Spatially distinct physiology of *Bacteroides fragilis* within the proximal colon of gnotobiotic mice

Gregory P. Donaldson1,5, Wen-Chi Chou2,5, Abigail L. Manson2, Peter Rogov2, Thomas Abeel2,3, James Bochicchio2, Dawn Ciulla2, Alexandre Melnikov2, Peter B. Ernst4, Hiutung Chu4, Georgia Giannoukos2, Ashlee M. Earl2,5 and Sarkis K. Mazmanian2,5

A complex microbiota inhabits various microenvironments of the gut, with some symbiotic bacteria having evolved traits to invade the epithelial mucus layer and reside deep within the intestinal tissue of animals. Whether these distinct bacterial communities across gut biogeographies exhibit divergent behaviours is largely unknown. Global transcriptomic analysis to investigate microbial physiology in specific mucosal niches has been hampered technically by an overabundance of host RNA. Here, we employed hybrid selection RNA sequencing (hsRNA-Seq) to enable detailed spatial transcriptomic profiling of a prominent human commensal as it colonizes the colonic lumen, mucus or epithelial tissue of mice. Compared to conventional RNA-Seq, hsRNA-Seq increased reads mapping to the *Bacteroides fragilis* genome by 48- and 154-fold in mucus and tissue, respectively, allowing for high-fidelity comparisons across biogeographic sites. Near the epithelium, *B. fragilis* upregulated numerous genes involved in protein synthesis, indicating that bacteria inhabiting the mucosal niche are metabolically active. Further, a specific sulfatase (BF3086) and glycosyl hydrolase (BF3134) were highly induced in mucus and tissue compared to bacteria in the lumen. In-frame deletion of these genes impaired in vitro growth on mucus as a carbon source, as well as mucosal colonization of mice. Mutants in either *B. fragilis* gene displayed a fitness defect in competing for colonization against bacterial challenge, revealing the importance of site-specific gene expression for robust host-microbial symbiosis. As a versatile tool, hsRNA-Seq can be deployed to explore the in vivo spatial physiology of numerous bacterial pathogens or commensals.

The mammalian gastrointestinal tract hosts an ecosystem of bacteria, protists, fungi and viruses, which is assembled following birth to establish lifelong symbiosis. The microbial community is important for host digestion and protection from enteric infection (that is, colonization resistance). Gut bacteria inhabit a variety of distinct microhabitats along the longitudinal and cross-sectional axes of the intestines. Along the entire length of the gut, mucus provides a physical barrier that partitions the gut lumen from the intestinal surface. Accordingly, the mucus and lumen compartments of the gut house distinct bacterial communities. Some bacteria, such as segmented filamentous bacteria, adherent *Lactobacillus* and a diverse community of crypt-resident bacteria also directly associate with the gut epithelial surface. Studying the relevance of spatial microbiome structure is a challenge due to the dynamic nature of the gut, the complexity of microbial communities and difficulty in accessing gut tissues. Most current knowledge about the microbiome derives from DNA-based sequence profiling of low-volume, homogenized faecal samples, which probably cannot capture information about bacteria closely associated with the host. While imaging can provide insights into the microbial biogeography of the gut, the functional relevance of localization is often difficult to infer. Approaches that can determine the physiology of individual bacterial species, spatially within specific gut microenvironments, could offer unprecedented insights into host-microbial interactions.

Attempts to investigate bacterial transcriptomes in host tissues or mucus are complicated by the paucity of bacterial RNA present at these sites compared to host RNA. Existing dual RNA sequencing (RNA-Seq) methods address this issue by either sequencing combined host and bacterial RNA to extremely high depths, assisted by host ribosomal RNA depletion, and/or enriching for bacterial cells before lysis, using methods such as fluorescence-activated cell sorting, density-gradient centrifugation or laser capture microdissection. Deep sequencing for in silico separation is impractical for confident assessment of bacterial transcriptomes in host niches where the ratios of bacteria to host RNA are exceedingly small. Bacterial cell enrichment before lysis can overcome the abundance limitation, but this involves greater processing time and conditions that induce bacterial stress responses and biases in RNA expression and degradation. Since typical bacterial messenger RNA has an in vivo half-life of just several minutes, RNA loss may be avoided by an alternative strategy, that is, by separating bacterial transcripts from host transcripts at the molecular level, following rapid nucleic acid extraction. Towards this objective, we used hybrid selection (HS) to enrich bacterial reads from total RNA pools isolated from intact and unprocessed host tissues, then used RNA-Seq to measure the enriched transcriptome, a method we termed ‘hybrid selection RNA sequencing’ (hsRNA-Seq). Following extraction of total RNA from various biogeographies in the gut of mice, reverse-transcribed complementary DNA (cDNA) was enriched for bacterial sequences using biotinylated probes complementary to the entire genome of the target gut commensal bacteria (Fig. 1a). After probe hybridization to bacterial cDNA and capture using biotin, host cDNA was washed away allowing the elution...
of enriched bacterial cDNA for sequencing. Originally developed for resequencing targeted regions of the human genome\textsuperscript{24}, HS has been adapted to enrich eukaryotic pathogen\textsuperscript{25} and RNA virus\textsuperscript{26} genomes in ex vivo clinical DNA samples dominated by human genetic material. In this study, we validated and assessed hsRNA-Seq to uncover in vivo interactions between a bacterial species of the gut microbiome and its mammalian host, with defined spatial resolution. We discovered distinct transcriptomes of the same microorganism depending on its biogeography, and further identified and validated specific genes that aid in mucosal association and colonization resistance. This technique can potentially be applied to any sequenced bacterial species to explore its physiology during symbiotic colonization or pathogen infection in the gut or other body sites.

**Results**

hsRNA-Seq enables spatial bacterial transcriptomics during commensal colonization. To explore the genetic underpinnings of host-microbial associations in vivo, we used a simplified model system of mice monocolonized with *B. fragilis* (three mice originating from different cages). RNA was purified from each sample type using the same protocol. After cDNA synthesis, biotinylated whole-genome baits (probes) were used to select *B. fragilis* cDNA to the exclusion of mouse cDNA. The eluted bacterial cDNA libraries were then sequenced. Percentage of RNA-Seq reads mapping to the *B. fragilis* genome increases with HS (mean and s.e.m., one-sided t-tests, \( n = 3 \), **\( P < 0.01 \)). Distribution of genes by average read coverage (in three animals) shifts dramatically with HS for mucus and tissue transcriptomes (median and interquartile range indicated; 4,306 genes plotted). Normalized gene expression levels with and without HS are highly correlated (Pearson’s \( r \) within microenvironments. Each gene is represented by a single dot. HS increases the number of genes identified in differential expression analyses between various sample sites.
RNA in the mucus and tissue samples, respectively (Fig. 1b). After HS, we observed a reduction in unaligned reads and reads mapping to the mouse genome, with a corresponding increase in reads mapping to the \textit{B. fragilis} genome (Supplementary Table 1); the percentage of total reads mapping to the \textit{B. fragilis} genome increased 48- and 154-fold in the mucus and tissue, respectively (Fig. 1b). This enrichment resulted in a dramatic improvement in \textit{B. fragilis} gene coverage in the mucus and tissue samples (Fig. 1c).

