neu5Ac was added first, and the addition of 10 μM Neu5Ac caused no further change in the fluorescence intensity (Fig. 2d), indicating that Neu5Ac is unable to displace 2,7-anhydro-Neu5Ac, and further supporting the specificity of the interaction between RsSBP and 2,7-anhydro-Neu5Ac.
The affinity of the interaction between RgSBP and sialic acid ligands was further assessed by ITC. RgSBP bound to 2,7-anhydro-Neu5Ac with a $K_d$ of 2.42 ± 0.27 μM (Fig. 3a) and no binding was observed when Neu5Ac was used as the ligand (Fig. 3b), in agreement with the findings from the fluorescence spectroscopy experiment. The binding of 2,7-anhydro-Neu5Ac...
revealed a thermodynamic signature with both entropic (temperature (K) × entropy change (−TΔS) = −7.05 ± 0.08 kcal mol⁻¹) and enthalpic (enthalpy change (ΔH) = −0.93 ± 0.03 kcal mol⁻¹) components contributing favourably to the binding process (free energy change (ΔG) = −7.99 ± 0.05 kcal mol⁻¹; Fig. 3a).

To gain structural insights into the unique ligand specificity of RgSBP, we conducted STD NMR studies with RgSBP in the presence of 2,7-anhydro-Neu5Ac or Neu5Ac. The transfer of magnetization as saturation from the protein to the ligand was clearly observed for 2,7-anhydro-Neu5Ac but not for Neu5Ac, confirming that RgSBP preferentially selects 2,7-anhydro-Neu5Ac (Supplementary Fig. 2). STD NMR epitope binding revealed that protons H3, H4 and H6 showed the highest STD (%) factors, indicating the close contacts present at the interface of binding (Fig. 3c). By contrast, protons H7, H8 and H9, and protons belonging to the CH₃ group showed lower percentage STD and are expected to be more exposed to solvent. For the differential epitope mapping (DEEP)-STD NMR experiment, 4-hydroxy-1-oxyl-2,2,6,6-tetramethylpiperidine (TEMPOL) was used to gain insights into the RgSBP binding pocket (Supplementary Fig. 3). We found that protons H4, H6, H7, H8 and H9’ were preferentially oriented toward aromatic residues, whereas H3 and protons belonging to the CH₃ group were oriented towards aliphatic residues (Fig. 3d).

Together these data demonstrate that RgSBP specifically binds to 2,7-anhydro-Neu5Ac but not to Neu5Ac, in line with the growth profile of R. gnavus ATCC 29149 on these substrates.

**Specificity of R. gnavus sialic acid aldolase for Neu5Ac.** The substrate specificity of recombinant sialic acid aldolase (RUMGNA_02692; RgNanA) was determined using a coupled activity assay in which pyruvate released during the conversion of Neu5Ac to ManNAc is converted to lactate by a lactate dehydrogenase, and the conversion of NADH to NAD⁺ is measured as a decrease in absorbance at 340 nm (A₃₄₀). RgNanA and EcNanA (E. coli Neu5Ac lyase/aldolase, used as a control) showed activity against Neu5Ac, whereas neither enzyme showed activity against 2,7-anhydro-Neu5Ac (Fig. 4a). The product of the reaction with Neu5Ac was confirmed to be ManNAc by high-performance liquid chromatography (HPLC; Supplementary Fig. 4). RgNanA showed a turnover number (kₑcat) of 2.757 ± 0.033 s⁻¹ and a Michaelis constant (Kₘ) of 1.473 ± 0.098 mM (Fig. 4b). These kinetic parameters are consistent with values from other bacterial sialic acid aldolases that have been described previously (Supplementary Table 2).

The RgNanA crystal structure presents as a (β/α)₈ triose-phosphate isomerase (TIM) barrel with an adjacent three-helix...
Impact of the *R. gnavus* nan cluster on in vitro growth and in vivo colonization of mice. The Clostron transformation method was successfully applied to *R. gnavus* ATCC 29149, enabling the generation of nan deletion mutants with an erythromycin resistance gene present in either the sense or antisense direction (relative to RgNanH). The recombination event was confirmed by PCR (Supplementary Fig. 7a) and the expression of the full cluster was tested using qPCR (Supplementary Fig. 7b). The level of expression of the genes flanking the cluster, RUMGNA_02690 and RUMGNA_02702, were comparable to the expression of these genes in the WT strain, as also observed for the first three genes of the nan cluster, RUMGNA_02701–02699; however, the nan cluster genes RUMGNA_02698–02691 showed significantly reduced expression compared with the WT strain. The *R. gnavus* ATCC 29149 WT strain was able to utilize both 3′SL and 2,7-anhydro-Neu5Ac as a sole carbon source, whereas no growth was detected when using the nan deletion mutants on these substrates (Supplementary Fig. 8).

