

**Supplementary Information for
“Phosphorothioate DNA modification by BREX Type 4 systems in
the human gut microbiome”**

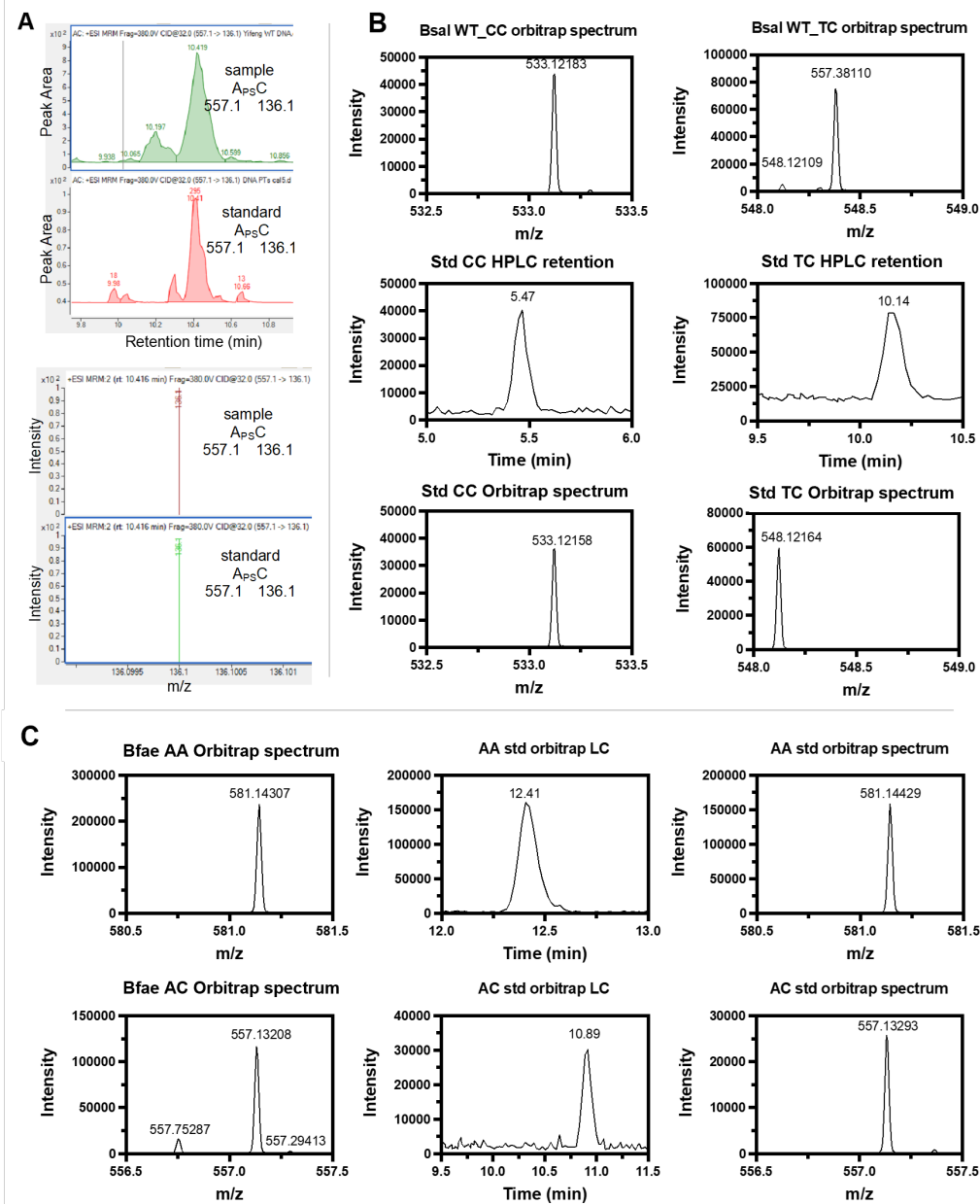
Yifeng Yuan, Michael S. DeMott, Shane R. Byrne, Katia Flores, Mathilde Poyet, Mathieu Groussin, Global Microbiome Conservancy, Brittany Berdy, Jenni Lehtimäki, Tommi Vatanen, Charles A Onyekwere, Ivan Emil Mwikarago, Yvonne Ayerki Nartey, Lewis R. Roberts, Le Thanh Tu Nguyen, Audax Z. P. Mabulla, Jesse Shapiro, John Rusine Bahunde, Catherine Girard, Laurie E. Comstock, Eric J. Alm, and Peter C. Dedon

