

# Analysis of a phase-variable restriction modification system of the human gut symbiont *Bacteroides fragilis*

Nadav Ben-Assa<sup>1,†</sup>, Michael J. Coyne<sup>2,†</sup>, Alexey Fomenkov<sup>3</sup>, Jonathan Livny<sup>4</sup>, William P. Robins<sup>5</sup>, Maite Muniesa<sup>6</sup>, Vincent Carey<sup>7</sup>, Shaqed Carasso<sup>1</sup>, Tal Gefen<sup>1</sup>, Juan Jofre<sup>6</sup>, Richard J. Roberts<sup>3</sup>, Laurie E. Comstock<sup>2,\*</sup> and Naama Geva-Zatorsky<sup>1,8,\*</sup>

<sup>1</sup>Department of Cell Biology and Cancer Science, Rappaport Faculty of Medicine, Technion – Israel Institute of Technology, Technion Integrated Cancer Center (TICC), Haifa, 3525422 Israel, <sup>2</sup>Division of Infectious Diseases, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA, <sup>3</sup>New England Biolabs, 240, County Rd., Ipswich, MA, USA, <sup>4</sup>Infectious Disease and Microbiome Program, Broad Institute of MIT and Harvard, Cambridge, MA, USA, <sup>5</sup>Department of Microbiology, Harvard Medical School, Boston, 02115, MA, USA, <sup>6</sup>Department of Genetics, Microbiology and Statistics, School of Biology, University of Barcelona, Avda. Diagonal 643 08028 Barcelona Spain, <sup>7</sup>Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA and <sup>8</sup>Canadian Institute for Advanced Research (CIFAR) Azrieli Global Scholar, MaRS Centre, West Tower 661 University Ave., Suite 505 Toronto, ON M5G 1M1, Canada

Received March 26, 2020; Revised September 10, 2020; Editorial Decision September 11, 2020; Accepted October 06, 2020

## ABSTRACT

The genomes of gut Bacteroidales contain numerous invertible regions, many of which contain promoters that dictate phase-variable synthesis of surface molecules such as polysaccharides, fimbriae, and outer surface proteins. Here, we characterize a different type of phase-variable system of *Bacteroides fragilis*, a Type I restriction modification system (R-M). We show that reversible DNA inversions within this R-M locus leads to the generation of eight specificity proteins with distinct recognition sites. *In vitro* grown bacteria have a different proportion of specificity gene combinations at the expression locus than bacteria isolated from the mammalian gut. By creating mutants, each able to produce only one specificity protein from this region, we identified the R-M recognition sites of four of these S-proteins using SMRT sequencing. Transcriptome analysis revealed that the locked specificity mutants, whether grown *in vitro* or isolated from the mammalian gut, have distinct transcriptional profiles, likely creating different phenotypes, one of which was confirmed. Genomic analyses of diverse strains of Bacteroidetes from both host-associated and environ-

mental sources reveal the ubiquity of phase-variable R-M systems in this phylum.

## INTRODUCTION

The human gut is colonized with numerous species and strains of microbes. *Bacteroides* is one of the most abundant bacterial genera in the human gut, with >20 human gut species. These Gram-negative bacteria have many unique properties, including the ability to invert an extensive number of DNA regions throughout their genomes. These DNA inversions include the promoter regions of the numerous capsular polysaccharide biosynthesis loci (1–3), the promoter for an extracellular polysaccharide (EPS) (4), promoters for putative fimbriae regions (5–8) and promoters for S-layer proteins and other outer surface proteins (9–12). Site-specific recombinases of either the serine or tyrosine families mediate these distinct DNA inversions (5,8,9). DNA inversions in the gut Bacteroidales have been best studied in *B. fragilis* NCTC 9343. The complete genome sequence of *B. fragilis* NCTC 9343 revealed a total of 28 invertible regions throughout this organism's genome (13). A few of these invertible regions do not contain promoters, but rather involve the inversion of full or partial genes (13). One region subject to DNA inversions identified by genome sequencing contains genes encoding specificity proteins for a Type I restriction-modification system (R-M) (13). Genetic

\*To whom correspondence should be addressed. Tel: +972 77 8875271; Email: naama\_gz@technion.ac.il  
Correspondence may also be addressed to Laurie E. Comstock. Tel: +1 617 732 5500; Email: lcomstock@rics.bwh.harvard.edu  
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

analyses of this region, the methylated sites, and the resulting phenotypes were not previously performed.

Type I R-M systems are pentameric protein complexes comprised of a specificity protein that dictates the DNA recognition sequence, two methyltransferase subunits, and two restriction enzyme subunits (reviewed (14)). The specificity proteins are composed of a modular architecture with two DNA binding domains, one in each half of the protein, known as target recognition domains (TRD). Each TRD recognizes a distinct DNA sequence. When the complex binds DNA, at the target sequence, it either methylates it or restricts it. The methylation occurs typically at an adenine residue within the recognition sequence creating N6-methyladenine (m6A) of hemimethylated DNA. Restriction occurs on unmethylated DNA, usually at a site distant from the recognition sequence. Each half of the specificity protein recognizes a distinct 3-4 bp sequence separated by approximately 5-7 unspecified nucleotides, making the recognition sequence in most cases non-palindromic.

Type I R-M is a ubiquitous defensive system to combat invasive foreign DNA. Bacteriophages have evolved numerous mechanisms to circumvent restriction by R-M systems, including encoding anti-restriction proteins in their genomes and limiting the number of recognition sites (15). Bacteria in turn continue to evolve systems to counteract phage escape. One such mechanism is to change the site where foreign DNA is restricted by changing the R-M recognition site. The modular nature of Type I specificity proteins allows the protein to change its recognition site by varying the N- and C-terminal TRD. Such domain shuffling and rearrangements of specificity-encoding genes has been demonstrated in numerous bacteria, and, in a few organisms, the shuffling is due to site-specific DNA inversions (11,16–19).

The extensive number of invertible regions in the genomes of gut Bacteroidales suggests that these DNA inversions contribute to fitness, likely by a bet-hedging strategy where distinct bacterial phenotypes arise in a population, leading to intra-strain diversity. In this study, we provide the first characterization of a phase-variable Type I restriction modification system of the gut Bacteroidales. We show that reversible DNA inversions at four different inverted repeats within a single R-M locus leads to the generation of eight distinct specificity genes all of which are present at the expression locus in a population of bacteria resulting in extensive methylome variation within a population. By creating a series of mutants, each of which can synthesize only one distinct specificity protein, we characterized the R-M recognition sites using SMRT sequencing. We also characterized the transcriptome of each of the locked specificity mutants from both *in vitro* and *in vivo* grown bacteria, and demonstrated that phase-variable R-M systems are ubiquitous in diverse members of the Bacteroidetes phylum.

## MATERIALS AND METHODS

All primers used in this study are listed in Supplementary Table S4.

### Bacterial strains and growth conditions

*B. fragilis* NCTC 9343 and isogenic mutants were grown in supplemented basal medium (20) or on supplemented brain

heart infusion (BHIS) plates. Erythromycin was added where appropriate (5 µg/ml). *Escherichia coli* strains were grown in L broth or L agar plates with antibiotics added where appropriate; ampicillin 100 µg/ml, trimethoprim 100 µg/ml, kanamycin 50 µg/ml.

### Expression analysis of specificity gene combinations

To determine if each of the six specificity half-genes could be present at the expression locus, primers were designed that bind the end of each half gene and were coupled in PCR with a forward primer at the end of the methylase encoding gene (BF9343\_1756) (Figure 1). To analyze the relative percentage of expression of each of the eight hybrid specificity genes from a bacterial population, RT-qPCR was performed with primers that amplify across the junctions of each of the eight different specificity gene combinations (Supplementary Table S4). Wild type (WT) bacteria were grown to OD<sub>600</sub> ~0.7, and RNA was isolated using Trizol. qRT-PCR was performed using the KAPA SYBR FAST kit (One step KK4651, KAPA Biosystems). Results were normalized to the housekeeping gene *rpsL* (BF9343\_3902). The  $\Delta\Delta C_t$  method was employed for the specificity fold change tests. For the quantification of the specificities, standard curves of each locked-ON mutants were generated. Samples were normalized to the same Ct number, and each specificity primer was normalized according to its primer efficiency compared to the *rpsL* primer efficiency. Ct levels of the four specificities were compared to get relative percentage of each product.

### Creation of locked specificity gene mutants and complementing plasmids

To create mutants locked with different specificity gene combinations at the expression site, the region was oriented to place each specific pattern in the expression locus *in silico*, and then primers were designed to delete the remaining specificity genes. DNA upstream and downstream of the region to be deleted were PCR amplified, and the products were digested and cloned by three-way ligation into *E. coli* DH5 $\alpha$  pNJR6 (21). The resulting plasmids were conjugally transferred into wild-type *B. fragilis* 9343 using helper plasmid R751 (22), and cointegrates were selected by erythromycin resistance. Double cross-out resolvants were screened by PCR to identify mutants.

Genes used for complementation studies were PCR amplified and cloned into expression vector pFD340 (23). Resulting plasmids were verified by sequencing and transferred to *B. fragilis* by conjugal mating using helper plasmid RK231.

### RNA isolation and RT-qPCR from colonic contents

For RNAseq and RT-qPCR analyzes from bacteria from the mouse colon, germ-free Swiss Webster mice (5-7-weeks) of both sexes were mono-colonized with WT *B. fragilis* or mutants for 10 days. Stool and cecal contents were harvested. These contents were suspended in 1 ml Trizol reagent (ThermoFisher Scientific), transferred to 2 mL FastPrep tubes (MP Biomedicals) containing 0.1 mm Zirconia/Silica beads (BioSpec Products) and bead beaten for 90 s at 10 m/s speed using the FastPrep-24 5G (MP

Biomedicals). After addition of 200  $\mu$ l chloroform, each sample tube was mixed thoroughly by inversion, incubated for 3 min at room temperature, and centrifuged for 15 min at 4°C. The aqueous phase was mixed with an equal volume of 100% ethanol, transferred to a Direct-zol spin plate (Zymo Research), and RNA was extracted according to the Direct-zol protocol (Zymo Research).