To assess the impact of the nan cluster on the fitness of *R. gnavus* in vivo, germ-free C57BL/6j mice were gavaged with $1 \times 10^8$ colony-forming units (CFU) of *R. gnavus* ATCC 29149 or *R. gnavus* antisense nan deletion mutant, or a mixture of the WT and nan mutant strains at $1 \times 10^8$ CFU each (Fig. 6). During the monoclonization experiments, both the WT and nan mutant strains were detectable in the faecal content at day 3, 7 and 14 after gavage at mean levels of between $1 \times 10^8$ and $1 \times 10^7$ bacteria mg$^{-1}$ of material (Fig. 6a). Both strains were also detected in the caecal content of monoclonized mice that were euthanized at day 14. The absence of the nan cluster did not affect the gene expression response of mice, as shown by RNA sequencing (RNA-seq; Supplementary Fig. 9). In competition experiments, the WT strain reached mean colonization levels that were comparable to the levels obtained during monoclonization, whereas the mutant strain was severely outcompeted, reaching only $2 \times 10^4$ copies mg$^{-1}$ at day 3, before decreasing further by day 7 and day 14 to below the level of detection in the faecal and caecal contents (Fig. 6b). The impact of the nan deletion on the location of *R. gnavus* within the mucus layer was determined in monoclonized mice by measuring the distance of the nan mutant or WT R. gnavus strains to the epithelial layer throughout the colon by fluorescence in situ hybridization (FISH) imaging using confocal microscopy. The data showed that the nan mutant resided 19.70 µm from the epithelial layer, which was 5.06 µm further away than the WT strain (14.64 µm; Fig. 6c,d).
Sialic acid comprises a family of nine-carbon acidic sugars that are predominantly found on the cell-surface glycans of humans and other animals. Neu5Ac, the most common form of sialic acid in humans, is predominantly found on the cell-surface glycans of humans and other animals. Neu5Ac, the most common form of sialic acid in humans, is predominantly found on the cell-surface glycans of humans and other animals.

**Discussion**

Sialic acid metabolism is vital to the ability of bacteria to utilize mucin as a nutrient source. The SAT2 transporter is expected to be coupled with an MsiK-like ATPase encoded elsewhere in the genome, with RUMGNA_03040 sharing 59% identity with the S. pneumoniae MsiK. Interestingly, in contrast to S. pneumoniae, R. gnavus does not encode SAT or SAT3 transporters, which are known to recognize Neu5Ac, and SAT3 is required for growth on Neu5Ac.

By studying the RgSBP subunit, we have discovered that SAT2 is a specific transporter for 2,7-anhydro-Neu5Ac with a $K_d$ of $2.42 \pm 0.27 \mu M$ that does not bind to Neu5Ac. Using STD NMR and DEEP-STD NMR, we characterized the orientation of the ligand in the binding site and the contribution of aromatic and aliphatic residues in the RgSBP 2,7-anhydro-Neu5Ac binding pocket. The lower affinity compared with bacterial SAT (SatA) transporters specific for Neu5Ac that have been previously...
described, which bind in the nM range\textsuperscript{14}, might be consistent with the ‘exclusive’ access of the bacteria to the 2,7-anhydro-Neu5Ac substrate. Taken together, these findings indicate that the ability of \textit{R. gnavus} strains to grow on 2,7-anhydro-Neu5Ac (and not on Neu5Ac) can be explained by the high level of specificity of RgSBP (RUMGNA\_02698). RgSBP is also orthologous (72% identity and 86% similarity) with the SBP from the putative sialic transporters in \textit{Streptococcus sanguinis} SK36 (SSA\_0076) and \textit{Streptococcus gordonii} str. Challis substr. CH1 (SGO\_0122)\textsuperscript{15}. It would therefore be of interest to determine the specificity of \textit{Streptococcus} SBPs towards 2,7-anhydro-Neu5Ac.

Once inside the cell, the 2,7-anhydro-Neu5Ac needs to be converted back into Neu5Ac to become a substrate for the sialic acid aldolase. RgNanOx (RUMGNA\_02695) was identified as the oxidoreductase that catalyses the conversion of 2,7-anhydro-Neu5Ac into Neu5Ac, following a mechanism of action that remains to be determined. Bioinformatic analysis identified close homologues of this protein in a range of bacterial species, including \textit{YjhcC} from \textit{E. coli} (Supplementary Fig. 1, Supplementary Table 1). Neu5Ac is then converted into ManNac and pyruvate through the action of RgNanA (RUMGNA\_02692), a Neu5Ac-specific aldolase with structural features that are conserved with NanA proteins from the \textit{nan} canonical pathway.

MultiGeneBlast analysis revealed that predicted homologues of the \textit{R. gnavus nan} cluster are shared by a limited number of species, including 37 homologous clusters in \textit{S. pneumoniae} isolates, \textit{Streptococcus suis} A7, \textit{Blautia hansenii} DSM 20583, \textit{Blautia} sp. YLS8 and \textit{Intestimonas butyricippens} AF211 (Supplementary Figs. 10 and 11, Supplementary Table 5). The presence of this cluster in \textit{S. pneumoniae} suggests that it can also transport this unusual sialic acid into the cell. A major difference between NanB and NanH \textit{IT}-sialidase and NanC \textit{II}-sialidase cluster types is the associated transporter class; NanB and NanH clusters are associated with a carbohydrate ABC transporter, whereas NanC clusters are associated with a sodium:solute symporter\textsuperscript{15}, which may indicate a difference in the form of sialic acid being transported. These bioinformatic analyses support the specialisation of the \textit{R. gnavus nan} cluster.