CONTENTS

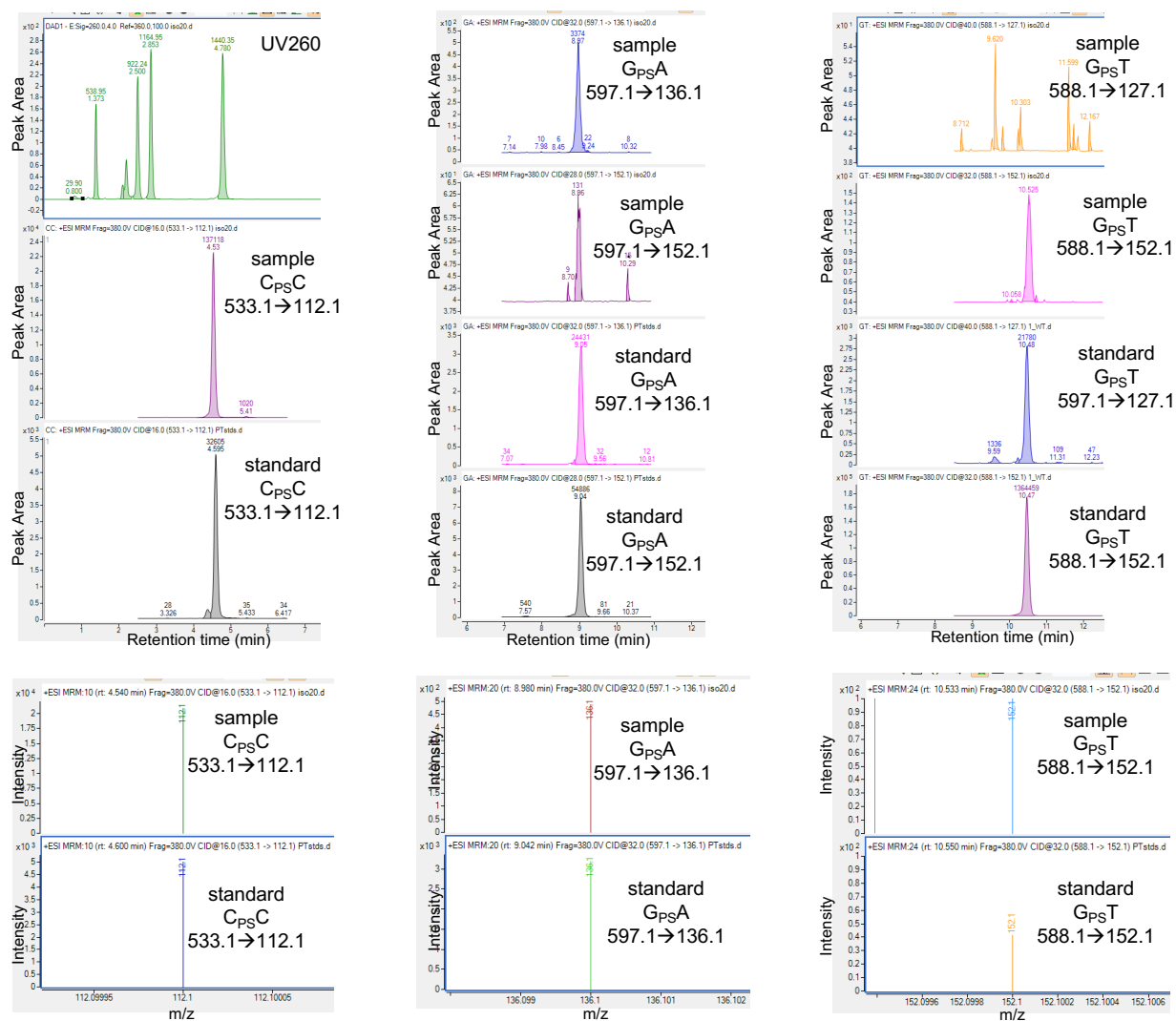
- **Supplementary Figure S1.** Mass spectrometric validation of PT dinucleotides.
- **Supplementary Figure S2.** LC-MS/MS QQQ characterization of PT dinucleotides.
- **Supplementary Figure S3.** Examples of microbiome isolates lacking critical PT synthesis genes.
- **Supplementary Figure S4.** Workflow for PT-seq.
- **Supplementary Figure S5.** The frequencies of known PT- consensus sequences were low at boundaries of coding regions.
- **Supplementary Figure S6:** The density of PT-modified consensus sequences in CDSs and the boundaries of CDSs.
- **Supplementary Figure S7.** The distribution of PTs in *E. coli* BW25113.