To assess the fidelity and performance of HS, we compared hybrid-selected and non-hybrid-selected transcriptomes from the same samples from monoclonized animals. Within sample sites, normalized gene expression levels with and without HS were highly correlated when analysed in bulk (Fig. 1d) or within individual mice (Extended Data Fig. 2 and Supplementary Table 2), indicating that the method did not globally skew the bacterial transcriptome. Our method also allowed simultaneous study of the host transcriptome, since it remained largely unbiased after HS (Extended Data Fig. 3), although we did not extend this analysis in the current study. In the lumen, which served as a control sample yielding a quality bacterial transcriptome without HS, correlations between the same sample with and without HS were as good or better than correlations between biological replicates (Supplementary Table 2), indicating that the process did not alter the bacterial transcriptome. Including all sample sites, only 20 of >5,000 genes in the \textit{B. fragilis} genome did not enrich as expected following HS. These were mostly short (<200 nucleotides) non-coding RNAs, including some transfer and 5S ribosomal RNAs (Supplementary Table 3). However, the majority of tRNAs were not skewed since 85% of them were enriched as expected (Extended Data Fig. 4). These results demonstrate that hsRNA-Seq provides a significantly accurate transcriptome.

hsRNA-Seq enabled the measurement of otherwise undetectable transcripts. Limited bacterial transcript coverage in mucus and tissue samples without HS led to many genes with near-zero expression levels (555 in mucus and 1,034 in tissue). Applying HS, the number of genes with near-zero expression levels decreased to 159 in mucus and 299 in tissue. HS also substantially improved the correlation of gene expression levels between the lumen and mucus, as well as the lumen and tissue (Extended Data Fig. 5). Importantly, HS dramatically increased the number of genes identified as differentially expressed, suggesting a unique, lumen-specific role for PSG in the physiology of \textit{B. fragilis}. We speculate that the induction of a stress response, genome protection and PSG expression observed in the lumen may collectively prepare \textit{B. fragilis} for survival outside of the gut on shedding.

Genes more highly expressed in mucus and tissue samples provide a glimpse into the biology of \textit{B. fragilis} during close association with the intestinal surface. Although an obligate anaerobe, \textit{B. fragilis} is well equipped to contend with reactive oxygen species (ROS)\(^{34}\), which may allow it to associate with the oxygen epithelial surface of the gut\(^{35}\). Indeed, both subunits of alkyl hydroperoxide reductase (an ROS resistance enzyme), \textit{ahpC} (BF1210 or BF9343_RS05610) and \textit{ahpF} (BF1209 or BF9343_RS05605)\(^{36}\) were induced in the mucus and tissue samples (Fig. 2c,e). The previously defined transcriptomic response of \textit{B. fragilis} to oxygen during growth in laboratory media includes suppression of dozens of genes involved in protein synthesis\(^{33}\). In contrast, we found evidence of increased protein synthesis by \textit{B. fragilis} in the mucus and tissue during colonization of mice. Fourteen tRNAs were upregulated in the mucus and 26 in the tissue compared to the lumen, but not a single tRNA was more highly expressed in the lumen (Fig. 2f and Supplementary Table 6). Ribonuclease P (BF0076 or BF9343_RS00335), the ribozyme that cleaves the precursor RNA on tRNAs to form mature tRNAs, was also upregulated in both mucus and tissue (Fig. 2f). Additionally, several ribosome-related genes were upregulated in tissue, including the 30S and 50S subunits (Fig. 2c). Importantly, none of the genes were outliers that appeared to be affected by the HS process (Supplementary Table 3). Although we did not measure protein synthesis directly, these data suggest that mucus-associated bacteria are metabolically active. We speculate that decreased protein synthesis in the lumen may reflect a relatively nutrient-poor environment, since a starvation response involving the shutdown of protein synthesis in the lumen by \textit{Bacteroides thetaiotaomicron} was recently found to be important for colonization\(^{40}\). In contrast, \textit{B. fragilis} tolerates the oxygenic stress of the epithelium through the deployment of alkyl hydroperoxide reductase while expanding its capacity for protein synthesis in the mucus and tissue.
Fig. 2 | \emph{B. fragilis} gene expression across gut microenvironments. \textbf{a}. Genes differentially expressed between the lumen and mucus (inner circle) and lumen and tissue (outer circle) during monocolonization. The blue and red bars indicate the magnitude of the fold change in gene expression. The squares in the innermost ring indicate genes differentially expressed in both (orange) or only one (grey) of the two comparisons. \textbf{b}, \textbf{c}. STRING\textsuperscript{31} network analysis of genes more highly expressed in the lumen compared to tissue (\textbf{b}) and those more highly expressed in the tissue compared to the lumen (\textbf{c}). The thickness of the connecting lines indicates the confidence in relationships between genes and colours are arbitrary. \textbf{d}–\textbf{f}. Fold change in the expression of individual genes in mucus and tissue with respect to the lumen. The indicated genes are highlighted within each figure: PSG biosynthesis genes and genes with lumen-specific expression patterns (\textbf{d}); genes with tissue-specific expression patterns and all annotated sulfatases (\textbf{e}); tRNAs and the tRNA processing ribozyme, ribonuclease P (\textbf{f}). \textbf{g}–\textbf{l}, RT-qPCR (\(\Delta\Delta Ct\) method normalized to \(gyrB\)) confirms differential expression of genes identified by hsRNA-Seq (grey horizontal line at \(y=1\) represents equal expression to the lumen). Genes more highly expressed in the lumen: BF1252 DNA methyltransferase (\textbf{g}); BF3379 histone (\textbf{h}); polysaccharide G flippase (PSG) (\textbf{i}). Genes more highly expressed in mucus and tissue: alkyl hydroperoxide reductase (\(ahp\)) (\textbf{j}); BF3086 sulfatase (\textbf{k}); BF3134 glycosyl hydrolase (\textbf{l}) (mean and s.e.m., ANOVA with Dunnett’s test, \(n=4\) animals). The fold change between sample sites was quantified within each mouse individually; *\(P<0.05\), **\(P<0.01\), ***\(P<0.001\).
We observed spatially differentiated expression patterns in a small, defined set of individual genes, which we chose to further explore in the remainder of this study. The two, most upregulated genes in both mucus and tissue were BF3086 (or BF9343_RS14795) (Fig. 2c,k), 1 of 17 annotated sulfatases encoded in the B. fragilis genome (Fig. 2e), and BF3134 (or BF9343_RS15035), annotated as a glycosyl hydrolase (Fig. 2e,l). Functional annotation and structural modelling indicated that BF3134 probably encodes a cyclo-maltodextrinase, a member of the glycosyl hydrolase family 13 (Extended Data Fig. 6a), BF3086 probably encodes an acetylglucosamine-6-sulfatase (Extended Data Fig. 6a,b), an enzyme previously implicated in mucin desulfation, and important for mucosal glycan foraging by B. thetaiotaomicron (Extended Data Fig. 7b) upstream of both BF3086 and BF3134 genes. When initially assessed using mucus collected from the colons of germ-free mice and digested with BF3086 and BF3134, we observed a common potential regulatory feature, a 36-base pair (bp) motif (Extended Data Fig. 7b) upstream of both BF3086 and BF3134 that was conserved in 13 other B. fragilis genomes, suggesting that both genes are part of the same regulon. Together, the unique spatial expression patterns of these genes, their conservation across multiple strains of B. fragilis and their consistent expression profile across biological replicates (Extended Data Fig. 8a) motivated us to investigate the function of BF3086 and BF3134 during colonization of mice.