We confirmed the importance of this metabolic pathway (Supplementary Fig. 12) by generating an \textit{R. gnavus nan} deletion mutant that was tested in vitro and in vivo using germ-free mice. In the in vivo competition experiments, the fitness of the mutant was impaired compared with the WT strain, showing a reduced ability to colonize the mucus layer. The \textit{nan} cluster is therefore important to maintain the spatial distribution of \textit{R. gnavus} strains in the gut. The ability of \textit{R. gnavus} strains that harbour a \textit{nan} cluster to penetrate further down into the mucus layer may contribute to protecting the bacteria from the constant mucus turnover. This mechanism may serve as a determinant that underlies the success of \textit{R. gnavus} as one of the most commonly shared species among individuals\textsuperscript{29,31}.

Together, these findings provide robust biochemical and in vivo evidence for the role of the \textit{R. gnavus nan} cluster in the adaptation of this important gut symbiont to the mucosal environment in the gut, providing defined molecular targets for bio-markers and therapeutic strategies.

**Methods**

**Materials.** All of the chemicals were obtained from Sigma-Aldrich unless otherwise stated. d-Glucose and Neu5Ac were purchased from Sigma-Aldrich and 3\' SL was purchased from Carbosynth. We prepared 2,7-anhydro-Neu5Ac as previously described\textsuperscript{14,15}.

**Bacterial strains and media.** \textit{R. gnavus} ATCC 29149 was routinely grown in an anaerobic cabinet (Don Whitley) in brain–heart infusion with yeast extract and hemin (BHI-YH) as previously described\textsuperscript{46,64}. Bacterial strains and media.

RT-qPCR. Total RNA was extracted from 3 ml of mid- to late-exponential-phase cultures of \textit{R. gnavus} ATCC 29149 in YCFA medium supplemented with one carbon source (d-Glucose, 3\' SL or 2,7-anhydro-Neu5Ac). Three biological replicates were performed for each carbon source. The cDNA was prepared before extraction using RNAProtect Bacteria Reagent (Qiagen) 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 the manufacturer’s instructions with an on-column DNase treatment. The purity and quantity of the extracted RNA was assessed using the NanoDrop spectrophotometer (Thermo Fischer Scientific) and a Qubit 2.0 fluorometer (Invitrogen).

qPCR was performed using an Applied Biosystems 7500 real-time PCR system (Life Technologies). One pair of primers was designed for each target gene using ProbeFinder v.2.45 (Roche Applied Science) to obtain an amplicon of around 60–80 bp. The primers were between 18 and 23 nucleotides, with a melting temperature (\textit{Tm}) of 59–60 °C (Supplementary Table 6). Calibration curves were prepared in triplicate for each pair of primers using 2.5-fold serial dilutions of \textit{R. gnavus} ATCC 29149 genomic DNA. The standard curves showed a linear relationship of log-transformed input DNA versus the threshold cycle (\textit{Ct}), with acceptable values for the slopes and the regression coefficients (\textit{R}²).

Total protein concentrations were also performed to check the specificity of the ampiclons. Each DNase-treated RNA sample (1 \textmu g) was converted into cDNA using the QuantiTect Reverse Transcription kit (Qiagen) according to the manufacturer’s instructions. DNase-treated RNA was also treated in the same way but without addition of the reverse-transcriptase (–RT). Each qPCR amplification curve was also performed to check the specificity of the ampiclons. Each DNase-treated RNA sample (1 \textmu g) was converted into cDNA using the QuantiFast SYBR Green PCR kit (Qiagen) according to the manufacturer’s instructions (except for the combined annealing/extension step, which was extended to 35 s). Data obtained with cDNA were analysed only when \textit{Ct} values of above 36 were obtained for the corresponding –RT. For each cDNA sample, the \textit{Ct} values obtained for each gene were used to calculate the \textit{2^-\Delta \Delta Ct} method using the housekeeping gene \textit{gyrB} (RUMGNA\_00867) as a reference gene and \textit{Ct} value as a reference condition. For each gene in each condition, the final value of the relative level of transcription (expressed as a fold change in gene transcription compared with glucose) is an average of three biological replicates; statistical analysis was performed using one-way ANOVA with Graphpad Prism (v.5.03).

**Cloning, expression, mutagenesis and purification of recombinant proteins.** \textit{R. gnavus} ATCC 29149 genomic DNA was purified from the cell pellet of a bacterial overnight culture (1 ml) following centrifugation (5,000g for 5 min) using the GeneJET Genomic DNA Purification Kit (ThermoFisher) according to the manufacturer’s instructions.

Full-length RgSBP excluding the signal sequence (residues 1–29), full-length RgNanA and full-length RgNanOx were amplified from \textit{R. gnavus} ATCC 29149 genomic DNA, and cloned into the pEHISTEV\textsuperscript{60} expression system, introducing a His-tag at the N terminus using primers, details of which are provided in Supplementary Table 6. DNA manipulation was performed in \textit{E. coli} DH5a cells. The plasmid DNAs were amplified and the DNA sequencing was performed by Eurofins Genomics, after plasmid preparation using the Monarch Plasmid Miniprep kit (New England Biolabs). The RgNanA active site mutant, K167A, was generated using the QuickChange Lightning mutagenesis kit (Agilent) and the primers listed in Supplementary Table 6. \textit{E. coli} BL21 (New England Biolabs) cells were transformed with the recombinant plasmid harbouring the gene of interest according to the manufacturer’s instructions. Expression was carried out in 800 ml autoinduction media (Terrific Broth Base with Trace Elements, ForMedium); the cells were grown for 3 h at 37 °C and then at 16 °C for 48 h, with shaking at 250 r.p.m. The cells were collected by centrifugation at 10,000 g for 20 min. The His-tagged proteins were purified by immobilized metal affinity chromatography and further purified by gel filtration (Superdex 75 column) on an Akta system (GE Health Care Life Sciences). Protein purification was assessed by standard SDS–polyacrylamide gel electrophoresis using NuPAGE Novex 4–12% Bis-Tris gels (Life Technologies). Protein concentrations were measured using NanoDrop spectrophotometer (Thermo Scientific) and using an extinction coefficient, which was calculated using ProtParam (ExPASy-Artimo, 2012) from the peptide sequence.