Samples were treated by TURBO DNA-free kit (Thermo scientific) in accordance with the manufacturer's instructions. cDNA was synthesized using Verso cDNA synthesis kit (Thermo scientific). RT-qPCR was performed with 20–50 ng of cDNA per well (in a 96-well plate), using the same conditions stated in 'Expression analysis of specificity gene combinations'.

### RNAtag-Seq

Illumina cDNA libraries were generated using a modified version of the RNAtag-seq protocol (24). Briefly, 500 ng–1  $\mu$ g of total RNA was fragmented, depleted of genomic DNA, dephosphorylated, and ligated to DNA adapters carrying 5'-AN<sub>(8)</sub>-3' barcodes of known sequence with a 5' phosphate and a 3' blocking group. Barcoded RNAs were pooled and depleted of rRNA using the RiboZero rRNA depletion kit (Epicentre). Pools of barcoded RNAs were converted to Illumina cDNA libraries in two main steps: (i) reverse transcription of the RNA using a primer designed to the constant region of the barcoded adaptor with the addition of an adapter to the 3' end of the cDNA by template switching using SMARTScribe (Clontech) as described (25); (ii) PCR amplification using primers whose 5' ends target the constant regions of the 3' or 5' adaptors and whose 3' ends contain the full Illumina P5 or P7 sequences. cDNA libraries were sequenced on the Illumina [Nextseq 2500] platform to generate paired end reads.

Sequencing reads from each sample in a pool were demultiplexed based on their associated barcode sequence using custom scripts. Up to one mismatch in the barcode was allowed provided it did not make assignment of the read to a different barcode possible. Barcode sequences were removed from the first read as were terminal G's from the second read that may have been added by SMARTScribe during template switching.

Reads were aligned to the *B. fragilis* NCTC 9343 genome (RefSeq accessions NC\_003228.3 (chromosome) and NC\_006873.1 (plasmid)) using BWA (26) and read counts were assigned to genes and other genomic features using custom scripts. Differential expression analysis was conducted with DESeq2 (27) and edgeR (28). The *in vivo* and *in vitro* heatmaps of *B. fragilis* NCTC 9343 genes found to be differentially expressed (Figure 5) were created using the R package pheatmap (v. 1.0.12 . <https://cran.r-project.org/package=pheatmap>).

### Competitive colonization assays in gnotobiotic mice

Mouse studies were approved by the Harvard Medical Area Standing Committee on Animals, or by the Institutional Animal Care and Use Committee (IACUC), Brigham & Women's Hospital or the Technion's Institutional Animal Care and Use Committee, in accordance with the NRC

guide and comply with all relevant ethical regulations for animal testing and research. Swiss-Webster germ-free mice (4–6 weeks old) were obtained from the Geva-Zatorsky lab germ-free colony at the Technion, or the Harvard Digestive Diseases Center gnotobiotic facility. Mice were kept in individually ventilated cages (ISOcage N System, Tecniplast, Buguggiate, Italy or Optimice) with a maximum of five mice per cage. Both males and females were used and housed separately. Each of the four locked-ON mutants were grown to an OD<sub>600</sub> of 0.4, and mixed at a 1:1:1:1 ratio (actual ratio shown in Figure 3A), and then were either diluted 1:15 in fresh medium and grown to an OD<sub>600</sub> of 0.4 and plated for quantification for the *in vitro* competition, or gavaged into germ-free mice. The feces were collected after 18 h and 10 days, diluted in PBS and plated for single colonies. Strains were differentiated and quantified by PCR. Each reaction was performed with four sets of primers producing different sized amplicons, each corresponding to one of the four specificities. Only samples that showed a single PCR product were counted. A one sample t-test on arcsine-transformed values was used to determine if the ratios of strains in the inoculum differed from those in feces.

### Pac-Bio sequencing and methylation pattern detection

Genomic DNA from WT and the four locked specificity gene mutants were sequenced. SMRTbell template libraries were prepared as previously described (29,30). Briefly, genomic DNA samples were sheared to an average size of ~10–20 kb using the G-tubes protocol (Covaris; Woburn, MA, USA), additionally purified using a PowerClean DNA Clean up kit (MoBio laboratories, Inc., Carsbad, CA, USA), end repaired, and ligated to hairpin adapters. Incompletely formed SMRTbell templates were digested with a combination of Exonuclease III (New England Biolabs; Ipswich, MA, USA) and Exonuclease VII (Affymetrix; Cleveland, OH, USA). Genomic DNA fragments and SMRTbell library qualification and quantification were performed using a Qubit fluorimeter (Invitrogen, Eugene, OR, USA) and 2100 Bioanalyzer (Agilent Technology, Santa Clara, CA, USA). SMRT sequencing was carried out on the PacBio RSII (Pacific Biosciences; Menlo Park, CA, USA) using standard protocols for large insert SMRTbell libraries. Sequencing reads were processed, mapped and assembled on the Pacific Biosciences' SMRT Analysis pipeline (<http://www.pacbiodevnet.com/SMRT-Analysis/Software/SMRT-Pipe>) using the HGAP protocol (31).

Interpulse durations (IPDs) were measured as previously described (32) and processed as described (29) for all pulses aligned to each position in the reference sequence. To identify modified positions, we used Pacific Biosciences' SMRTPortal analysis platform, v. 1.3.1, which uses an *in silico* kinetic reference, and a t-test based kinetic score detection of modified base positions (details are available at [https://www.pacb.com/wp-content/uploads/2015/09/WP\\_Detecting\\_DNA\\_Base\\_Modifications\\_Using\\_SMRT\\_Sequencing.pdf](https://www.pacb.com/wp-content/uploads/2015/09/WP_Detecting_DNA_Base_Modifications_Using_SMRT_Sequencing.pdf)). MTase target sequence motifs were identified by selecting the top one thousand kinetic hits and subjecting a  $\pm 20$  base window around the detected base to MEME-ChIP (33). To measure the extent of methylation

for each motif in a genome, a kinetic score threshold was chosen such that 1% of the detected signals were not assigned to any MTase recognition motifs.

### Methylation site frequency analysis

Enumeration and localization of the four methylation patterns was performed using fuzznuc from the EMBOSS suite of programs, and by custom Perl scripts using regular expressions. Each molecule (the *B. fragilis* NCTC 9343 chromosome (accession no. NC\_003228), the pBF9343 resident plasmid (NC\_006873)) was searched on both strands, allowing for overlapping sites.

### PCR-digestion of polysaccharide promoter regions

A previously described PCR-digestion technique was used to quantify the percentage of *B. fragilis* bacteria with the seven invertible polysaccharide (PS) biosynthesis loci promoters oriented ON or OFF (1,34). Chromosomal DNA was isolated from all specificity gene locked-ON mutants and PCRs were performed using primers that annealed outside the IRs flanking each invertible PS promoter region. Each PS promoter region PCR product was digested with a restriction enzyme that cleaves asymmetrically (34) resulting in four differently sized fragments, two from a promoter in the ON orientation and two from a promoter in the OFF orientation.

### Fluorescence microscopy

Bacteria were grown in basal media to OD<sub>600</sub> of 0.6, washed, and incubated with a rabbit antiserum specific to PSB (1). Cells were washed and mixed with Alexa Fluor 488-labeled goat anti-rabbit IgG antibody (Molecular Probes). Bacteria were washed and imaged using an AxioPlan 2 imaging system (Zeiss) at 600×.

### Flow cytometry

*B. fragilis* was grown to OD<sub>600</sub> = 0.5. 10 μl of the samples were added to 1 ml of PBS, and was washed twice with FACS buffer (2% fetal bovine serum, 1 mM EDTA, 0.1% sodium azide in PBS). PSB-specific antiserum was added and the sample was incubated for 1 h at 4°C. The sample was washed again twice with FACS buffer and incubated with Alexa Fluor 488-labeled goat anti-rabbit IgG antibody (Molecular Probes) for 1 h at 4°C in the dark. The cells were washed with PBS and fixed with 4% of paraformaldehyde in the dark for 1h, following analysis by flow cytometry. Data were collected on LSR Fortessa (BD Biosciences) and analyzed with the 'Kaluzsa' software (Beckman Coulter). Bacteria cells positive for PSB were gated.

### Prevalence of putative phase-variable Type I restriction-modification systems in Bacteroidetes

Our locally maintained collection of bacterial genomic sequences includes 3528 genomes identified by the National Center for Biotechnology Information as belonging to the phylum Bacteroidetes and that also include depositor-supplied or NCBI-generated protein sequence information.

This collection of proteomes was searched using the hmm-search program (v. 3.2.1) and the profile hidden Markov model PF01420.18 (Pfam v. 31) to detect protein sequences potentially encoding a Type I restriction modification DNA specificity domain. To be further considered, the bit score of each initial match had to equal or exceed the gathering threshold score (20.70) included in the HMM model. This initial survey detected one or more such matches encoded by each of 2329 genomes. Further parsing of these initial scan results required that each retained genome had two or more such matches on the same contig (1126 genomes) and that these matches be within 10 kb of each other (955 genomes). Finally, the match data were further parsed by comparing the amino acid sizes of the putative restriction modification DNA specificity proteins, requiring that the largest protein in the retained loci be at least 40% larger than one or more of the smaller proteins in the same loci (686 genomes).