**Discovery of candidate mucosal colonization factors in B. fragilis.** To explore potential mechanisms in B. fragilis colonization of the gut, we generated in-frame deletion mutants in the sulfatase (ABF3086) and glycosyl hydrolase (ABF3134) genes. When initially assessed in minimal media with defined carbon sources, mutant strains exhibited similar growth in minimal glucose (Fig. 3a), several dietary polysaccharides (Extended Data Fig. 8b–d) and pig stomach mucins (Extended Data Fig. 8e) compared to their parent wild-type (WT) B. fragilis. We modelled mucosal growth using mucus collected from the colon of germ-free mice and discovered that WT B. fragilis grew robustly in culture, whereas both mutants exhibited growth defects, as measured by colony-forming units (c.f.u.) per ml (Fig. 3b). While WT and complemented B. fragilis strains reached a stable stationary phase, both ABF3086 and ABF3134 entered a death phase after logarithmic growth on mucus (Fig. 3b). Lack of persistence in the stationary phase may reflect the inability of mutant strains to use less accessible or lower abundance mucosal glycans that remain after log-phase growth, because in minimal glucose all strains entered a rapid death phase (Fig. 3a). Ectopic expression of BF3086 or BF3134 under the control of their native promoters on a plasmid to complement the respective
Fig. 4 | BF3086 and BF3134 promote the robustness of *B. fragilis* colonization. **a**, Germ-free mice were orally gavaged with a 1:1 mixture of $10^8$ c.f.u. each of WT and ΔBF3134 *B. fragilis*. Colonization of each strain was monitored using antibiotic resistance during microbiological plating (geometric mean and 95% confidence interval, two-way ANOVA with Šidák's correction of log-transformed data, $n=4$ animals representative of two independent experiments). **b**, Germ-free mice were orally gavaged with a 1:1 mixture of $10^8$ c.f.u. each of WT and ΔBF3086 *B. fragilis*. Colonization of each strain was monitored using antibiotic resistance during microbiological plating (geometric mean and 95% confidence interval, two-way ANOVA with Šidák's correction of log-transformed data, $n=4$ animals representative of two independent experiments). **c**, Horizontal transmission between pairs of mice that had been monocolonized for four weeks with either WT *B. fragilis* (WT initial) or ΔBF3134 (BF3134 initial). The levels of initial and invading strains are shown two weeks after cohousing and separating the mice (mean and s.e.m., two-way ANOVA with Tukey's test of log-transformed data, $n=6$ animals pooled from two independent experiments). **d**, Horizontal transmission with WT- and ΔBF3086-colonized mice (mean and s.e.m., two-way ANOVA with Tukey's test of log-transformed data, $n=6$ animals pooled from two independent experiments). **e**, Colonization by indicated *B. fragilis* strains in mice with a complex microbiome two weeks after a single gavage of $10^8$ c.f.u. **f, g**, Mice were killed to assess colon lumen (f) and colon mucus (g) colonization (mean and s.e.m., ANOVA with Tukey's test, $n=4$ animals representative of two independent experiments; *$P<0.05$, **$P<0.01$, ***$P<0.001$).

mutants fully recovered the mucus growth phenotypes similar to WT levels (Fig. 3b), ruling out polar effects of gene deletions.

To determine whether BF3086 or BF3134 play a role in *B. fragilis* colonization, groups of germ-free mice were associated with WT or mutant bacteria and all strains reached the same stable c.f.u. level in faeces after 4 weeks of monocolonization (Fig. 3c). Spatially, although there were similar bacterial numbers in the colonic lumen (Fig. 3d), both ABF3086 and ΔBF3134 were reduced in the colonization fitness of the colonic mucus (Fig. 3e). This defect was fully recovered by transcomplementation of BF3134, with a trend recovery for BF3086 (Fig. 3e). Impaired mucosal colonization was not due to effects on the expression of genes previously shown to be involved in mucosal colonization: the commensal colonization factors (ccf) or capsular polysaccharides B and C (Extended Data Fig. 8f–h)29. Collectively, these data reveal genes used by *B. fragilis* for growth on mucus in vitro and robust mucosal colonization in vivo.

**B. fragilis** genes for mucosal association enhance competitive colonization. As a stringent test for in vivo function, we determined whether BF3086 or BF3134 contribute to *B. fragilis* fitness in competitive colonization studies. When mice were orally gavaged with an equal mixture of WT and mutant *B. fragilis*, ABF3134 steadily decreased as a proportion of the total population (Fig. 4a). In contrast, ΔBF3086 did not show a competitive disadvantage in this direct competition model (Fig. 4b), possibly because the desulfating activity of WT bacteria liberated enough mucosal glycans to also support the cocolonizing mutants.

Other *B. fragilis* mutants exhibiting defects in mucosal colonization cannot exclude competitors of the same species28,29. In horizontal transmission assays between pairs of mice, animals colonized for four weeks with WT *B. fragilis* (WT initial) displayed robust colonization resistance, while mice initially colonized for four weeks with either ABF3086 or ΔBF3134 were substantially invaded by WT bacteria following cohousing (Fig. 4c,d). This outcome is consistent with a previously proposed model whereby saturation of a mucosal niche prevents invasion by a foreign strain28,29. Accordingly, the model predicts long-term colonization by a single strain of *B. fragilis*, which remarkably has been observed in longitudinal microbiome profiling studies in humans30,44.

*B. fragilis* has been shown to protect from symptoms and gut pathology in multiple preclinical models of colitis, via the induction
of anti-inflammatory interleukin-10 (IL-10) production by Foxp3\(^+\) regulatory T (\(T_{reg}\)) cells\(^{15,16}\). This effect requires delivery to intestinal dendritic cells of \(B.\) \(fragilis\) capsular polysaccharide A\(^{16}\), which was not differentially expressed across microenvironments (Extended Data Fig. 10a). \(B.\) \(fragilis\) strains were tested in the dinitrobenzene sulfonic acid model of experimentally induced colitis. Compared to mice monoclonized with WT \(B.\) \(fragilis\) or BF3086 mutants, those colonized with \(\Delta\)BF3134 exhibited worsened weight loss (Extended Data Fig. 9a), gross inflammation of the tissue (Extended Data Fig. 9b), shorter colon length (Extended Data Fig. 9c) and higher pathology scoring (Extended Data Fig. 9d). Following sham treatment, animal weight and colon length were similar in mice colonized with WT, \(\Delta\)BF3086 or \(\Delta\)BF3134 strains (Extended Data Fig. 10b,c). Both mutants expressed a critical gene for polysaccharide A production at levels equal to or higher than WT bacteria (Extended Data Fig. 9e) and promoted the development of similar amounts of pro-inflammatory IL-17-producing \(T_{reg}\) cells (Extended Data Figs. 9f and 10d,e), but significantly fewer anti-inflammatory IL-10-producing Foxp3\(^+\) \(T_{reg}\) cells (Extended Data Figs. 9g and 10d,e). We were unable to test whether transcomplementation of \(\Delta\)BF3134 restored protection from colitis. These data suggest that expression of BF3134 may assist in polysaccharide A delivery to (or sensing by) the animal host\(^{15,16}\) by positioning the bacteria in closer proximity to surveilling gut immune cells.