**Fluorescence spectroscopy.** All of the protein fluorescence experiments were performed using a FluoroMax 3 fluorescence spectrometer with the connecting water bath at 37 °C. Owing to the presence of 15 tyrosine residues, the protein was excited at 297 nm with slit widths of 5 nm. Under these conditions, the protein had a maximal emission at 331 nm. RgNanOx was excited at 0.2 \textmu M in 50 mM Tris pH 7.5 for all of the fluorescence experiments. Cumulative fluorescence changes from titration of the protein with ligand were plotted in
GraphPad and fitted to a single rectangular hyperbola. The \( K_a \) values were reported from three separate ligand titration experiments.

ITC. ITC experiments were performed using the PEAQ-ITC system (Malvern) with a cell volume of 200 \( \mu \)l. Before titration, protein samples were exhaustively dialysed into 50 mM Tris-HCl pH 7.5. The ligand was dissolved in the dialysis buffer. The cell protein concentration was 100 mM and the syringe ligand concentration was 2 mM. Controls with titrant (sugar) injected into the buffer only were subtracted from the data. The analysis was performed using the Malvern software, using a single-binding site model. The experiments were performed in triplicate.

Sialic acid aldolase activity assays. Aldolase activity was measured by monitoring the decrease in \( A_{280} \), while NADH was converted to NAD\(^+\) by lactate dehydrogenase in a coupled reaction in which pyruvate is released from sialic acid by the aldolase. The reactions were performed in a 100 \( \mu \)l volume with final concentrations of 150 mM NADH (Sigma-Aldrich), 0.5 U LDH (Sigma-Aldrich), 10 mM sialic acid (Neu5Ac or 2,7-anhydro-Neu5Ac) and 1.5 \( \mu \)g purified RgNanA or RgNanB (E. coli aldolase; 9027-60-5, Carboxylsyn) in 50 mM sodium phosphate buffer (pH 7.0). The reactions were performed at 37 °C and monitored using FLUOstar OPTIMA (BMG LABTECH). For kinetics experiments, the sialic acid concentration was varied at 20 mM, 10 mM, 5 mM, 4 mM, 2 mM, 1 mM, 0.4 mM, 0.2 mM and 0.1 mM, and the initial rate of reaction was determined for each concentration in triplicate before analysis was performed by fitting the data to a Michaelis–Menten plot using Graphpad Prism (v.5.03).

To monitor the production of ManNAc during the aldolase-catalysed reactions, 2-AB labelling was performed on the products from the above reactions. In brief, 50 ng GlcNAc was added to 10 \( \mu \)l reactions, 2-AB labelling was performed on the products from the above reactions, 2-AB labelling was performed on the products from the above reactions at 65 °C for 3 h. The labelling reagent was prepared by dissolving 50 mg 2-aminobenzamide in a solution containing 300 \( \mu \)l acetic acid and 700 \( \mu \)l dimethyl sulfoxide before 60 mg sodium cyanoborohydride was added. Following the addition of \( \text{H}_2\text{O} \) up to a total volume of 100 \( \mu \)l, the sample was transferred into an HPLC vial and 10 \( \mu \)l was loaded onto a HyperClone 3 \( \mu \)m ODS (C18) 120A 150 \( \times \)4.6 mm column. Mobile phases of 0.25% \( \text{H}_3\text{PO}_4 \) and 0.5% phosphoric acid and 1.5% tetrahydrofuran, as well as 50% acetonitrile, were used at a flow rate of 0.7 ml min\(^{-1}\).

Bioinformatics analyses. SSN. The InterPro families for RgNanH (glycoside hydrolase, family 34; IPR001860) and RgNanG (N-acetylmuramicylase; IPR005264) were identified using the UniProt database, and this family identifier was used to extract protein sequences using the Enzyme Function Initiative Enzyme Similarity Tool\(^{11}\). For the other proteins, the families found in the InterPro database were too large to be analysed, so the BLAST tool was therefore used with a maximum of 2,500 protein sequences extracted. The sequence BLAST tool of the InterPro database was too large to be analysed; the sequence BLAST tool of the InterPro database was too large to be analysed; the sequence BLAST tool of the InterPro database was too large to be analysed; the sequence BLAST tool of the InterPro database was too large to be analysed; the sequence BLAST tool of the InterPro database was too large to be analysed; the sequence BLAST tool of the InterPro database was too large to be analysed.

Cluster analysis. Homologous gene clusters were identified for the R. gnavus ATCC 29149 nan cluster\(^{47,70}\), and the clusters are summarized by organism and predicted gene content in Supplementary Table 5.