This retained set (Supplementary Table S3) is for illustrative purposes, and likely does not reflect the true count of such loci in Bacteroidetes for several reasons. Examples include: (i) the draft nature of most of the genome assemblies likely resulted in proteins containing the motif which were in reality in close proximity to one another to be contained on different contigs, thus excluding them from further consideration, (ii) the smaller, N-terminal alternative specificity domain regions must have been translated as an open reading frame by the depositor to even appear in the genome's proteome for possible detection, (3) the PFam HMM model was constructed using a seed alignment of 31 protein domains, none of which originated from a Bacteroidetes species, thus using the gathering threshold bit score cut-off might be less applicable to Bacteroidetes proteins, (4) both the ≤10 kb separation defining the necessary proximity of these proteins to one another and the requirement that one or more of the smaller protein sequences in a loci be no >60% of the size of the largest protein in the same loci are arbitrary cut-off choices, and may erroneously include or exclude such loci.

### MinION library preparation and sequencing

The specificity region of BF9343\_1757–1760 was amplified by PCR from a population of bacteria grown *in vitro* and *in vivo* from fecal content as described above. The primers annealed outside the invertible region (Supplementary Table S4). The amplicons were purified by Wizard SV Gel and PCR Clean-Up System (Promega), and measured by nanodrop.

DNA quantity was measured again using Qubit fluorometry (Thermo Fisher Scientific, Waltham, MA, USA). Nanopore sequencing libraries were prepared from 200 fmol purified amplicons using Ligation Sequencing Kit 1D (SQK-LSK109) and PCR-free Native Barcoding Expansion Kit (EXP-NBD104) (Oxford Nanopore Technologies, Oxford, England). The barcoded libraries were loaded and sequenced on the MinION device controlled by MinKNOW software (v.19.12.5) using MinION flow cells (FLO-MIN106D R9.4.1, Oxford Nanopore Technologies, Oxford, England) after quality control runs. The raw data

were base called and demultiplexed by Guppy Basecalling Software (v. 3.3.3+fa743a6).

### MinION data analysis

Adapters and barcodes sequences were removed from the reads using Porechop (v0.2.4, available from <https://github.com/rwick/Porechop>). Reads were oriented using the ‘Preparing Reads for Stranded Mapping’ protocol (Eccles, D.A. (2019). *Protocols.io*, Vol. 2019, pp. Protocol). We aligned the reads to the PCR forward primer using LASTAL (v.1060) (35), and then reverse-complemented the reverse-oriented reads. The reads were combined to an all forward oriented file and cropped to the first 1300 bases using Trimmomatic (v.0.39) (36). The reads were then split according to their alignment to the 1757-57 or 1757-60 5’ half sequences using LASTAL. Reads were mapped to the full sequences with Minimap2 (v.2.17-r941) (37) using the –for-only and asm20 options. Mapped read counts were extracted from the Minimap2 SAM output using SAMtools (v.1.7) (38).

MinION sequence data has been deposited in the NCBI sequence read archive (SRA) under the BioProject accession number: PRJNA648829.

## RESULTS

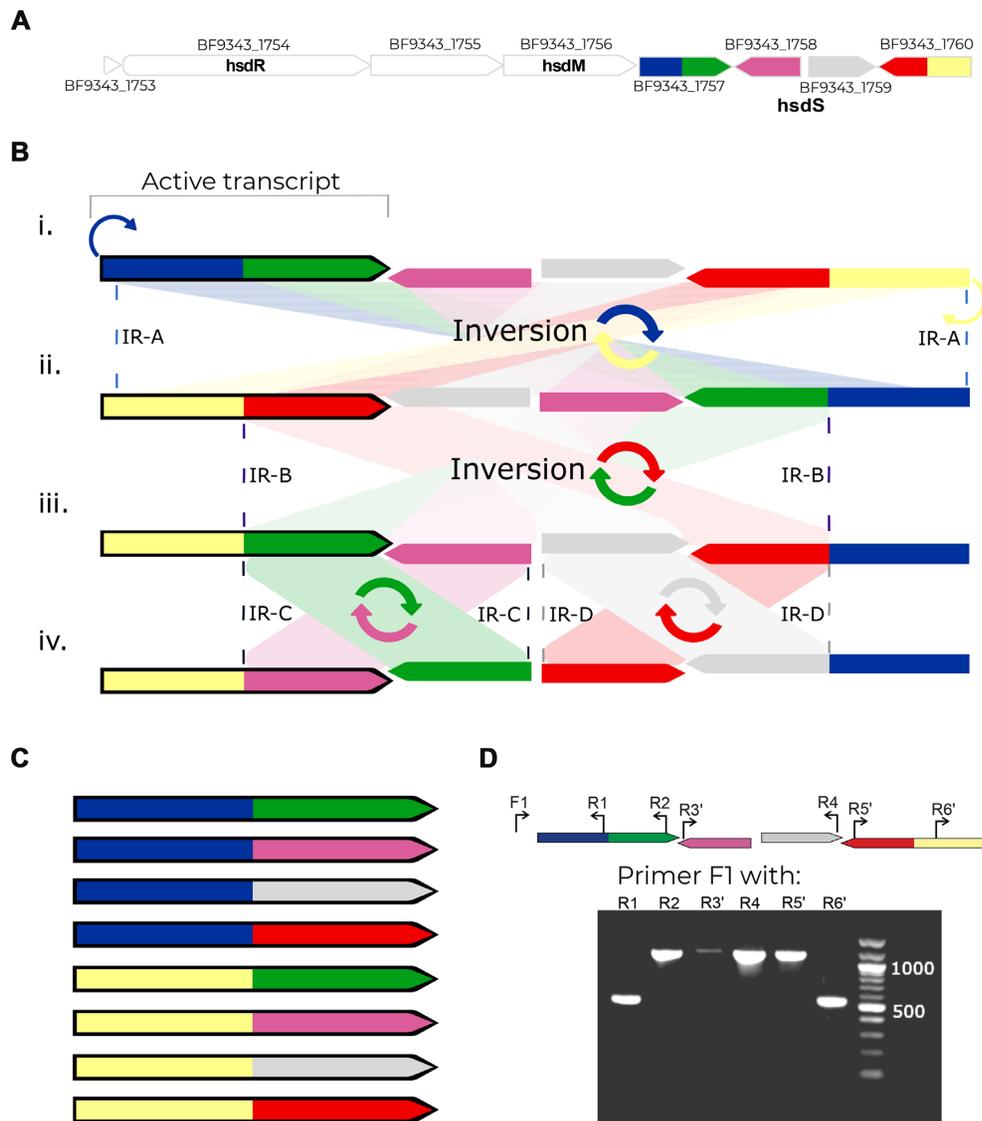
### A phase-variable Type I restriction modification system of *B. fragilis*

A genetic region encoding a Type I R-M system of *B. fragilis* NCTC 9343 includes genes BF9343\_1753 through BF9343\_1760 (Figure 1A). BF9343\_1753 encodes a putative XRE family transcriptional regulator, BF9343\_1754 encodes a Type I restriction enzyme (*hsdR*), BF9343\_1755 encodes a protein with endonuclease motifs, BF9343\_1756 encodes a Type I methyltransferase (*hsdM*), and the next four genes or partial genes, BF9343\_1757 through BF9343\_1760, encode six distinct modules (N-terminal or C-terminal half) of Type I specificity proteins (Figure 1A, B). The genome sequence revealed the presence of four distinct inverted repeats (IRs) of between 16 and 47 bp within this specificity genes region (Figure 1B dashed lines, S1b) (13). The entire region, beginning with BF9343\_1753 through to the specificity gene BF9343\_1757, is a predicted operon with very little intergenic DNA (32 bp or less between genes). Only one specificity gene is actively transcribed - the gene in the expression locus just downstream of *hdsM* (Figure 1B, C, outlined in black). The other genes in the locus either lack the 5’ end of the gene or are not downstream of a promoter (Figure 1B). Based on the genetic architecture, there are two different half-genes that could encode the N-terminal portion of the specificity protein (shown in blue and yellow, Figure 1A, B) and four different half-genes that could invert into the expression locus and comprise the C-terminal portion of the specificity protein (green, pink, gray, and red, Figure 1A,B). An inversion event of IR-A, (Figure 1B, dashed lines), would change the 5’ end of the gene, for example, from blue (BF9343\_1757 5’ half) to yellow (BF9343\_1760 3’ half). Inversions between IR-B, C and D, would each change the 3’ module of the gene in the expression locus. Figure 1B (i), represents the orientation as

present in the published *B. fragilis* NCTC 9343 genome sequence. From this position, in order for the yellow-pink module (BF9343\_1760-1758) to be present at the expression locus, three inversion events are required. First, the IR-A inversion, then the IR-B, and lastly, the IR-C inversion (Figure 1B). This region thus potentially encodes eight different combinations of specificity proteins, depending on the orientation of the DNA (Figure 1C, Supplementary Figure S1). To determine whether each of the six specificity subunit genes are present at the expression locus within a population of *in vitro* grown bacteria, we performed PCR. The forward primer (F1) is located upstream of the invertible regions, at the 3’-end of the methylase-encoding gene BF9343\_1756, while the reverse primers (R1–6), are located at the end of each of the gene halves. Elongation time was short in order to amplify only the gene present at the expression locus (Figure 1B, outlined in black). Hence, only gel bands of ~1200 and ~500 bp are present. These analyses revealed that within this *in vitro* grown population, all specificity half-genes are detected at the expression site (Figure 1D), although the 1758 half-gene (pink in Figure 1B) amplicon is in low abundance (Figure 1D, lane 3).