To expand these findings beyond gnotobiotic mice, which have poorly developed mucus\(^15\), we gavaged 8-week-old excluded flora mice, which have a complex but controlled microbiota, with 10\(^6\) c.f.u. of \(B.\) \(fragilis\) strains. Two weeks later, we assessed colonization. In the faeces (Fig. 4e) and lumen of the proximal colon (Fig. 4f), \(B.\) \(fragilis\) \(\Delta\)BF3134 colonized at a higher level than BF3086, while neither differed from WT bacteria. Importantly, both mutants colonized the mucus of the proximal colon at significantly lower levels than WT \(B.\) \(fragilis\) (Fig. 4g), indicating that mucosal colonization defects observed in gnotobiotic mice are maintained in the context of a complex microbial community.

Discussion

Despite numerous sequence-based studies exploring the structure of faecal microbial communities across geographies, diets, diseases and lifestyles, investigations into fundamental bacterial functions have largely not accounted for the spatial organization of the microbiome within the gut\(^1\). This is in large part due to technical limitations. In this study, we show that hsRNA-Seq can be used to profile global RNA expression for \(B.\) \(fragilis\) in samples that contain overwhelming amounts of host RNA at various mucosal sites without skewing transcriptomes, thereby providing a glimpse into bacterial physiologies with spatial resolution. Compared to previously published dual RNA-Seq studies of host-associated bacteria\(^{16,18}\), the proportions of bacterial-to-host RNA was lower in this study, by an order of magnitude on average. To measure the transcriptomes reported in this study without using HS would require at least 400 billion additional bp of sequencing per sample. Thus, hsRNA-Seq provides an alternative to ultra-deep sequencing that is preferable when studying host-associated niches where the bacterial load is comparably low. A small number of RNAs did not enrich as expected, most of which were short (<200 nucleotides) non-coding RNAs; therefore, a degree of caution is warranted when interpreting the short transcript data. However, hsRNA-Seq requires only the minimal amount of RNA needed to create a cDNA sequencing library. While we used 2\(\mu\)g of cDNA from a pool of libraries as the input for HS, lower amounts would probably work as well. Given that hsRNA-Seq effectively captured cDNAs across a wide dynamic range and that adjustments could be made to the protocol to accommodate a lower input (for example, increasing the amount of bait, reducing the hybridization volume or increasing the hybridization time), we expect that input amounts could be reduced substantially, giving hsRNA-Seq a further advantage over existing methods.

hsRNA-Seq could be extended to distinguish transcriptomes from individual species within a complex community, potentially enabling the study of low-abundance species behaviours that cannot be resolved from meta-transcriptomes. The level of enrichment of the targeted sequences from polymicrobial communities can potentially be increased with multiple rounds of the HS procedure. Homologous genes in different organisms present a challenge for HS, although synthetic bait designs that consider sequence similarities among cohabiting microorganisms\(^9\) could help overcome this potential issue. hsRNA-Seq could also be applied to other tissues to explore the microbiomes of the oral cavity, lungs, skin and other interfaces of host-microbial symbiosis.

The spatial gene expression patterns revealed in this study provide several insights into the biology of the model gut symbiont \(B.\) \(fragilis\). Previous transcriptomics studies with members of the prominent \(Bacteroides\) genus have revealed sets of genes that are upregulated during in vivo colonization compared to growth in laboratory culture\(^9,50\), during growth in purified mucin glycans\(^17\) or in animals exposed to different diets\(^9,52\). The gene expression profiles of \(B.\) \(theetaitotaomicron\) were investigated in the lumen and mucus of the colon using density-gradient centrifugation to separate bacteria from host cells, and confirmed the upregulation of polysaccharide utilization loci in the mucus layer that were previously shown to be induced by mucins in vitro\(^9\). We found that in the case of \(B.\) \(fragilis\), bacterial populations proximal to tissue display increased metabolic activity, which is perhaps indicative of replication and/or production of microbial molecules that interact with the host. In some cases, genes previously shown to be induced by mucins in vitro and highly expressed in vivo compared to growth in laboratory culture (such as don\(^+\) and cef\(^15,53\)) did not exhibit spatial expression patterns, either because the activating signal for these genes is present in the lumen or they are suppressed in culture. During colon mucus and tissue colonization, we discovered that \(B.\) \(fragilis\) upregulates a set of candidate colonization factors, including genes encoding a sulfatase (BF3086) and a glycosyl hydrolase (BF3134). BF3086 was previously found to be induced during growth in mucosal O-glycans\(^9,54\), indicating its importance to host mucus degradation. Expression of several other sulfatasas was also induced in the previous in vitro model, while the glycosyl hydrolase BF3134 was not induced, highlighting the importance of in vivo models for the study of gut bacterial physiology. We observed that in vivo mucosal association enabled by these gene products is beneficial to \(B.\) \(fragilis\) through increased colonization robustness, which is consistent with previous reports\(^9,39\). While association with the intestinal surface may be perilous for other bacteria, \(B.\) \(fragilis\) appears well adapted to thrive within the mucosal environment since it exhibits lower stress responses, increases protein synthesis while tolerating oxygen and performs poorly in competitive colonization assays without genes enabling robust host association. We conclude that \(B.\) \(fragilis\) has evolved to dynamically modulate its behaviour at distinct sites within the gut, deploying a specific genetic programme during intimate association with its host. This concept and the tools to determine spatial transcriptomes in the gut can be applied to study the many bacterial species, both commensal and pathogenic, that associate with the mucosal surfaces of animals\(^1\).