RgNanOx enzymatic activity assay. To assay RgNanOx activity against 2,7-anhydro-Neu5Ac, the purified recombinant protein was incubated in 100 \( \mu \)l reactions at 37 °C overnight with 1 mM 2,7-anhydro-Neu5Ac, 50 mM sodium phosphate buffer pH 7.0 and 500 mM NADH, NAD\(^+\), NAAD\(^+\), or no cofactor. The reaction was then stopped and protein (500 \( \mu \)g) in the presence and in the absence of TEMPOL (2.5 mM) was added, and the samples were processed using the Topspin v.3.1 compressed sensing routine. The final selected data points using 256 or 512 scans. Owing to the absence of a three-dimensional structure, it was impossible to derive the resonances for saturation of aliphatic and aromatic residues found in the binding site as required by the DEEP-STD NMR technique. Moreover, as RgSBP is a high-molecular-mass protein, NMR structure assignment was precluded. We therefore adopted a search-for-druggable-sites strategy using TEMPO as shown in Supplementary Table 5.

STD NMR spectroscopy. An Amicon centrifuge filter unit with a 10kDa molecular mass cut-off was used to exchange the protein in 25 mM \( d_2\)-2,2-bis(hydroxymethyl)-2′,2′-nitrotrilohethanol pH 7.4 (the pH value was uncorrected for the deuterium isotope effect on the pH glass electrode) \( \text{D}_2\text{O} \) buffer and 50 mM NaCl. 2,7-anhydro-Neu5Ac and Neu5Ac were dissolved in 25 mM \( d_2\)-2,2-bis(hydroxymethyl)-2′,2′-nitrotrilohethanol pH 7.4 and 50 mM NaCl. Characterization of ligand binding by STD NMR spectroscopy was performed by using a Bruker Avance 800.23 MHz at 298 K. The on- and off-resonance spectra were acquired using a train of 50 ms Gaussian selective saturation pulses using a variable saturation time from 0.5 s to 4 s for binding epitope mapping determination, whereas only 0.5 s of saturation time for each selected frequency was used to perform the DEEP-STD NMR experiments\(^{11}\). The water signal was suppressed using the excitation sculpting technique\(^{11}\), whereas the remaining protein resonances were filtered using a \( T\)\(_{1}\) filter of 40 ms. All of the spectra were performed with a spectral width of 10 kHz and 32,768 data points using 256 or 512 scans. Owing to the absence of a three-dimensional structure, it was impossible to derive the resonances for saturation of aliphatic and aromatic residues found in the binding site as required by the DEEP-STD NMR technique. Moreover, as RgSBP is a high-molecular-mass protein, NMR spectra assignment was precluded. We therefore adopted a search-for-druggable-sites strategy using TEMPO as previously described\(^{11}\).

We acquired \( 1^1\)H-1H total correlated spectroscopy (TOCSY) spectra of the protein (500 \( \mu \)M) in the presence of ligand (2.5 mM) and the protein (50 \( \mu \)M) in the presence of ligand (2.5 mM). The spectra were performed with a spectral width of 10 kHz using a time of 2,056 data points in the dimension and 32 scans. The indirect dimension was acquired using a non-uniform sampling technique, acquiring a non-uniform sampling amount of 50% of the original 256 increments and resulting in 64 hypercomplex points. The spectra were processed using the Topspin v.3.1 compressed sensing routine. The final selected resonances were those that were identified by the TEMPO PRE effect, and not overlapping with ligand signals. The DEEP-STD NMR data obtained were used to derive the average orientation of the ligand bound to RgSBP by averaging the DEEP-STD factors obtained from each saturated region. The DEEP-STD NMR chemical binding epitope mapping analysis were performed using previously published procedures\(^{11}\).

Determination of crystal structure. Sitting drop vapour-diffusion crystallization experiments of RgNanA WT were set up at a concentration of 20 mg ml\(^{-1}\) and monitored using the VMXIII beamline at Diamond Light Source\(^{11}\). The described
In vivo colonization and analyses. The impact of the nan deletion mutation on the fitness of R. gnavus was assessed by its ability to colonize germ-free C57BL/6J mice. Groups of 7–9-week-old germ-free mice (two male and two female) were gavaged with 1 × 10⁷ CFU of R. gnavus ATCC 29149 WT or antisense nan mutant in 100 µL PBS, either individually or in combination. The sample size was selected following the 3Rs principles of reduction, replacement, and refinement, while ensuring that data collected allowed for statistical analysis; randomization was not possible owing to the constraints of germ-free isolators. Scientists were blinded for the FISH analysis. Care and treatment of animals was in accordance guidelines from, and approval by the University of East Anglia Disease Modelling Unit, and all of the animal experiments were conducted in strict accordance with the Home Office Animal (Scientific Procedures) Act 1986. Faecal samples were collected from each mouse at 3, 7 and 14 days after gavage, and cecal content was taken at day 14. DNA was extracted from these samples using the MP Biomedicals Fast DNA SPIN kit for Soil DNA extraction with the following modifications. The samples were resuspended in 978 µL of sodium phosphate buffer before being incubated at 4 °C for 1 h following the addition of 122 µL MT Buffer. The samples were then transferred to the lysing tubes and homogenized in a FastPrep Instrument (MP Biomedicals) three times for 40 s at a speed setting of 6.5 with 5 min on ice between each bead-beating step. The protocol was then followed as recommended by the supplier.