### Analysis of specificity gene transcription from *in vitro* and *in vivo* grown bacteria

We next sought to quantify the relative percent of cells expressing each specificity gene combination to examine whether certain combinations are more prevalent than others. To do so, we attempted to generate a set of mutants in which each of the eight gene combinations (Figure 1C) would be locked in place by deleting all the other specificity genes. We were successful in creating mutants where the BF9343\_1757 5’ half-gene was locked in the expression locus with each of the four 3’ half-genes (Figure 2A). We made numerous attempts to obtain mutants with the BF9343\_1760 5’ half-gene (yellow 5’ in Figure 1C) locked in the expression locus in combination with any of the 3’ half-genes, but were unable to create these mutants by the methods used. The four locked mutants we did obtain (Figure 2A) allowed us to compare the percent expression from the WT to the locked mutants (100% expression). RT-qPCR analysis of WT *B. fragilis* NCTC 9343 grown to OD<sub>600</sub> ~ 0.7 in rich medium revealed unequal expression of each of these four combinations. The 1757-59 half-gene pairing was at >60% expression level compared to the locked-ON mutant (Figure 2B). The 1757-57 and 1757-60 pairings were in approximate abundance of 10% and 20% relative to their locked-ON counterparts, respectively, and transcript of the 1757-58 specificity gene configuration was barely detectable, a result predicted by the previous PCR analysis (Figure 1D). To confirm these data, the entire specificity region was amplified by PCR from the *in vitro* grown population, and sequenced for quantitative analysis (Figure 2C). The results were similar to the RT-qPCR data (Figure 2C). Unlike the RT-qPCR analysis, this method also revealed the frequency of the 1760 5’ half-gene at the expression locus. Less than 12% of the *in vitro* grown bacteria have the 1760 gene at the expression locus with any of the four C-terminal genes (Supplementary Figure S2a, summation of the four rightmost bars). Therefore, there is an uneven distribution of



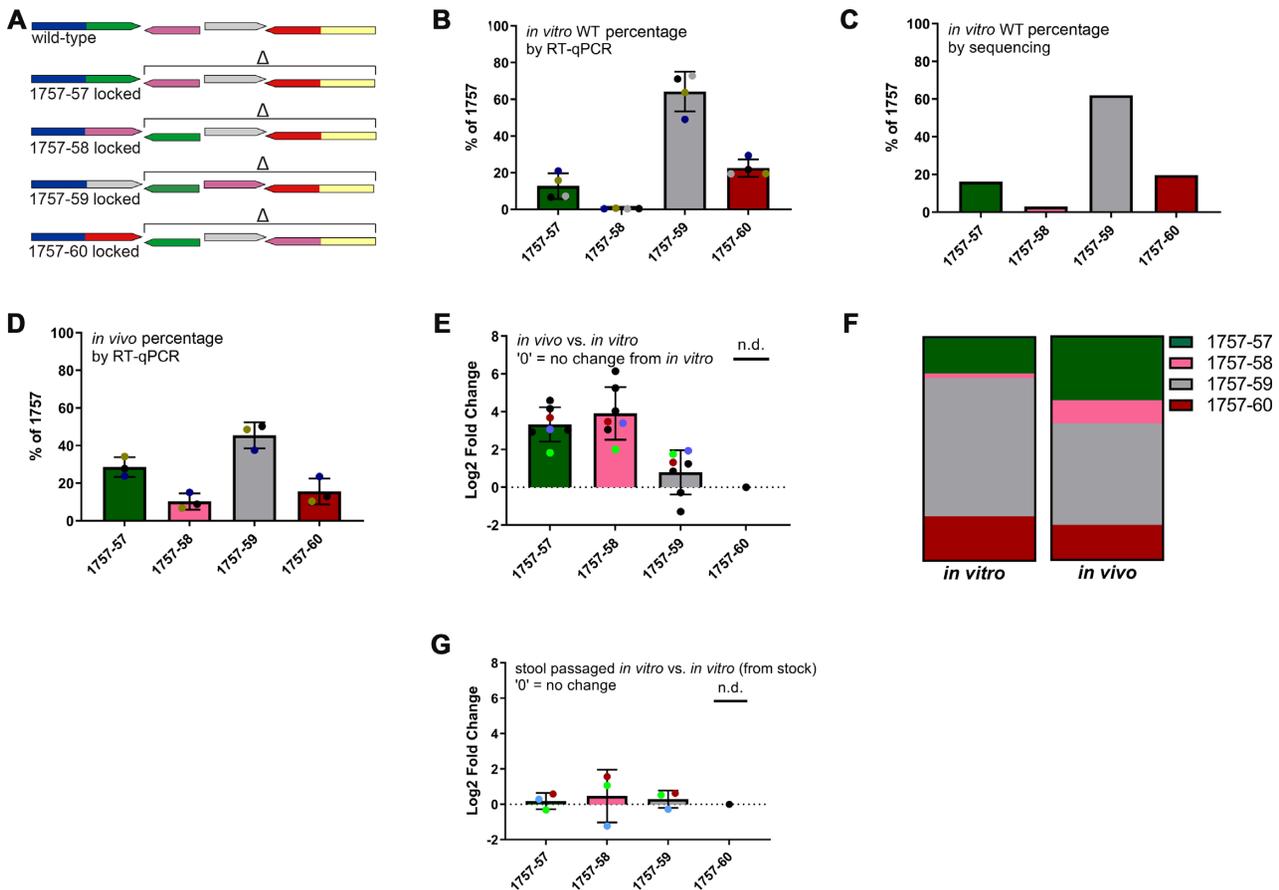
**Figure 1.** A phase-variable Type-I R-M system in *B. fragilis*. (A) Gene map of the region. *hsdR* – restriction enzyme; *hsdM*– methylase; The colored region of BF9343\_1757-60, represents the phase-variable inverting regions of *hsdS* (specificity genes). (B) Genetic regions showing the specificity genes. i-iv show the different DNA inversions that can occur in this region between inverted repeats shown as dashed lines. Only the outlined gene is transcribed. (C) The eight specificity gene combinations possible from the different DNA inversions. (D) EtBr-stained agarose gel of the amplicons detecting each of the six specificity half genes in the expression locus. Each fragment was amplified using primer F1 paired with each of the six reverse primers.

specificity gene combinations present at the expression locus from the population of *in vitro* grown bacteria.

To determine whether the relative percentages of these specificity gene combinations at the expression locus are the same from *in vivo* isolated bacteria, we mono-colonized six sets of two gnotobiotic mice with WT bacteria. Colonic contents were harvested and the transcripts of each of the four combinations were quantified by qRT-PCR (Figure 2D). Compared to their abundances from bacteria grown *in vitro*, there was a significant increase in the fold-change expression levels of 1757-57 ( $P$ -value < 0.0001) and 1757-58 ( $P$ -value < 0.0001) (Figure 2E) ('0' means no fold change from *in vitro* grown bacteria). There was no significant change from the *in vitro* level of the 1757-59 combination and the sequencing data also showed very similar results to

those of the RT-qPCR (Figure 2F). The 1757-60 transcript was not detected in the *in vivo* sample, but showed a slight decrease in the sequencing analysis. There was an increase in the prevalence of the 1760 gene (5' end) at the expression locus compared to *in vitro* grown bacteria (Supplementary Figure S2b), with approximately 22% of the population having this half-gene at the expression locus (summation of the four rightmost bars).

Lastly, we examined whether the percentage of the various S-gene combinations from *in vivo* isolated bacteria revert back to the percentages detected *in vitro* after *in vitro* passage. Three stool samples from three mice of different cages were passaged *in vitro* five times before qRT-PCR analysis. The data showed that these combinations reverted back to the transcriptional profiles of the *in vitro* sample



**Figure 2.** RT-qPCR and sequencing analysis of four of the specificity gene transcripts from WT bacteria grown *in vitro* and *in vivo* (A) Diagram of the deletions made to lock each specificity gene combination at the expression locus. (B, C) Relative percentages of each of the four specificity gene transcripts with the 1757 5' end (blue) from *in vitro* grown WT bacteria by RT-qPCR (B) or by sequencing (C). (D) Relative percentages of each of the four specificity gene transcripts (with the 1757 5' end) from stool of mice monocolonized with WT bacteria. (E) RT-qPCR fold-change of specificity gene expression from bacteria isolated from *in vivo* stool compared to bacteria grown *in vitro*. (F) Abundance of each specificity gene at the expression locus *in vitro* compared to *in vivo* by sequencing. (G) Fold-change expression of specificity genes from bacteria isolated from *in vivo* stool and passed five times compared to bacteria grown *in vitro*.

(Figure 2G, '0' means no change compared to *in vitro* bacteria from stock).