Methods

Bacterial strains and media. \(B.\) \(fragilis\) NCTC 9343 was cultured in Brain Heart Infusion (BD) with 5\(\mu\)g ml\(^{-1}\) hemin (Frontier Scientific) and 5\(\mu\)g ml\(^{-1}\) vitamin K1 (Sigma-Aldrich) or in a defined minimal medium\(^9,55\) in an 80% nitrogen, 10% carbon dioxide and 10% hydrogen atmosphere. For growth in mouse mucus, crude mucus was isolated as described in the Separation of colon lumen, mucus, and tissue section from the entire colon of germ-free mice into defined minimal media at a concentration of one whole colon of mucus per 5 ml of final medium. Where appropriate, 200\(\mu\)g ml\(^{-1}\) gentamycin, 10\(\mu\)g ml\(^{-1}\) erythromycin, 2\(\mu\)g ml\(^{-1}\) tetracycline
or 10 µg ml⁻¹ chloramphenicol were used in selective media. For competitive colonization and horizontal transmission, the marker plasmids pFD340-Chlor (providing resistance to erythromycin and chloramphenicol) or pFD340-Tet (providing resistance to erythromycin and tetracycline) were used to distinguish two strains, as described previously.15,16 Scarce, in-frame deletions of B. fragilis (1.305 bp deleted) and BF3134 (1.686 bp deleted) were made using allelic exchange with the pKNOCK suicide vector17 using a method described previously.18

Briefly, flanking regions were cloned into the pKNOCK backbone, which was then conjugated into B. fragilis using erythromycin selection. Cotuitrates were passaged without erythromycin until they lost resistance (following a second recombination event), and these colonies were screened for loss of the targeted gene using PCR. The mutants were complemented by expressing the full-length gene using PCR. The mutants were complemented by expressing the full-length gene using PCR.

Preparation of whole-transcriptome fragment libraries (pond) for HS. Libraries for barcoding Illumina reads were created as described previously.15 The isolated RNA was first quantified and qualified by Qubit and Agilent Bioanalyzer. The RNA was then fractionated by FastAP (Thermo Fisher Scientific) and linked to barcoded adapters. The fragmented and barcoded RNA was pooled by sample type (lumen, mucus, tissue) to perform ribosomal RNA depletion using the Ribo-Zero Magnetic Gold (Epicentre/Illumina). The cDNA was generated from the RNA through template-switching RT–PCR. After exonuclease I (New England Biolabs) treatment and PCR enrichment, the cDNA was used for HS.

HS probes (bait) construction. The whole-genome bait was generated at the Broad Institute. For input, 3 µg of B. fragilis NCTC 9343 DNA was sheared for 4 min on a Covaris E210 instrument set to duty cycle 5, intensity 5 and 200 cycles per burst. The mode of the resulting fragment size distribution was 250 bp. End repair, addition of a 3′ A base, adapter ligation and reaction clean-up followed the Illumina's Genomic DNA Sample Preparation Kit protocol, except that the adapter consisted of oligonucleotides 5′-TGTTAATCCATAGAATCAGCGCCCGTCGATGATGTGCGCCATCAGT-3′ ('x' refers to an exonuclease I-resistant phosphorothioate linkage) and 5′-[PHOS]GAGCTGATGCGCCACTACGACTACATTG3′. The ligated products were cleaned up (QIAGEN), amplified by 8–12 cycles of PCR on an ABI GeneAmp 9700 thermocycler in Phusion High-Fidelity PCR Master Mix with HF buffer (New England Biolabs) using the PCR forward primer 5′-GGCTCAGGCAG GGCGCATACGCCGCTCATG3′ and reverse primer 5′-GCTGCAAGGCCG CCTGATGCTGCCGCTCATG3′ (ABI). Initial denaturation was 95 °C at 30 s. Each cycle was 10 s at 98 °C, 30 s at 50 °C and 30 s at 68 °C. PCR products were size-selected on a 4% NuSieve 3:1 agarose gel (Lonza) followed by QIAquick gel extraction. To add a 17 promoter, size-selected PCR products were reamplified as outlined earlier using the forward primer 5′-GGATTCAATATACGACTCATATAG GGCGCTACGGCGCCGACATCCGAC-3′ and reverse primer 5′-CGCTACGGCGCCGACATCCGAC-3′. The QIAGEN-purified PCR product was used as the template for whole-genome biotinylated RNA bait preparation with the MEGAShortscript T7 Transcription Kit (Invitrogen).

HS. Using the designed baits, hybridization was conducted at 65°C for 66 h with 2 µg of pellet libraries carrying staphylococcus ill and mrsa genomes were performed by Bowtie 2 v.2.2.9 (ref. 63) and STAR v.2.7.3a64, respectively. The fragmented and barcoded RNA was pooled by sample type (lumen, mucus, tissue) to perform ribosomal RNA depletion using the Ribo-Zero Magnetic Gold (Epicentre/Illumina). The cDNA was generated from the RNA through template-switching RT–PCR. After exonuclease I (New England Biolabs) treatment and PCR enrichment, the cDNA was used for HS.

RNA-Seq read processing and mapping. The identifiers for all RNA-Seq experiments are listed in Supplementary Table 1. The RNA-Seq reads were trimmed with a Phred quality score cut-off of 20 by the program fastq_quality_trimmer from the FASTX-Toolkit v.0.0.13 (http://hannonlab.cshl.edu/fastx_toolkit/). Reads shorter than 20 bp after adapter and poly(A) trimming were discarded before mapping. Trimmed RNA-Seq reads were aligned to the B. fragilis NCTC 9343 genome (https://www.ncbi.nlm.nih.gov/nuccore/NC_003228) and the mouse genome (genomic adapter sequences and 500 ng of bait in a volume of 30 µl. After hybridization, captured DNA was pulled down using streptavidin Dynabeads (Invitrogen). Beads were washed once at room temperature for 15 min with 0.5 ml of 1× SSC/0.1% SDS, followed by three 10-min washes at 65°C with 0.5 ml pre-warmed 0.1× SSC/0.1% SDS, resuspending the beads once at each washing step. Hybrid-selected DNA was eluted with 30 µl of 1 M NaOH. After 5 min at room temperature, the beads were pulled down, the supernatant was transferred to a tube containing 70 µl of 1 M Tris·HCl pH 7.5, and the neutralized DNA was desalted and concentrated on a QIAquick MinElute column (QIAGEN) and eluted in 20 µl.

Sequencing. Pooled, indexed samples were sequenced on an Illumina HiSeq 2500 System at the Broad Institute to produce 101-bp paired-end reads. Sequence data have been deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) under accession number PRJNA438372.

Differentiable gene expression analysis. The read counts for each bacterial gene were used to analyze differentiable gene expression using the edger v.3.12.1 package.19 Bacterial genes with more than ten uniquely mapped reads in each of three replicates were considered differentially expressed. The differentially expressed gene expression analysis. Genes with a P < 0.05 in the edger analysis were considered differentially expressed.
of interest in B. fragilis NCTC 9343. Protein-coding genes were predicted with Prodigal v2.6.3 (ref.71) and filtered to remove genes with ≥20% overlap to tRNAs or rRNAs. The gene product names were assigned based on top BLAST hits against the Swiss-Prot protein database (ref.76) and Enzyme Commission databases. Structural modelling was performed using Phyre v2.0 (ref.77).

**Functional enrichment and other statistical analysis.** We identified predicted protein domains within the bacterial genes using Pfam categories (ref.28) as part of the Broad Institute's prokaryotic annotation pipeline (ref.71). To assess functional enrichment of differentially expressed genes, we calculated statistical significance using the hypergeometric function with adjustment for multiple hypothesis testing. \( \gamma/0.05 \) was considered enriched. We calculated Pearson’s correlation coefficients to determine if the expression of genes (in TPM) were comparable between different samples. We used the Wilcoxon signed-rank test to compare the expression of a gene family in two different samples. \( P<0.05 \) was considered significant. STRING analysis of relatedness of differentially expressed genes was performed using the online database (ref.78). Default settings were used: minimum interaction score of 0.4 with interaction sources including text mining, experiments, databases, co-expression, neighbourhood, gene fusion and co-occurrence.