Colonization was quantified using qPCR performed using an Applied Biosystems 7500 Real-Time PCR system (Life Technologies). In competition experiments, primers were targeted to the insertions in the nan cluster, and one pair of primers was designed to specifically amplify the inserted DNA and, therefore, target the nan mutant (Supplementary Table 6). The primers were between 18 and 23 nucleotides with a melting temperature (Tm) of 59–60 °C. Standard curves were prepared in triplicate for both primer pairs using a tenfold serial dilution of DNA corresponding to 1 × 10⁶ copies of RgNaH per 2 µL to 1 × 10⁴ copies per 2 µL diluted in 5 mg/mL herring sperm DNA. The standard curves showed a linear relationship of log input DNA versus the Ct, with acceptable values for the slopes and Rs. The dissociation curves were also performed to check the specificity of the primers. Each qPCR reaction (10µL) was then carried out in triplicate with 2 µL of 1 ng/mL DNA (diluted in 5 µg/mL herring sperm DNA) and 0.2 µM of each primer, using the QuantFast SYBR Green PCR kit (Qiagen) according to the manufacturer’s instructions (except that the combined annealing-extension step was extended to 35 s instead of 30 s). Data were analysed using the prepared standard curves.

RNA-seq analysis. For RNA-seq analysis, the colonized mice from monocolonized mice were gently killed and stored in RNAlater at −80 °C until extraction. RNA extraction was performed using the RNeasy mini kit (QIAGEN) following the manufacturer’s instructions for purification of total RNA from animal tissues, including the on-column DNase digestion. Homogenization was achieved with acid-washed glass beads using a FastPrep-24 (MP Biomedicals) by three intermittent runs of 30 s at a speed of 6 ms⁻¹ every 5 min, at room temperature. Elution was performed as recommended with 50 µL Rnase-free water. The quality and concentration of the RNA was assessed using a NanoDrop2 spectrophotometer, a Qubit RNA HS assay using a Qubit 2.0 fluorometer (Life Technologies) and an Agilent RNA 600 Nano kit using an Agilent 2100 Bioanalyzer (Agilent Technologies).

RNA-seq was performed by Novogene. In brief, mRNA was enriched using oligo(dT) beads, fragmented randomly in fragmentation buffer, followed by cDNA synthesis using random hexamers and reverse transcriptase. After first-strand synthesis, a custom second-strand synthesis buffer (Illumina) was added with dNTPs, RNase H and E. coli polymerase I to generate the second strand by nick-translation. The final cDNA library was obtained after a round of purification, terminal repair, A-tailing, ligation of sequencing adapters, size selection and PCR enrichment. Library concentration was first quantified using a Qubit 2.0 fluorometer (Life Technologies) and then diluted to 1 ng/µL before assessing the insert size using an Agilent 2100 Bioanalyzer and quantifying, to greater accuracy, by qPCR (library activity > 2 nM). Sequencing of the library was performed using an Illumina HiSeq platform, and 125 bp and 150 bp paired-end reads were generated.

FASTQ files containing base calls and quality information for all of the reads that passed quality filtering were generated. Reads were mapped to the mouse reference genome using TopHat2 (v. 2.1.0). The mismatch parameter was set to 2, and other parameters were set as default. Appropriate parameters were also set, such as the longest intron length. Filtered reads were used to analyse the mapping status of RNA-seq data to the reference genome. HTSeq was used to analyse the gene expression levels, using the union model. For the gene-expression levels estimated from different genes and experiments to be comparable, the fragments per kb of transcript sequence per million bp sequenced (FPKM) analysis was used to take into account the effects of both subsequent depth and gene length. The differential gene expression analysis was performed using the DESeq package and then adjusted from gene-expression-level analysis as input data. An adjusted P value cut-off of 0.05 was used to determine differentially expressed transcripts.
References

1. Donaldson, G. P., Lee, S. M. & Mazmanian, S. K. Gut biogeography of the bacterial microbiota. *Nat. Rev. Microbiol.* 14, 20–32 (2016).

2. Johansson, M. E. V., Larsson, I. M. H. & Hansson, G. C. The two mucus layers of colon and organized by the MUC2 mucin, whereas the outer layer is a mediator of host-microbial interactions. *Proc. Natl Acad. Sci. USA* 108, 4659–4665 (2011).

3. Etienne-Mesmin, L. et al. Experimental models to study intestinal microbes-mucus interactions in health and disease. *FEBS Microbiol. Rev.* 43, 457–489 (2019).

4. Jensen, P. H., Kolarich, D. & Packer, N. H. Mucin-type O-glycosylation—putting the pieces together. *FEBS J.* 277, 81–94 (2010).

5. Ndeh, D. & Gilbert, H. J. Biochemistry of complex glycan depolymerisation by the gut microbiota. *FEBS Microbiol. Rev.* 42, 164–168 (2018).

6. Tailford, L. E., Crost, E. H., Kavanagh, D. & Juge, N. Mucin glycan foraging in the human gut microbiome. *Front. Genet.* 6, 81 (2015).

7. Martens, E. C., Chiang, H. C. & Gordon, J. I. Mucosal glycan foraging enhances fitness and transmission of a saccharolytic human gut bacterial symbiont. *Cell Host Microbe* 4, 447–458 (2008).

8. Robbe, C., Capon, C., Goddeldev, B. & Michalski, J. C. Structural diversity and specific distribution of O-glycans in normal human mucins along the intestinal tract. *Biochem. J.* 384, 307–316 (2004).