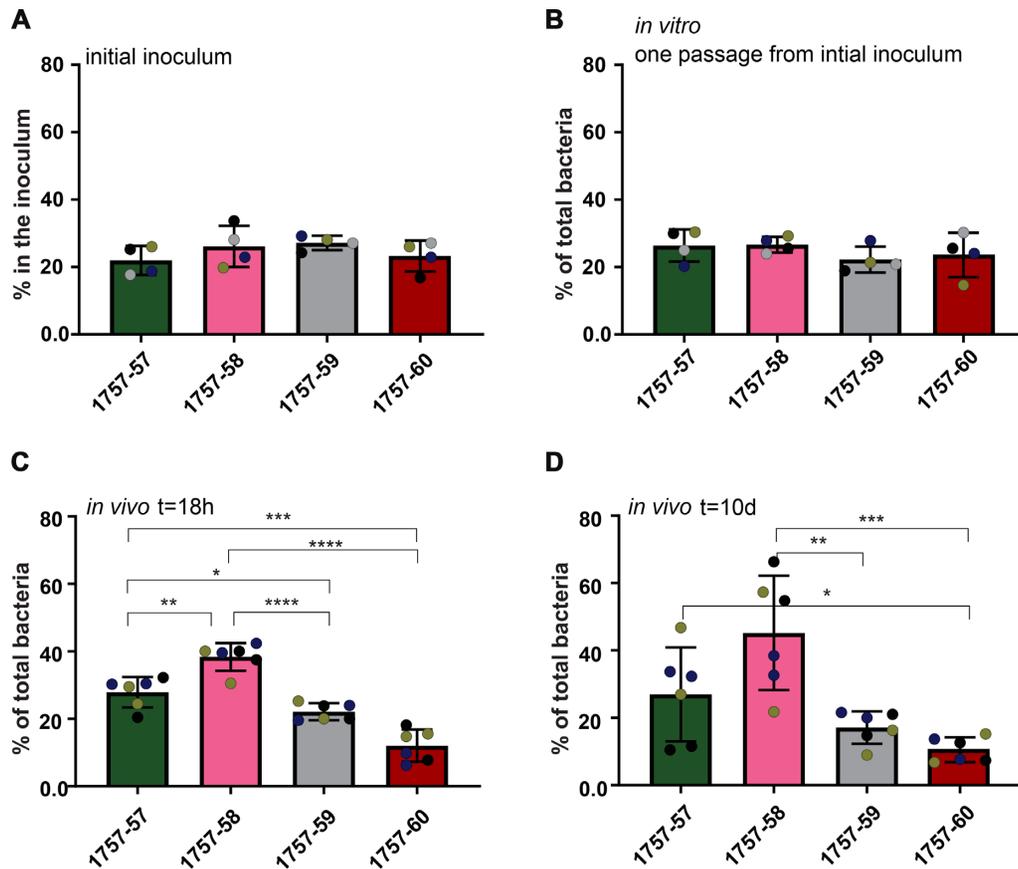
### *In vitro* and *in vivo* competition assays

The WT population showed an uneven distribution of specificity gene expression both *in vitro* and *in vivo*. We speculated that the mutants locked-ON for the specificity genes that were up-regulated *in vivo* may have a competitive advantage in the mammalian gut. Therefore, we conducted a competition assay *in vitro* and in the mammalian gut. A 1:1:1:1 ratio of each of the four mutants was mixed and grown in rich medium or gavage into gnotobiotic mice (Figure 3A). In the *in vitro* assay, none of the mutants demonstrated a competitive advantage over the others (Figure 3B). For the mouse experiments, stool samples were plated at two time points: 18 h after gavage (Figure 3C) and 10 days after gavage (Figure 3D) and analyzed to determine the percentage of each mutant. Eighteen hours after gavage, there was a statistically significant difference in the percentages of the mutants in the stool, and this difference was maintained at day 10. The 1757-58 mutant, which showed

the greatest upregulation *in vivo* compared to *in vitro*, had a competitive advantage over the 1757-59 and 1757-60 mutants (*P*-values 0.0029 and 0.006).

### Methylation site analysis

To determine the DNA sequences that are recognized by each specificity subunit combination and the methylated base, the four mutants were sequenced using SMRT sequencing. Each of the four mutants derives the 5' portion of the specificity gene from 1757, thus all four mutants recognize the same 5' sequence, which we determined was GAC. The 5' GAC is followed by five or six Ns and the 3' recognition site is variable between mutants and contains either three or four bp for a total of 11–12 bp including the intervening Ns (Table 1). The recognition site of the 3' BF9343\_1759 protein is GRTY, therefore, four different tetranucleotide sites are recognized by this half protein (GATC, GATT, GGTC, and GGTT), with the methylation occurring on the A of the complementary strand. The recognition sites for the other 3' half proteins are each a specific three nucleotide sequence TCC, CTG or TGC. All



**Figure 3.** Competitive growth and colonization assay among the four locked-ON mutants (A) the percent of each of the four mutants in the initial inoculum (B) percentages of each mutant after dilution of 1:15 and regrowth *in vitro*. (C) percentages of each mutant from the fecal samples of mice 18 h post-gavage, (D) or after 10 days of competitive competition in gnotobiotic mice. Three sets of two mice were used for the competitive colonization assay and are differentiated by colored circles. *P*-values that were calculated are marked as follows: \**P*-value <0.05, \*\**P*-value <0.005, \*\*\**P*-value <0.0005, \*\*\*\**P*-value < 0.0001.

half recognition sites with the exception of GRTY have  $\frac{2}{3}$  of the nucleotides G or C. The number of methylation sites in the *B. fragilis* NCTC 9343 genome for each of these specificity proteins ranges from 1408 to 1931, with a total of 6479 sites for all four specificity proteins of this R-M system. The number of these four recognition sites in each gene or intergenic region for the entire genome of *B. fragilis* NCTC 9343 is shown in Supplementary Table S1.

Per kilobase, the number of sites in coding regions is greater than in intergenic regions with 1.30 total sites/kb in coding regions and 0.67 sites/kb in intergenic regions (Table 1). This difference likely reflects the %GC in these recognition sites and the lower %GC in intergenic regions compared to coding regions. We found that tRNA genes contain a disproportionate number of recognition sites per kilobase with a total of 21 sites reflecting 3.8 sites/kb, which may be due in part to the high GC content of the 73 annotated tRNA genes (54.3 %GC, on average). (Table 1). As R-M systems are used for bacterial defense to restrict incoming DNA, we analyzed the number of sites found in DNA horizontally acquired by *B. fragilis* NCTC 9343. To this end, we analyzed three elements annotated as integrative conjugative elements and one region annotated as an integrated phage. In these four combined regions, there are a total of

126 methylation sites, with a per kb average of 1.84, with these regions collectively having a %GC content just slightly greater than the overall genome (46.6% versus 43.2%).

### RNA-Seq analyses from *in vitro* and *in vivo* grown WT and specificity gene locked mutants

DNA methylation can have a profound effect on gene transcription, both up-regulating and down-regulating transcript levels in bacteria (39–41). To study the potential transcriptional effects of the different methylation patterns in the locked mutants, we performed RNA-Seq analysis. We analyzed biological duplicates of WT bacteria and all four locked mutants from bacteria grown *in vitro* and from cecal contents of monocolonized mice. Using DESeq2 (27) and EdgeR (28), we identified 220 genes differentially expressed (>2-fold, *P*-value < 0.05) in at least one of the four mutants compared to WT for *in vitro* grown bacteria, including 136 and 84 genes upregulated and down-regulated in the mutants, respectively (Supplementary Table S2). For bacteria from the mouse cecum, 332 genes were differentially regulated in at least one mutant compared to WT, with 236 genes upregulated and 96 genes down-regulated. Under *in vitro* conditions, mutant 1757-60 exhibited the greatest number of differentially regulated genes compared to WT, while un-

**Table 1.** Analysis of specific methylation sites in *B. fragilis* NCTC 9343 genome<sup>a</sup>

Configuration <sup>b</sup>	BF9343_1757–1757	BF9343_1757–1758	BF9343_1757–1759 <sup>c</sup>	BF9343_1757–1760	Totals
Pattern	(GACN6TCC)	(GACN5CTG)	(GACN5GRTY)	(GACN6TGC)	
Unique sites <sup>d</sup>	1931 (958, 973)	1408 (723, 685)	1636 (802, 820, 14)	1504 (776, 728)	6479
Coding gene <sup>e</sup>	1776 (791, 985)	1306 (815, 491)	1513 (929, 572, 12)	1410 (885, 525)	6005
Intergenic	107	62	85 (total, 2 palindromic)	75	329
rRNA	18 (6, 12)	18 (6, 12)	12 (12, 0, 0)	6 (0, 6)	54
tRNA	5 (3, 2)	1 (0, 1)	15 (3, 12, 0)	0 (0, 0)	21
ncRNA	0	0	0	0	0
tmRNA	0	0	0	0	0
Pseudo genes	28 (13, 15)	23 (15, 8)	14 (6, 8, 0)	14 (9, 5)	79
Integrated phage	12	10	15	5	42
ICE	18	16	28	22	84
Total mapped sites <sup>f</sup>	1934 (813, 1014, 107)	1410 (836, 512, 62)	1639 (956, 598, 14, 85)	1505 (894, 536, 75)	6488
coding gene per kb	0.381	0.281	0.325	0.303	1.290
Intergenic per kb	0.217	0.126	0.173	0.152	0.668
rRNA per kb	0.657	0.657	0.438	0.219	1.971
tRNA per kb	0.904	0.181	2.713	0.000	3.798
pseudo per kb	0.397	0.326	0.199	0.199	1.121
int. phage per kb	0.249	0.207	0.311	0.104	0.871
ICE per kb	0.208	0.185	0.324	0.255	0.972

<sup>a</sup>The genome consists of a 5205140 bp chromosome (NC\_003228.3) and a 36 560 bp plasmid (NC\_006873.1).

<sup>b</sup>The configuration is named based on which S subunit is active.

<sup>c</sup>Note that BF9343\_1757-1759 methylation pattern can be palindromic.

<sup>d</sup>Reported as total (forward strand, reverse strand, both strands [BF9343\_1757-1759 only]).

<sup>e</sup>All genes (CDS, RNA) and pseudo genes are reported as total (coding strand, non-coding strand, both strands [BF9343\_1757-1759 only]).

<sup>f</sup>Total sites mapped exceeds unique sites due to overlapping DNA. Reported as total (coding strand, non-coding strand, both strands [BF9343\_1757-1759 only], intergenic).

der *in vivo* conditions, mutant 1757-59 displayed the greatest number of both upregulated and down-regulated genes. Figure 4 shows heat maps of the 50 most upregulated and down-regulated genes for the composite four mutants under each growth condition. Notable is the presence of capsular polysaccharide biosynthesis loci, of which there are many in the *B. fragilis* NCTC 9343 genome (1). Each of these polysaccharide biosynthesis regions is organized as an operon, and for seven of these regions, the promoter driving PS synthesis is contained between IRs and undergoes inversion, leading to phase-variable expression of the operons (1).