**Comparative genomics and motif scanning of 92 Bacteroides and Parabacteroides genomes.** Using a comparative analysis of 92 diverse genome sequences related to B. fragilis (ref.70), which included 23 Bacteroides and 5 Parabacteroides species, we identified 43 BF3134 orthologues in 43 strains and 117 BF3086 orthologues in 83 strains. We constructed multiple alignments of the nucleotide sequences of these groups of orthologues for BF3086 and BF3134, which we used to calculate pairwise sequence identities to measure conservation levels. We searched the upstream regions of BF3086 and BF3134 in B. fragilis NCTC 9343 for conserved motifs, or potential binding sites for transcription factors, using GLAM2 (ref.78) packaged in MEME Suite v4.12.0 (ref.70). We further examined the presence and conservation of this potential regulatory motif using GLAM2Scan on our set of 92 diverse Bacteroides and Parabacteroides genome sequences (Extended Data Fig. 7) (ref.70). A presence of the motif was defined by having two or fewer mismatches. Of the 16 B. fragilis genomes, we used multiple sequence alignments to confirm that the 3 divergent B. fragilis were missing the predicted motif (Extended Data Fig. 7). Excluding 3 same-patient samples, the B. fragilis genomes containing the motif had a pairwise average nucleotide identity value of 98%, indicating that these strains were members of the same species, but not clonally related. In contrast, pairwise average nucleotide identity values between B. fragilis strains with and without the motif averaged 86%, below the threshold commonly used to describe species (ref.79,80).

**RT-qPCR.** RNA purified as described earlier was used to generate cDNA with the iScript cDNA Synthesis Kit (Bio-Rad) according to the manufacturer’s instructions. RT–PCR reactions with Power SYBR Green Master Mix (Applied Biosystems) were run on an ABI Prism 7500 Fast Real-Time PCR System (Applied Biosystems). Relative quantification was calculated using the \( 2^{-\Delta\Delta Ct} \) method with one control (WT or lumen) sample as the calibrator (set to 1) and DNA gyrase \((\text{gyrB})\) as the reference gene. Ct values were converted to log[2] before statistical testing. In all figures, * \( P<0.05 \), ** \( P<0.01 \) and *** \( P<0.001 \). Details of all statistical comparisons, including effect sizes, confidence intervals and exact \( P \) values are included in Supplementary Table 10.

**Code availability.** The code used in the analysis is available at https://github.com/wenchichou/bugInHost.

Received: 25 July 2018; Accepted: 3 February 2020; Published online: 9 March 2020

**References.**

1. Topping, D. L. & Clifton, P. M. Short-chain fatty acids and human colonic function: roles of resistant starch and nonstarch polysaccharides. *Physiol. Rev.* **81**, 1031–1064 (2001).
2. Buffie, C. G. & Pamer, E. G. Microbiota-mediated colonization resistance against intestinal pathogens. *Nat. Rev. Immunol.* **13**, 790–801 (2013).
3. Donaldson, G. P., Lee, S. M. & Mazmanian, S. K. Gut microbiome of the bacterial microbiota. *Nat. Rev. Microbiol.* 14, 20–32 (2016).
4. Johansson, M. E. V. & Hansson, G. C. Immunological aspects of intestinal mucus and mucins. *Nat. Rev. Immunol.* **16**, 639–646 (2016).
5. Li, H. et al. The outer mucus layer hosts a distinct intestinal microbial niche. *Nat. Commun.* **6**, 8292 (2015).
6. Wang, Y. et al. Regional mucosa-associated microbiota determine physiological expression of TLR2 and TLR4 in murine colon. *PloS ONE* **5**, e13607 (2010).
7. Yasuda, K. et al. Biogeography of the intestinal mucusal and luminal microbiome in the rhesus macaque. *Cell Host Microbe* **17**, 385–391 (2015).
8. Albenberg, L. et al. Correlation between intraluminal oxygen gradient and radial partitioning of intestinal microbiota. *Gastroenterology* **147**, 1053–1063 (2014).
9. Davis, C. P. & Savage, D. C. Habitat, succession, attachment, and morphology of segmented, filamentous microbes indigenous to the murine gastrointestinal tract. *Infect. Immun.* **10**, 948–956 (1974).
10. Ivano, I. et al. Induction of intestinal Th17 cells by segmented filamentous bacteria. *Cell* **139**, 485–498 (2009).
11. Vélez, M. P., De Keersmaecker, S. C. J. & Vanderliden, J. Adherence factors of *Lactobacillus* in the human gastrointestinal tract. *FEMS Microbiol. Lett.* **276**, 140–148 (2007).
12. Savage, D. C. Microbial ecology of the gastrointestinal tract. *Annu. Rev. Microbiol.* **61**, 107–133 (1977).
13. Rowan, F. et al. Bacterial colonization of colonic crypt mucous gel and disease activity in ulcerative colitis. *Annu. Surg.* **252**, 869–875 (2010).
14. Pédron, T. et al. A crypt-specific core microbiota resides in the mouse colon. *mBio* **5**, e0116-12 (2012).
15. Earle, K. A. et al. Quantitative imaging of gut microbiota spatial organization. *Cell Host Microbe* **18**, 478–488 (2015).
16. Mark Welch, J. L., Hasegawa, Y., McNulty, N. P., Gordon, J. I. & Borisy, G. S. Spatial organization of a model 15-member human gut microbiota established in gnotobiotic mice. *Proc. Natl Acad. Sci. USA* **114**, E9105–E9114 (2017).
17. Marsh, J. W., Humphrys, M. S. & Myers, G. S. A. Laboratory methodology for dual RNA-sequencing of bacteria and their host cells in vitro. *Front. Microbiol.* **8**, 1830 (2017).
18. Westerman, A. J., Barquist, L. & Vogel, J. Resolving host–pathogen interactions by dual RNA-seq. *PloS Pathog.* **13**, e1006033 (2017).
19. Humphrys, M. S. et al. Simultaneous transcriptional profiling of bacteria and their host cells. *PloS ONE* **8**, e86957 (2013).
20. Mayrmoser, V. C. et al. The co-transcriptional dynamics of uropathogenic *Escherichia coli*-infected mouse macaques reveals new insights into host–pathogen interactions. *Cell Microbiol.* **17**, 730–746 (2015).
Author contributions
G.P.D. and S.K.M. conceived the study. G.P.D., W.-C.C., G.G., A.M.E. and S.K.M. designed the study. G.P.D. prepared the samples for sequencing and performed the mouse colonization and microbiology experiments. D.C., P.R., J.B., A.M. and G.G. performed the hybrid capture and sequencing experiments. W.-C.C., A.L.M. and T.A. performed the computational analysis. H.C. performed the colitis model and flow cytometry. P.B.E. scored the sections for histology. G.P.D., W.-C.C. and A.L.M. created the figures. A.M.E. and S.K.M. supervised the work. G.P.D., W.-C.C., A.L.M., A.M.E. and S.K.M. wrote the paper. All authors provided input on the paper.