9. Robbe, C. et al. Evidence of regio-specific glycosylation in human mucin mucus—presence of an acidic gradient along the intestinal tract. *J. Biol. Chem.* 278, 46337–46348 (2003).

10. Ouge, N., Tailford, L. & Owen, C. C. Sialidases from gut bacteria: a mini-review. *Biochem. Soc. Trans.* 44, 166–175 (2016).

11. Lewis, A. L. & Lewis, W. G. Host sialoglycans and bacterial sialidases: a convergent pathway for utilization of the amino sugars N-acetylgalactosamine, N-acetylmannosamine, and N-acetylenuraminic acid by *Escherichia coli.* *J. Bacteriol.* 181, 47–54 (1999).

12. Almagro-Moreno, S. & Boyd, F. E. Insights into the evolution of sialic acid catabolism among bacteria. *BMC Evol. Biol.* 9, 118 (2009).

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

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

15. Brigham, C. et al. Sialic acid (N-acetyl neuraminic acid) utilization by *Bacteroides fragilis* requires a novel N-acetylmannosamine epimerase. *J. Bacteriol.* 191, 3629–3638 (2009).

16. Vimr, E. R., Kalvoda, K. A., Deso, E. L. & Steenbergen, S. M. Diversity of microbial sialic acid metabolism. *Microbiol. Mol. Biol. Rev.* 68, 132–153 (2004).

17. Thomas, G. H. Sialic acid acquisition in bacteria—one substrate, many transporters. *Biochem. Soc. Trans.* 44, 760–765 (2016).

18. Mulligan, C. et al. The substrate-binding protein imposes directionality on an electrochemical sodium gradient-driven TRAP transporter. *Proc. Natl Acad. Sci. USA* 106, 1778–1783 (2009).

19. Walsheg, W. Y. et al. Substrate-bound outward-open structure of a Na+-coupled sialic acid symporter reveals a new Na+ site. *Nat. Commun.* 9, 1753 (2018).

20. Severi, E., Hosie, A. H., Hawkhead, J. A. & Thomas, G. H. Characterization of a novel sialic acid transporter of the sodium solute symporter (SSS) family and in vivo comparison with known bacterial sialic acid transporters. *FEBS Microbiol. Lett.* 304, 47–54 (2010).

21. Gangi Setty, T., Cho, C., Govindappa, S., Apicella, M. A. & Ramaswamy, S. Bacterial periplasmic sialic acid-binding proteins exhibit a conserved binding site. *Acta Crystallogr. D* 70, 1801–1811 (2014).

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

23. Muller, A. et al. Conservation of structure and mechanism in primary and secondary transporters exemplified by SiaPQM. N-acytelyneuraminic acid binding virulence factor from *Haemophilus influenzae*. *J. Biol. Chem.* 281, 22212–22222 (2006).

24. Severi, E. et al. Sialic acid transport in *Haemophilus influenzae* is essential for lipopolysaccharide sialylation and serum resistance and is dependent on a novel tripartite ATP-independent periplasmic transporter. *Microbiol. Mol. Biol.* 58, 1173–1185 (2005).

25. Post, D. M., Mungur, R., Gibson, B. W. & Munson, R. S. Ir. Identification of a novel sialic acid transporter in *Haemophilus ducreyi*. *Infect. Immun.* 73, 6727–6735 (2005).

26. North, R. A. et al. Structure and inhibition of *N*-acytelyneuraminic acid lyase from *methicillin-resistant Staphylococcus aureus*. *FEBS Lett.* 590, 4414–4428 (2016).

27. Timms, N. et al. Structural insights into the recovery of aldolase activity in *Escherichia coli* by the human gut microbiome. *Biochem. J.* 483, 457–464 (2011).

28. Huynh, N. et al. Structural basis for substrate specificity and mechanism of *N*-acytelyneuraminic acid lyase from *Pseudomonas maltophilia*.*Biochemistry* 52, 8570–8579 (2013).

29. Barbosa, J. A. et al. Active site modulation in the *N*-acytelyneuraminic acid lyase sub-family as revealed by the structure of the inhibitor-complexed *Haemophilus influenzae* enzyme. *J. Mol. Biol.* 303, 405–421 (2000).

30. Danides, A. D. et al. Reaction mechanism of *N*-acytelyneuraminic acid lyase revealed by a combination of crystallography, QM/MM simulation, and mutation. *ACS Chem. Biol.* 9, 1025–1032 (2014).

31. Huynh, N. et al. The ClosTron: mutagenesis in *Clostridium* refined and streamlined. *J. Microbiol. Methods* 80, 49–55 (2010).
81. Emsley, P. Tools for ligand validation in Coot. Acta Crystallogr. D 73, 203–210 (2017).
82. Smart, O. S. et al. Exploiting structure similarity in refinement: automated NCS and target-structure restraints in BUSTER. Acta Crystallogr. D 68, 368–380 (2012).
83. Langer, G., Cohen, S. X., Lamzin, V. S. & Perrakis, A. Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. Nat. Protoc. 3, 1171–1179 (2008).
84. Winn, M. D., Murshudov, G. N. & Papaz, M. Z. Macromolecular TLS refinement in REFMAC at moderate resolutions. Methods Enzymol. 374, 300–321 (2003).
85. Williams, C. J. et al. MolProbity: more and better reference data for improved all-atom structure validation. Protein Sci. 27, 293–315 (2018).
86. McCoy, A. J. et al. Gyre and gimble: a maximum-likelihood replacement for Patterson correlation refinement. Acta Crystallogr. D 74, 279–289 (2018).
87. Vonrhein, C. et al. Data processing and analysis with the autoPROC toolbox. Acta Crystallogr. D 67, 293–302 (2011).
88. STARANISO (Global Phasing Ltd.; 2018); http://staraniso.globalphasing.org/cgi-bin/staraniso.cgi.
89. Kim, D. et al. TopHat2: accurate alignment of transcripts in the presence of insertions, deletions and gene fusions. Genome Biol. 14, R36 (2013).
90. Anders, S. & Huber, W. Differential expression analysis for sequence count data. Genome Biol. 11, R106 (2010).
91. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. Nat. Methods 9, 676–682 (2012).
## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- **n/a**
- Confirmed