### Genetic and phenotypic analyses of specificity gene locked mutants

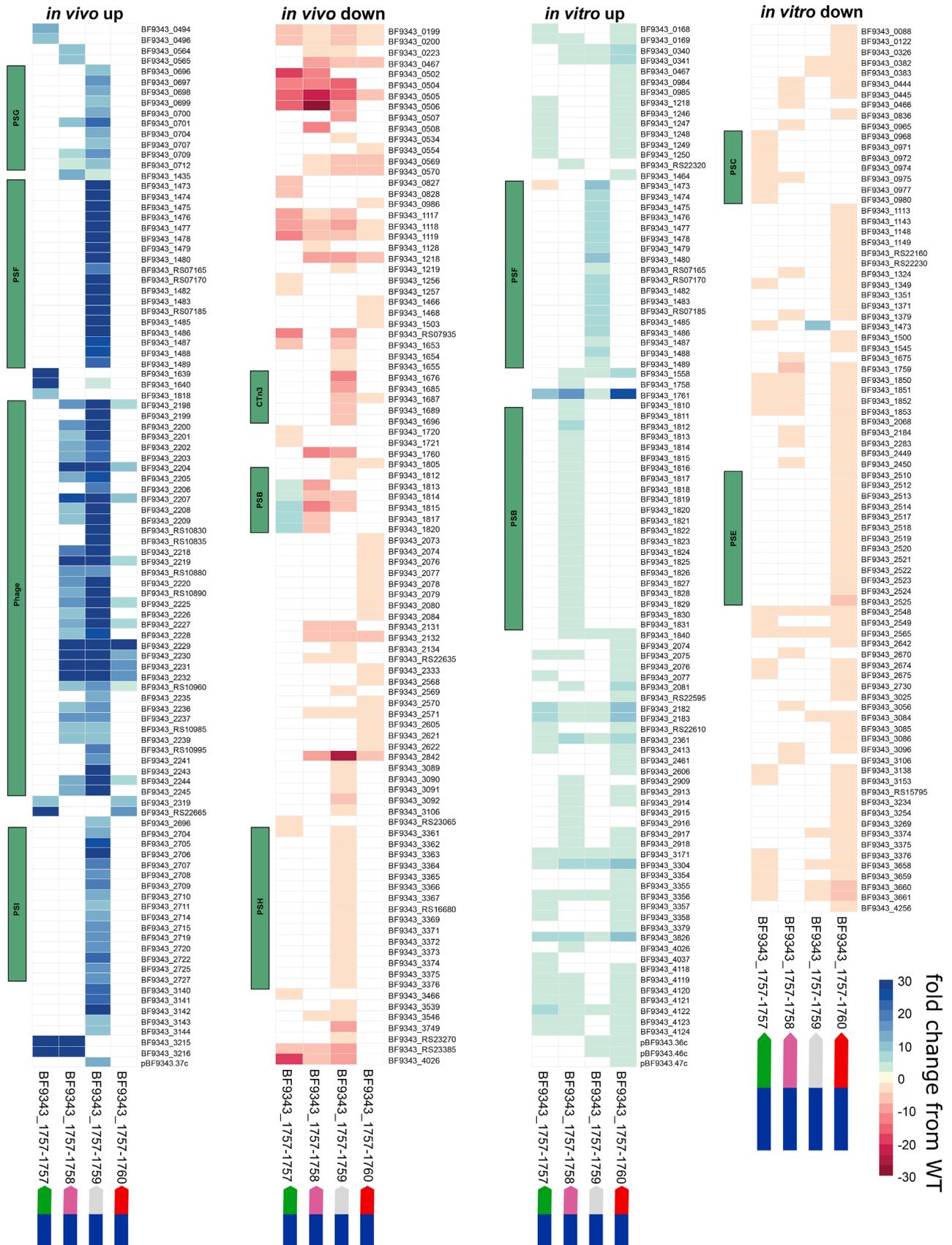
The RNA-Seq data show that there is more polysaccharide F (PSF) transcript in the 1757-59 mutant than WT bacteria and more polysaccharide B (PSB) transcript from the 1757-58 mutant than WT from *in vitro* grown bacteria. We sought to determine if these increases might be due to a greater percentage of the population having these invertible promoters in the ON orientation. We performed a PCR-digestion analysis of all seven invertible polysaccharide promoters in WT and all four mutants to quantify the percentage of bacteria with each promoter in the ON and OFF orientation. We show that the increases in transcripts correlate with the ON/OFF orientations of the relevant PS promoters in these two mutants (1757-58 and 1757-59). (Figure 5A). In addition, the PSE transcript was down-regulated relative to WT in the 1757-60 mutant and this also correlated with most of the PSE promoter being in the OFF orientation in this mutant. The only promoter orientation that was distinct from that of WT that did not correlate with differential transcript levels was the PSE promoter in the 1757-57 mutant, which was in the ON orientation in the vast majority of these bac-

teria. However, we previously showed that PSE transcription is terminated downstream of the promoter when the PSA locus is transcribed by the action of UpaZ (42), and therefore, this PSE phenotype is as expected based on this second level of regulation. These collective data show that the increases and decreases in polysaccharide transcription in these mutants occurs in most cases at the level of promoter orientation.

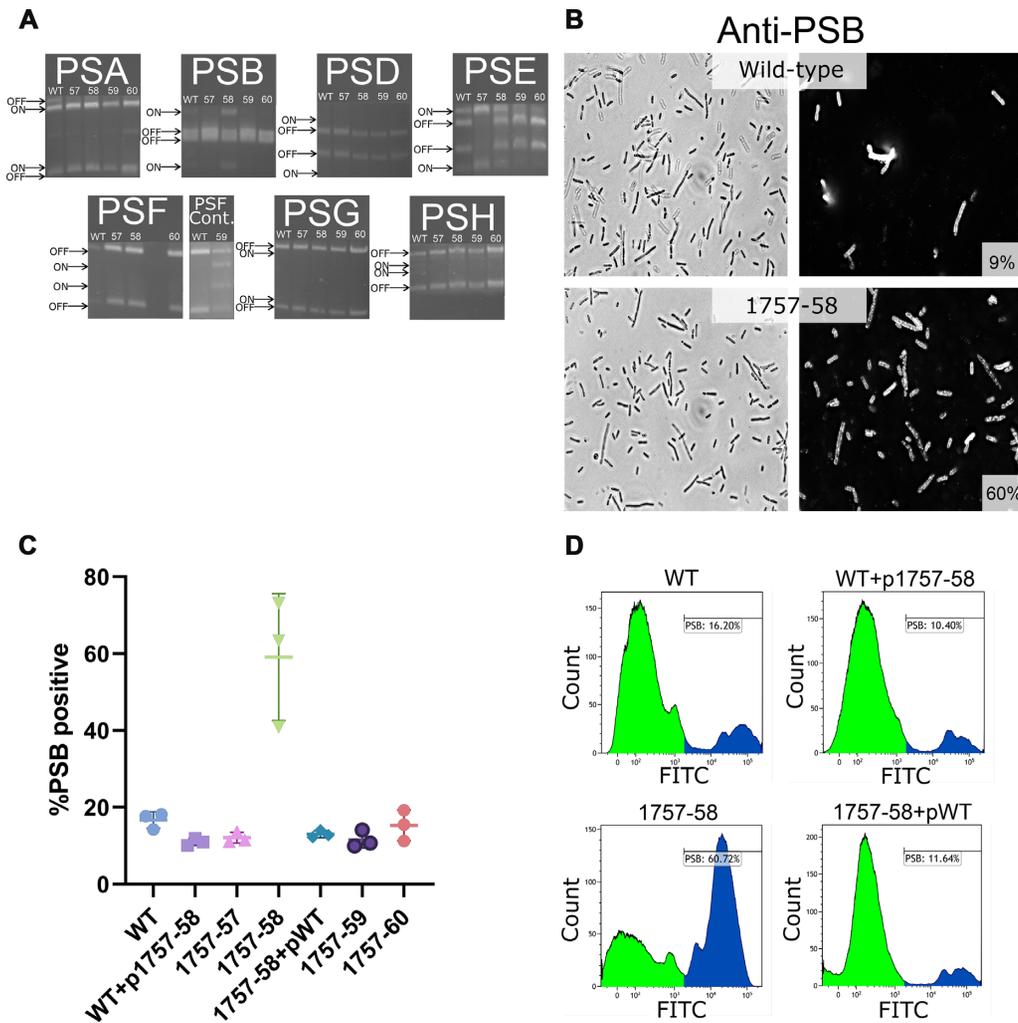
To determine if the RNA-Seq data and the promoter orientations are predictive of phenotype, we chose to analyze synthesis of PSB using an antiserum to this capsular polysaccharide (1). By the EdgeR analysis, the expression of genes in the PSB biosynthesis locus are upregulated 4.41–7.91-fold in the 1757-58 mutant compared to WT from *in vitro* grown bacteria. Indeed, PSB was present on the surface in 60% of the population from this mutant compared to WT (9 or 16%), quantified both by imaging (immunofluorescence, Figure 5B) and by flow-cytometry (Figure 5C, D). Restoration of the full complement of specificity genes with the upstream methylase gene transcribed from a plasmid-borne promoter revealed a reversal of this phenotype, which did not occur when the plasmid was added alone. In addition, when an expression plasmid containing the single locked 1757-58 specificity gene was added to WT, it failed to reproduce the PSB phenotype of the mutant (Figure 5C, D). Therefore, this phenotype may be dependent on methylation only at 1757-58 sites with methylation at other sites precluding the phenotype.