Competing interests
The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41564-020-0683-3.

Supplementary information is available for this paper at https://doi.org/10.1038/s41564-020-0683-3.

Correspondence and requests for materials should be addressed to G.P.D., A.M.E. or S.K.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020
Extended Data Fig. 1 | Intestinal biogeography of *Bacteroides fragilis* during mono-colonization. a, CFU per gram of lumen content and b, CFU per cm of mucus from indicated regions of intestine after 4 weeks of mono-colonization with wild-type *B. fragilis* (mean and standard error, n = 4 animals). c, CFU per sample in lumen, mucus, and tissue samples of the proximal intestine of mice mono-colonized for 4 weeks with wild-type *B. fragilis* (mean and standard error, n = 4 animals). These samples were collected using the same dissection method used to prepare samples for RNA-Seq (Fig. 1a).
Extended Data Fig. 2 | Individual mouse correlation plots to assess hybrid selection performance. Correlation plots for HS vs non-HS in individual mice (3 individual-mouse samples from lumen, 3 from mucus, and 3 from tissue, Pearson's r). Each dot represents a single gene.
Extended Data Fig. 3 | Host gene expression comparisons between samples with and without hybrid selection. Total RNA-Seq reads were mapped to mm10 mouse genome using STAR, and the mapped reads were converted into read counts for each gene by HTSeq. After excluding genes with <10 reads mapping across any sample, the read counts for each sample were normalized by TPM (Transcripts Per Million). Each dot represents a single gene. The average TPM for each gene is shown from non-hybrid selected libraries (x-axis) and hybrid selected libraries (y-axis) (n = 3 animals, Pearson’s r).
Extended Data Fig. 4 | Normalized gene expression levels with and without hybrid selection are highly correlated with few outliers. Each gene is represented by a single dot. The correlation coefficients for lumen, mucus, and tissue are 0.99, 0.96, and 0.98, respectively. Outliers where the difference between the HS and non-HS values is larger than three standard deviations are numbered and listed in Supplementary Table 3. These represent primarily short genes (median length 110 nucleotides), particularly tRNA and 5 s rRNA genes. Short genes (<200 nt) are colored blue, showing that most protein-coding genes are enriched properly.
Extended Data Fig. 5 | Correlation in gene expression between different sample sites was improved with hybrid selection. Each dot represents a single gene with all genes plotted (n = 3 animals, Pearson’s r).
Extended Data Fig. 6 | Structural modeling for genes of interest using Phyre. a, The predicted structure for BF3134, modeled using Phyre, indicated that BF3134 is a likely cyclo-malto-dextrinase, closely related to neopullulanase and maltogenic amylase and a member of glycosyl hydrolase family 13 (96% of the sequence was modeled with 100% confidence to the cyclo-malto-dextrinase template c3edeB, with 42% identity). b, Secondary structure prediction for BF3134 using Phyre. Pfam domain analysis for BF3134 also indicated the presence of an N-terminal cyclo-malto-dextrinase domain (PF09087), a central alpha-amylase domain (glycosyl hydrolase family 13; PF00128), and a C-terminal cyclo-malto-dextrinase domain (PF10438). c, The predicted structure for BF3086 indicated a role as an acetylglucosamine-6-sulfatase (93% of the sequence was modeled with 100% confidence by the single highest scoring template, c5g2va, an N-acetylglucosamine-6-sulfatase, with 51% identity). d, Secondary structure prediction for BF3086. Pfam domain analysis indicated the presence of a sulfatase domain, in addition to a domain of unknown function (DUF4976) downstream of the sulfatase domain. The region aligned by Phyre with the c5g2va template included both the regions encompassed by the Pfam sulfatase domain, as well as the Pfam domain of unknown function (DUF4976).
Extended Data Fig. 7 | BF3086 and BF3134 are conserved and share a potential regulatory motif. a, Phylogeny of 92 Bacteroides and Parabacteroides strains showing the presence of BF3086 and BF3134 orthologues, with horizontal bar graphs indicating the percent protein sequence identity to the studied type strain (NCTC9343, highlighted with red font). The teal box indicates strains that can be confidently assigned to the B. fragilis species (average pairwise ANI between them is 98%, whereas it falls below 95% for the next-closest strains also labeled as B. fragilis). The black squares indicate the presence of the conserved upstream motif (0-2 mismatches), using the GLAM2Scan algorithm. b, Sequence of the conserved motif upstream of both genes. The asterisk (*) at position 18 indicates a position that differs between the upstream regions of the glycosyl hydrolase (BF3086) and the sulfatase (BF3134). The glycosyl hydrolase upstream region has an “A” at this position, whereas the sulfatase upstream region has a deletion at this position.
Extended Data Fig. 8 | Additional *in vitro* and *in vivo* phenotypes of ΔBF3086 and ΔBF3134. 

**a**. BF3086 and BF3134 biological replicates. Fold-change for individual mice indicate consistently induced expression of BF3086 and BF3134 in the mucus and tissue relative to the lumen. 

**b-e**. Growth of individual *B. fragilis* strains in a defined minimal medium with **b**, inulin, **c**, pullulan, **d**, mannan, or **e**, pig mucin (mean and standard error, n = 8 independent cultures). 

**f-h**. Quantitative RT–PCR (ΔΔCt method normalized to gyrB) on fecal samples of mice mono-colonized with indicated strains of *B. fragilis*, assessing the expression of **f**, ccfC (BF3581), **g**, PSB flippase (BF1900), and **h**, PSC flippase (BF1014) (mean and standard error, Tukey ANOVA, n = 4 animals).
Extended Data Fig. 9 | BF3134 is required for B. fragilis protection from experimental colitis. a, Mice were mono-colonized with B. fragilis strains at weaning (3 weeks of age) before inducing DNBS colitis at 7 weeks of age. Body weights of mice were measured every 24 hours and are represented as a percentage of their starting weight on day 0 (Tukey 2-way ANOVA, n = 10, 9, 9, representative of two independent experiments). b, 72 hours after colitis induction, mice were sacrificed and the length of the colon from rectum to the cecal junction was dissected (representative images of 3 colons per group, images normalized to size using rulers and then cropped around the colon) and c, colon length measured (Tukey ANOVA, n = 10, 9, 9). d, Histopathologic scores of whole colons (max 48, mean and interquartile range, Tukey ANOVA, n = 10, 9, 9). e, Quantitative RT-qPCR (ΔΔCt method normalized to gyrB) on fecal samples of mice mono-colonized with indicated strains of B. fragilis, assessing the expression of the PSA flippase (BF1369) (Tukey ANOVA, n = 4 animals). f, Lymphocytes isolated from mesenteric lymph nodes of mono-colonized, DNBS-induced mice were analyzed using flow cytometry. IL-17A-producing T cells quantified as a percent of total CD4+Foxp3+ regulatory T cells (Tukey ANOVA, n = 10, 9, 9 animals). g, IL-10-producing T cells quantified as a percent of total CD4+Foxp3+ regulatory T cells (Tukey ANOVA, n = 10, 9, 9 animals, representative of two independent experiments) (all panels unless noted: mean and standard error, * p < 0.05, ** p < 0.01, *** p < 0.001).
Extended Data Fig. 10 | Control experiments and flow cytometry methods for DNBS colitis. **a**, Quantitative RT-qPCR (ΔΔCt method normalized to gyrB) for PSA flippase (BF1369) in lumen, mucus and tissue samples (mean and standard error, n = 4 animals). Fold-change between sample sites was quantified within each mouse individually. **b**, Mice mono-colonized with indicated strains of *B. fragilis* for one month were treated with 50% ethanol, the vehicle control for DNBS colitis induction. Mice were weighed every 24 hours, graphed as a percentage of their weight at day 0 (Tukey 2-way ANOVA, n = 5, 4, 4). **c**, 72 hours after treatment the mice were sacrificed and the length of the colon was measured from rectum to the cecal junction (Tukey 2-way ANOVA, n = 5, 4, 4). **d**, Example live cell gating for flow cytometry in Extended Data 9f and 9g (representative from two independent experiments with similar results). **e**, Example flow plots (1 from each group) for assessing the proportion of IL-10 and IL-17 positive regulatory T cells, as quantified in Extended Data 9f and 9g (representative from two independent experiments with similar results, mean and standard error in graphs, *p < 0.05*).
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.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