- [X] The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- [X] A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- [X] The statistical test(s) used AND whether they are one- or two-sided
- [X] Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- [X] A description of all covariates tested
- [X] A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- [X] A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- [X] For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
  
  *Give P values as exact values whenever suitable.*

- [X] For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- [X] For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- [X] Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

*Our web collection on statistics for biologists contains articles on many of the points above.*

### Software and code

Policy information about availability of computer code

**Data collection**  
Magellan (V 7.0), Multi-user reader control (BMG Labtech), Malvern PEAQ ITC control software, Aimless (V 0.6.2), dials (V 1.9.2), xia2 (V 0.5.542), pointless (V 1.11.8), phaser (V 2.8.2), EFI-EST(https://efi.igb.illinois.edu/efi-est/), multigene BLAST(http://multigeneblast.sourceforge.net/), All software used is publicly available.

**Data analysis**  
Microsoft Excel, Graph Pad Prism (V 5.03), Cytoscape (V 3.6), Fiji (V 2.9), R (V 3.5.3), Topspin (V 3.1), TopHat2, MARS Data analysis software (BMG Labtech), Malvern PEAQ ITC analysis software, PDB_extract (V 3.25), buster (V 2.10.3), xds, refmac (V 5.8.0238), molrep (V 11.6.04), staraniso (V 1.7.2), CCP4 (V 7.0.052), xtriage, , pymol (V 1.7.4.1), ARP/wARP (V 8.0), PDB redo (V 7.25), coot (V 0.8.9.1), all software used is publicly available. Computer code for statistical analysis is available on request.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

### Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Genome and protein sequences are available from NCBI and referenced within the text or supplementary information. RNAseq data is available at XXX. All other data are available in the Supplementary Information or upon request from the corresponding author.
**Field-specific reporting**

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- **Life sciences**
- **Behavioural & social sciences**
- **Ecological, evolutionary & environmental sciences**

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](http://nature.com/documents/nr-reporting-summary-flat.pdf)

---

**Life sciences study design**

All studies must disclose on these points even when the disclosure is negative.

- **Sample size**: Sample size for the in vivo study was selected following the 3 R’s principles, or reduction, replacement, refinement, whilst ensuring sample size (number of mice/cage) allowed for meaningful statistical analysis.

- **Data exclusions**: No data was excluded.

- **Replication**: Both biological and technical replicates are included in the data with the exception of the mouse in vivo study which was performed once but with sample size allowing statistical analysis (and replicates of downstream analyses) and of the in vitro oxidoreductase kinetics which were performed in technical replicates but not biological replicates due to limitation in substrate availability. All attempts at replication were successful and presented in the data.

- **Randomization**: For the germ free experiments, mice were split into four groups, and each group allocated to an isolator. The treatments given to the mice were determined by which isolator they were housed in. Therefore, randomization of treatments was not possible for this experiment.

- **Blinding**: The scientists were blinded for the FISH analysis of the in vivo study. Blinding was not relevant for other experiments as the way data was obtained is not subjective.

---

**Reporting for specific materials, systems and methods**

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

- **n/a**
- **Involved in the study**
  - Antibodies
  - Eukaryotic cell lines
  - Palaeontology
  - Animals and other organisms
  - Human research participants
  - Clinical data

### Methods

- **n/a**
- **Involved in the study**
  - ChIP-seq
  - Flow cytometry
  - MRI-based neuroimaging

---

**Antibodies**

- **Antibodies used**: Rabbit Mucin 2 antibody H-300 from Santa Cruz (Dallas/TX, US, SC-15334) Lot K2912. Dilution 1:100

- **Validation**: Application statement from antibody data sheet: Mucin 2 (H-300) is recommended for detection of Mucin 2 of mouse, rat and human origin by immunofluorescence and immunohistochemistry (including paraffin-embedded sections) (starting dilution 1:50, dilution range 1:50-1:500). Positive Controls: Normal human colon tissue. The website includes 39 referenced works that utilized the antibody [https://www.scbt.com/scbt/product/mucin-2-antibody-h-300](https://www.scbt.com/scbt/product/mucin-2-antibody-h-300)

---

**Animals and other organisms**

- **Policy information about studies involving animals**: ARRIVE guidelines recommended for reporting animal research

- **Laboratory animals**: 12 gnotobiotic C57BL/6J, 6 male, 6 female, 7 - 9 weeks old

- **Wild animals**: the study did not involve wild animals

- **Field-collected samples**: the study did not involve field samples
Ethics oversight

University of East Anglia Disease Modeling Unit

Note that full information on the approval of the study protocol must also be provided in the manuscript.