### Identification of hundreds of putative phase-variable Type I R-M systems in diverse Bacteroidetes species

Bacteroidetes is a diverse phylum containing not only species that are members of mammalian associated microbial communities such as the gut, oral cavity, vagina and rumen, but also environmental species of marine and terres-



**Figure 4.** Heatmap of differentially regulated genes in the locked-ON mutants compared to WT. The top 100 (84 in the case of *in vitro* down) differentially regulated genes (minimum 2-fold change,  $P$ -value  $\leq 0.05$  by both DESeq and EdgeR) for each of the four locked-ON mutants under both *in vivo* and *in vitro* conditions. The scale's endpoints are not absolute, any gene exhibiting greater or less than 30 fold change is colored as the endpoint color.



**Figure 5.** Phenotypic validation of RNA-Seq analysis. (A) Quantification of relative promoter orientations for the seven invertible PS promoters from each population, (WT, and each locked-ON mutant), using PCR-digestion. Each assay results in four bands, two corresponding to the ON orientation and two to the OFF orientation. (B) Immunofluorescence of bacteria using an antibody to PSB from WT and locked 1757-58. Percentage of labeled bacteria is shown in the bottom right (600 X magnification). (C) percentage of PSB-positive cells by FACS for each of four mutants and the WT. All strains contained either the vector control or the vector with the relevant specificity gene(s). p1757-58 contains the methylase gene and 1757-58 and pWT contains the entire WT locus 1756 through 1760 behind a constitutive promoter. (D) Flow cytometry histograms of WT, and 1757-58, with either vector control or p1757-58 and pWT, respectively.

trial ecosystems. To determine how prevalent these phase-variable R-M systems are among the members of this phylum, we searched the genomes of 3528 Bacteroidetes. The search was performed by querying the genomes for genes encoding proteins with the Type I R-M DNA specificity domain model embodied by Pfam (PF01420.18). Based on these systems having full length genes next to half-genes that can recombine into the expression locus, we searched for regions with at least two specificity genes <10 kb apart and for which one had to be 60% or less the size of any of the other S-genes. These search criteria identified 1399 distinct loci present in 686 different genomes, which were of 152 different species and 79 different genera of the Bacteroidetes phylum (Supplementary Table S3). In most cases, the S-genes are adjacent and, in many genomes, there were more than two adjacent S-genes, similar to this Type I R-M system in *B. fragilis* NCTC 9343. This broad and prevalent

distribution of these putative phase-variable Type I R-M systems in diverse species of very different ecosystems suggests that they are biologically relevant to these organisms. Phase-variation by DNA inversion of Type I R-M systems may be another conserved property of this phylum of bacteria that contributes to the ability of these organisms to adapt to changing conditions or threats in their ecosystems.

## DISCUSSION

Here we show that a Type I R-M system of *B. fragilis* is able to extensively diversify its methylation targeting so that within a population of bacteria, eight different sites are methylated, one per bacterial cell. The property of phase varying the synthesis of numerous molecules in the gut Bacteroidales has long been predicted to be a bet-hedging strategy to ensure survival of a strain by creating diversity in the population (8). The widespread presence of phase-variable

Type I R-M systems throughout the Bacteroidetes phylum suggests a fitness benefit to these diverse host-associated and environmental organisms in the complex microbial communities in which they must adapt and compete under changing environmental conditions.

Phase-variable Type I R-M systems with shuffling target domain genes were first described in *Mycoplasma pulmonis* (43) and have since been described in other bacteria (reviewed (44)). In addition to contributing to the protection of a bacterial population from phage infection (16,45), these phase-variable systems have also been shown to affect other phenotypes due to differing epigenetic states within a bacterial population. Manso *et al.* showed that locked specificity gene mutants in *Streptococcus pneumoniae* had differing virulence capabilities in a mouse model of infection (18), and this difference could be attributed to differential expression of the capsule operon. We similarly found here that the capsular polysaccharide biosynthesis genes were among the most differentially regulated in the specificity gene locked mutants relative to WT. These phenotypes correlated with orientation of the invertible promoters of these polysaccharide biosynthesis regions. Analysis of these promoter regions for the presence of methylation sites that may affect their inversions did not reveal any candidate sites. Indeed, in *S. pneumoniae*, the epigenetic change leading to differential expression of the capsule genes is still not understood. Due to the global nature of the epigenetic changes in these locked specificity gene mutants, the regulation may be multifactorial. Nevertheless, by complementary phenotypic analyses, we confirmed that PSB is synthesized in a greater percentage of the population of the 1757-58 locked mutant compared to WT, correlating with the transcriptomics data. Another phenotypic difference between the specificity genes mutants was in their ability to compete for colonization of the gnotobiotic mouse gut. After just 10 days of colonization, there was a significant difference in their proportions in the stool. We similarly found that the proportions of the specificity gene combinations in the expression locus varies depending on the environment (*in vitro* versus *in vivo*). Indeed, *in vivo* isolated bacteria return to the *in vitro* specificity gene proportions if subsequently passaged *in vitro*.

DNA inversions in *Bacteroides* are typically mediated by tyrosine site-specific recombinases that act locally on nearby inverted repeats. One gene upstream of the Type I R-M locus (BF9343\_1751) and two downstream genes (BF9343\_1761 and BF9343\_1763) encode putative tyrosine type site-specific recombinases. It is likely that all or some of these act to mediate inversions in this region. As there are four different IRs in this regions that are not similar to each other, it is possible that there is at least one additional site-specific recombinase that may also be involved.

A remaining unanswered question is how can an inversion of the specificity region occur with subsequent alteration of the methylation/ restriction site, without self-restriction. In the organism studied here, the genome encodes an ArdA anti-restriction protein, which specifically inhibits the action of restriction enzymes of Type I RM systems (46), allowing only the modification enzyme to function. Therefore, in this strain, the main function of this system is DNA methylation. This *ardA* gene, BF9343\_1161, is contained on an integrative and conjugative element as is

typical of *ardA* genes (47), and therefore, this function could be conjugally transferred to similar strains and species within an ecosystem. Of the 686 Bacteroidetes genomes identified as likely having phase-variable Type I RM systems (Supplementary Table S3), 224 also have a gene encoding a protein with the ArdA pfam PF07275.12. Therefore, in these strains, it is likely that the main function of these Type I RM systems is for phase-variable methylation, possibly modulating phenotypes in these organisms.

## DATA AVAILABILITY

The Illumina reads generated via RNAtag-Seq and used for differential expression analysis have been deposited to the NCBI Sequence Read Archive (SRA) database and assigned BioProject accession no. PRJNA622493. The MINION sequence data also has been deposited to the SRA under the BioProject accession no. PRJNA648829.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

We thank V. Yeliseyev and R. Lavin for assistance with mouse experiments, R. Shofti for assistance with mice at the Technion, D. Alvarez for help with flow cytometry analysis, A. Rasouly for help with RT-qPCR, T. Katz-Ezov and the Technion Genome Center for help in sequencing, S. Rich for operational assistance and I. Ben-Shoshan for help with design and figure illustration. RNA-Seq libraries were constructed and sequenced at the Broad Institute of MIT and Harvard by the Microbial 'Omics Core and Genomics Platform, respectively. The Microbial 'Omics Core also provided guidance on experimental design and conducted preliminary analysis for all RNA-Seq data.

## FUNDING

The HDDC Gnotobiotic facility is funded by [P30DK034854] to the BWH Center for Clinical and Translational Metagenomics; M.M. is supported by the Spanish Ministry project [AGL2016-75536-P]; Generalitat of Catalonia [2017SGR170]; Public Health Service [R01AI081843, R01AI120633] from the National Institutes of Health/National Institute of Allergy and Infectious Diseases (to L.E.C.); Generalitat of Catalunya [2005SGR00592 to J.J.]; HFSP [LT00079/2012]; EMBO [ALTF 251-2011] fellowships; Fulbright Award; UNESCO L'Oreal National & International Women in Science Awards; Weizmann Institute of Science–Revson National Postdoctoral Award Program for Advancing Women in Science; Canadian Institute for Advanced Research (CIFAR) [FL-000969]; Technion - Institute of Technology, including 'Keren Hanasi' and 'Cathedral for young PIs'; Technion Integrated Cancer Center; Israeli Science Foundation [ISF 1571/17]; ICRF Acceleration [1016142]; Human Frontiers Career Development Award [HFSP CDA00025/2019-C]; Applebaum family support [11916]; Gutwirth award (to N.G.-Z.); N.G.-Z. is an Azrieli global scholar at CIFAR, and is supported by Alon Fellowships for Outstanding Young Researchers and Horev Fellow of the Taub Foundation.

**Conflict of interest statement.** The authors declare no conflict of interest, except that A.F. and R.J.R. work for New England Biolabs, a company that sells research reagents, including restriction enzymes and DNA methyltransferases, to the scientific community.