| n/a | Confirmed |
|-----|-----------|
|     | - The exact sample size \((n)\) for each experimental group/condition, given as a discrete number and unit of measurement |
|     | - An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
|     | - The statistical test(s) used AND whether they are one- or two-sided |
|     | - Only common tests should be described solely by name; describe more complex techniques in the Methods section. |
|     | - A description of all covariates tested |
|     | - A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
|     | - A full description of the statistics 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) |
|     | - 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. |
|     | - For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
|     | - For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
|     | - Estimates of effect sizes (e.g. Cohen’s \(d\), Pearson’s \(r\)), indicating how they were calculated |
|     | - Clearly defined error bars |
|     | - State explicitly what error bars represent (e.g. SD, SE, CI) |

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about: availability of computer code

| Data collection |
|-----------------|
| RNAseq data was collected on an Illumina HiSeq2500 |
| Flow cytometry data was collected on a Miltenyi MACSQuant |

| Data analysis |
|---------------|
| Reads were trimmed using FASTX toolkit v.0.0.13 |
| Bacterial read mapping was performed with Bowtie2 2.2.9 |
| Read counts per gene were quantified using bedtools 2.25.0 |
| Differential expression analyses were performed with edgeR 3.12.1 |
| Genomic comparisons were made using the MEME suite 4.12.0 (GLAM2 and GLAM2Scan) |
| Structural modeling was done with Phyre 2.0 |
| Graphing and statistical analysis was performed with GraphPad Prism 7 |
| Flow Cytometry data were collected using MACSQuantify 2.0 |
| Flow Cytometry data were analyzed with FlowJo 10.0.8 |

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 upon request. 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

RNA-Seq and hsRNA-Seq data have been deposited in the NCBI Short Read Archive under a project accession number, PRJNA438372. The B. fragilis NCTC9343 genome used for mapping is available at Genbank GCA_000025985.1. Analysis code is available on Github: https://github.com/wenchichou/bugInHost. All other source data are provided with the paper.

Field-specific reporting

Please select 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/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

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

| Sample size | Sample sizes were based on previous published reports or empirically chosen based on pilot studies because expected effect sizes were unknown. |
| Data exclusions | Data were not excluded. |
| Replication | In vivo experiments were separately performed on at least two cohorts of mice (born months apart). Replication was successful in each case. |
| Randomization | Age-matched mice (within 1 week) were randomly sorted into experimental groups. |
| Blinding | Blinding was used for histological scoring (Fig. 5d). All other experiments were unblinded because the microbiological data collection methods are objective (ie CFU plating). |

Reporting for specific materials, systems and methods

Materials & experimental systems
- n/a Involved in the study
- □ Unique biological materials
- □ Antibodies
- □ Eukaryotic cell lines
- □ Palaeontology
- □ Animals and other organisms
- □ Human research participants

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

Antibodies

| Antibodies used | rat anti-mouse CD4 (clone RM4-5) PE-Cy7 (eBioscience ThermoFisher #25-0042-81) (1:200 dilution) |
| | rat anti-mouse IFN gamma (clone XMG1.2) FITC (eBioscience ThermoFisher #11-7311-41) (1:200 dilution) |
| | rat anti-mouse IL-10 (clone JES5-15E3) PE (eBioscience ThermoFisher #12-7101-81) (1:200 dilution) |
| | rat anti-mouse IL-17A (clone eBio17B7) PerCP-Cy5.5 (eBioscience ThermoFisher #45-7177-82) (1:200 dilution) |
| | rat anti-mouse FOXP3 (clone FJK-16s) APC (eBioscience ThermoFisher #17-5773-82) (1:200 dilution) |
| Validation | We did not keep record of the lot numbers |

We did not keep record of the lot numbers.

Validation

All were validated for flow cytometry using mouse samples by the manufacturer. Further validation details:
CD4: relative expression verification by manufacturer and 80 citations available on website
IFN gamma: cell treatment verification by manufacturer and 140 citations available on website
### Eukaryotic cell lines

**Policy information about cell lines**

| Cell line source(s) | NA |
|---------------------|----|
| Authentication | Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated. |
| Mycoplasma contamination | Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination. |
| Commonly misidentified lines (See ICLAC register) | Name any commonly misidentified cell lines used in the study and provide a rationale for their use. |

### Animals and other organisms

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

| Laboratory animals | Swiss Webster and C57BL/6 mice (Mus musculus) were purchased from Taconic Farms and bred at Caltech. For all experiments female 7-8 week-old germfree mice were used with the following exceptions: Fig. 3d-f (half male, half female), DNBS experiments (3 week-old, just-weaned mice were colonized to allow the induction of colitis at 7 weeks of age). For Fig. 4 mice with full microbiome, 7 week-old excluded flora Swiss Webster mice were purchased from Taconic. |
| Wild animals | No wild animals |
| Field-collected samples | No field-collected samples |

### Flow Cytometry

**Plots**

- The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- All plots are contour plots with outliers or pseudocolor plots.
- A numerical value for number of cells or percentage (with statistics) is provided.

**Methodology**

| Sample preparation | Mesenteric lymph nodes from gnotobiotic C57BL/6 mice were isolated and processed by dissociating tissues through a 70 μm cell strainer (BD Falcon) to generate single cell suspensions. |
| Instrument | Miltenyi MACSQuant 10 |
| Software | The Miltenyi software for the MACSQuant was used for acquisition of data. Flowjo 10.0.8 was used to analyze the data. |
| Cell population abundance | No sorting in this study |
| Gating strategy | A live/dead stain over SSC gate was used to identify live cells followed by a CD4+ gate to identify T cells. |

- Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.