## REFERENCES

- Krinos, C.M., Coyne, M.J., Weinacht, K.G., Tzianabos, A.O., Kasper, D.L. and Comstock, L.E. (2001) Extensive surface diversity of a commensal microorganism by multiple DNA inversions. *Nature*, **414**, 555–558.
- Xu, J., Bjursell, M.K., Himrod, J., Deng, S., Carmichael, L.K., Chiang, H.C., Hooper, L.V. and Gordon, J.I. (2003) A genomic view of the human-*Bacteroides thetaioetaomicron* symbiosis. *Science*, **299**, 2074–2076.
- Coyne, M.J. and Comstock, L.E. (2008) Niche-specific features of the intestinal *Bacteroidales*. *J. Bacteriol.*, **190**, 736–742.
- Chatzidaki-Livanis, M., Coyne, M.J., Roche-Hakansson, H. and Comstock, L.E. (2008) Expression of a uniquely regulated extracellular polysaccharide confers a large-capsule phenotype to *Bacteroides fragilis*. *J. Bacteriol.*, **190**, 1020–1026.
- Coyne, M.J., Weinacht, K.G., Krinos, C.M. and Comstock, L.E. (2003) Mpi recombinase globally modulates the surface architecture of a human commensal bacterium. *Proc. Natl. Acad. Sci. U. S. A.*, **100**, 10446–10451.
- Xu, Q., Shoji, M., Shibata, S., Naito, M., Sato, K., Elsliger, M.A., Grant, J.C., Axelrod, H.L., Chiu, H.-J., Farr, C.L. *et al.* (2016) A distinct type of pilus from the human microbiome. *Cell*, **165**, 690–703.
- Coyne, M.J. and Comstock, L.E. (2016) A new pillar in pilus assembly. *Cell*, **165**, 520–521.
- Weinacht, K.G., Roche, H., Krinos, C.M., Coyne, M.J., Parkhill, J. and Comstock, L.E. (2004) Tyrosine site-specific recombinases mediate DNA inversions affecting the expression of outer surface proteins of *Bacteroides fragilis*. *Mol. Microbiol.*, **53**, 1319–1330.
- Roche-Hakansson, H., Chatzidaki-Livanis, M., Coyne, M.J. and Comstock, L.E. (2007) *Bacteroides fragilis* synthesizes a DNA invertase affecting both a local and a distant region. *J. Bacteriol.*, **189**, 2119–2124.
- Fletcher, C.M., Coyne, M.J., Bentley, D.L., Villa, O.F. and Comstock, L.E. (2007) Phase-variable expression of a family of glycoproteins imparts a dynamic surface to a symbiont in its human intestinal ecosystem. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 2413–2418.
- Taketani, M., Donia, M.S., Jacobson, A.N., Lambris, J.D. and Fischbach, M.A. (2015) A phase-variable surface layer from the gut symbiont *Bacteroides thetaioetaomicron*. *mBio*, **6**, e01339–01315.
- Nakayama-Imaohji, H., Hirota, K., Yamasaki, H., Yoneda, S., Nariya, H., Suzuki, M., Secher, T., Miyake, Y., Oswald, E., Hayashi, T. *et al.* (2016) DNA inversion regulates outer membrane vesicle production in *Bacteroides fragilis*. *PLoS One*, **11**, e0148887.
- Cerdeno-Tarraga, A.M., Patrick, S., Crossman, L.C., Blakely, G., Abratt, V., Lennard, N., Poxton, I., Duerden, B., Harris, B., Quail, M.A. *et al.* (2005) Extensive DNA inversions in the *B. fragilis* genome control variable gene expression. *Science*, **307**, 1463–1465.
- Loenen, W.A., Dryden, D.T., Raleigh, E.A. and Wilson, G.G. (2014) Type I restriction enzymes and their relatives. *Nucleic Acids Res.*, **42**, 20–44.
- Rusinov, I.S., Ershova, A.S., Karyagina, A.S., Spirin, S.A. and Alexeevski, A.V. (2018) Avoidance of recognition sites of restriction-modification systems is a widespread but not universal anti-restriction strategy of prokaryotic viruses. *BMC Genomics*, **19**, 885.
- Dybvig, K., Sitaraman, R. and French, C.T. (1998) A family of phase-variable restriction enzymes with differing specificities generated by high-frequency gene rearrangements. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 13923–13928.
- Li, J., Li, J.W., Feng, Z., Wang, J., An, H., Liu, Y., Wang, Y., Wang, K., Zhang, X., Miao, Z. *et al.* (2016) Epigenetic switch driven by DNA inversions dictates phase variation in *Streptococcus pneumoniae*. *PLoS Pathog.*, **12**, e0105762.
- Manso, A.S., Chai, M.H., Atack, J.M., Furi, L., De Ste Croix, M., Haigh, R., Trappetti, C., Ogunniyi, A.D., Shewell, L.K., Boitano, M. *et al.* (2014) A random six-phase switch regulates pneumococcal virulence via global epigenetic changes. *Nature communications*, **5**, 5055.
- Sitaraman, R. and Dybvig, K. (1997) The hsd loci of *Mycoplasma pulmonis*: organization, rearrangements and expression of genes. *Molecular microbiology*, **26**, 109–120.
- Pantosti, A., Tzianabos, A.O., Onderdonk, A.B. and Kasper, D.L. (1991) Immunochemical characterization of two surface polysaccharides of *Bacteroides fragilis*. *Infect. Immun.*, **59**, 2075–2082.
- Stevens, A.M., Shoemaker, N.B. and Salyers, A.A. (1990) The region of a *Bacteroides* conjugal chromosomal tetracycline resistance element which is responsible for production of plasmidlike forms from unlinked chromosomal DNA might also be involved in transfer of the element. *J. Bacteriol.*, **172**, 4271–4279.
- Meyer, R.J. and Shapiro, J.A. (1980) Genetic organization of the broad-host-range IncP-1 plasmid R751. *J. Bacteriol.*, **143**, 1362–1373.
- Smith, C.J., Rogers, M.B. and McKee, M.L. (1992) Heterologous gene expression in *Bacteroides fragilis*. *Plasmid*, **27**, 141–154.
- Shishkin, A.A., Giannoukos, G., Kucukural, A., Ciulla, D., Busby, M., Surka, C., Chen, J., Bhattacharyya, R.P., Rudy, R.F., Patel, M.M. *et al.* (2015) Simultaneous generation of many RNA-seq libraries in a single reaction. *Nat. Methods*, **12**, 323–325.
- Zhu, Y.Y., Machleder, E.M., Chenchik, A., Li, R. and Siebert, P.D. (2001) Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *BioTechniques*, **30**, 892–897.
- Li, H. and Durbin, R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**, 589–595.
- Love, M.I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
- Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
- Clark, T.A., Murray, I.A., Morgan, R.D., Kislyuk, A.A., Spittle, K.E., Boitano, M., Fomenkov, A., Roberts, R.J. and Korlach, J. (2012) Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. *Nucleic Acids Res.*, **40**, e29.
- Travers, K.J., Chin, C.S., Rank, D.R., Eid, J.S. and Turner, S.W. (2010) A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res.*, **38**, e159.
- Chin, C.S., Alexander, D.H., Marks, P., Klammer, A.A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E.E. *et al.* (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods*, **10**, 563–569.
- Flusberg, B.A., Webster, D.R., Lee, J.H., Travers, K.J., Olivares, E.C., Clark, T.A., Korlach, J. and Turner, S.W. (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods*, **7**, 461–465.
- Machanic, P. and Bailey, T.L. (2011) MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*, **27**, 1696–1697.
- Zitomersky, N.L., Coyne, M.J. and Comstock, L.E. (2011) Longitudinal analysis of the prevalence, maintenance, and IgA response to species of the order *Bacteroidales* in the human gut. *Infect. Immun.*, **79**, 2012–2020.
- Frith, M.C., Hamada, M. and Horton, P. (2010) Parameters for accurate genome alignment. *BMC Bioinformatics*, **11**, 80.
- Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
- Li, H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
- Estibariz, I., Overmann, A., Ailloud, F., Krebes, J., Josenhans, C. and Suerbaum, S. (2019) The core genome m5C methyltransferase JHP1050 (M.Hpy99III) plays an important role in orchestrating gene expression in *Helicobacter pylori*. *Nucleic Acids Res.*, **47**, 2336–2348.
- Gorrell, R. and Kwok, T. (2017) The *Helicobacter pylori* methylome: roles in gene regulation and virulence. *Curr. Top. Microbiol. Immunol.*, **400**, 105–127.
- Casselli, T., Tourand, Y., Scheidegger, A., Arnold, W.K., Proulx, A., Stevenson, B. and Brissette, C.A. (2018) DNA methylation by

- restriction modification systems affects the global transcriptome profile in *Borrelia burgdorferi*. *J. Bacteriol.*, **200**, e00395-18.
42. Chatzidaki-Livanis, M., Weinacht, K. and Comstock, L. (2010) *Trans* locus inhibitors limit concomitant polysaccharide synthesis in the human gut symbiont *Bacteroides fragilis*. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 11976–11980.
  43. Dybvig, K. and Yu, H. (1994) Regulation of a restriction and modification system via DNA inversion in *Mycoplasma pulmonis*. *Mol. Microbiol.*, **12**, 547–560.
  44. De Ste Croix, M., Vacca, I., Kwun, M.J., Ralph, J.D., Bentley, S.D., Haigh, R., Croucher, N.J. and Oggioni, M.R. (2017) Phase-variable methylation and epigenetic regulation by type I restriction-modification systems. *FEMS Microbiol. Rev.*, **41**, S3–S15.
  45. O’Sullivan, D., Twomey, D.P., Coffey, A., Hill, C., Fitzgerald, G.F. and Ross, R.P. (2000) Novel type I restriction specificities through domain shuffling of HsdS subunits in *Lactococcus lactis*. *Mol. Microbiol.*, **36**, 866–875.
  46. Delver, E.P., Kotova, V.U., Zavilgelsky, G.B. and Belogurov, A.A. (1991) Nucleotide sequence of the gene (*ard*) encoding the antirestriction protein of plasmid *colIb-P9*. *J. Bacteriol.*, **173**, 5887–5892.
  47. Serfiotis-Mitsa, D., Roberts, G.A., Cooper, L.P., White, J.H., Nutley, M., Cooper, A., Blakely, G.W. and Dryden, D.T. (2008) The *Orf18* gene product from conjugative transposon *Tn916* is an *ArdA* antirestriction protein that inhibits type I DNA restriction-modification systems. *J. Mol. Biol.*, **383**, 970–981.