- **Supplementary Table S1.** Bacterial strains used in the present studies.
- **Supplementary Table S2.** Primers used in the present studies.
- **Supplementary Table S3.** Mass spectrometry detection parameters for PT-linked dinucleotides (“d(XpsX)”).
- **Supplementary Table S4.** PT dinucleotide high-resolution MS detection parameters used in the present studies.
- **Supplementary Table S5.** PT dinucleotides (A_{PS}C, C_{PS}C, T_{PS}C) per 10⁶ nucleotides in the *Bacteroides* wild-type and engineered strains.
- **Supplementary Table S6.** Statistical comparison of PT dinucleotide levels between the wild-type and engineered strains.

Supplementary Figures



Supplementary Figure S1. Mass spectrometric validation of PT dinucleotides. **(A)** Comparison of the elution times and low-resolution mass transitions of $A_{PS}C$ PT dinucleotide in the *B. salyersiae* microbiome isolate (“sample”) and synthetic standard (“standard”). **(B)** Orbitrap-derived exact masses of $C_{PS}C$ and $T_{PS}C$ PT dinucleotides in *B. salyersiae* microbiome isolate (“Bsal WT”) and HPLC retention and exact masses for synthetic standards (“Std”). **(C)** Orbitrap-derived exact masses of $A_{PS}A$ and $A_{PS}C$ in putative *Butyricimonas faecalis* isolate (“Bfae”) and the exact masses and HPLC retention times for synthetic standards (“std”).



Supplementary Figure S2. LC-MS/MS characterization of PT dinucleotides. The elution time and mass of $C_{PS}C$ (left), $G_{PS}A$ (middle) and $G_{PS}T$ (right) found in DNA from a *Prevotella* strain ("sample"; **Table S4**, Genome ID: 5645FP_0918_058_A7) and PT standards ("standard").

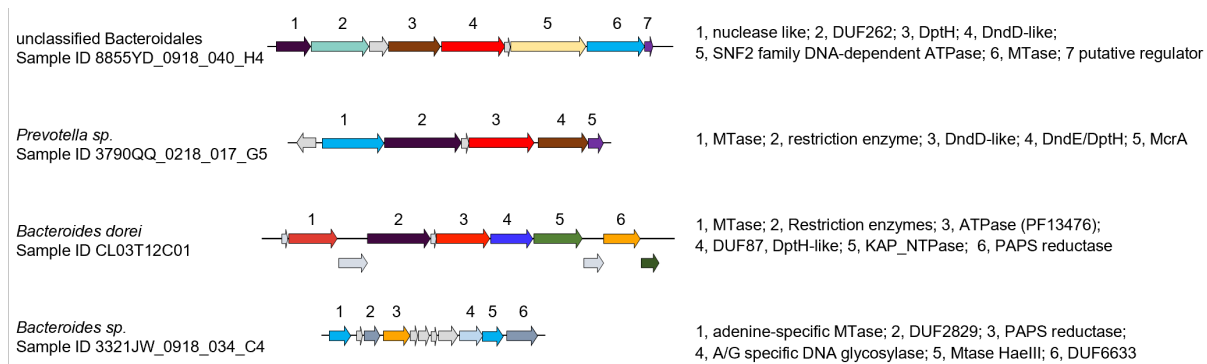


Figure S3. Examples of microbiome isolates lacking critical PT synthesis genes. The four examples shown lack one or more of the essential enzymes needed for PT synthesis: *dndCD*, *sspBCD*, and *brxPC* genes.

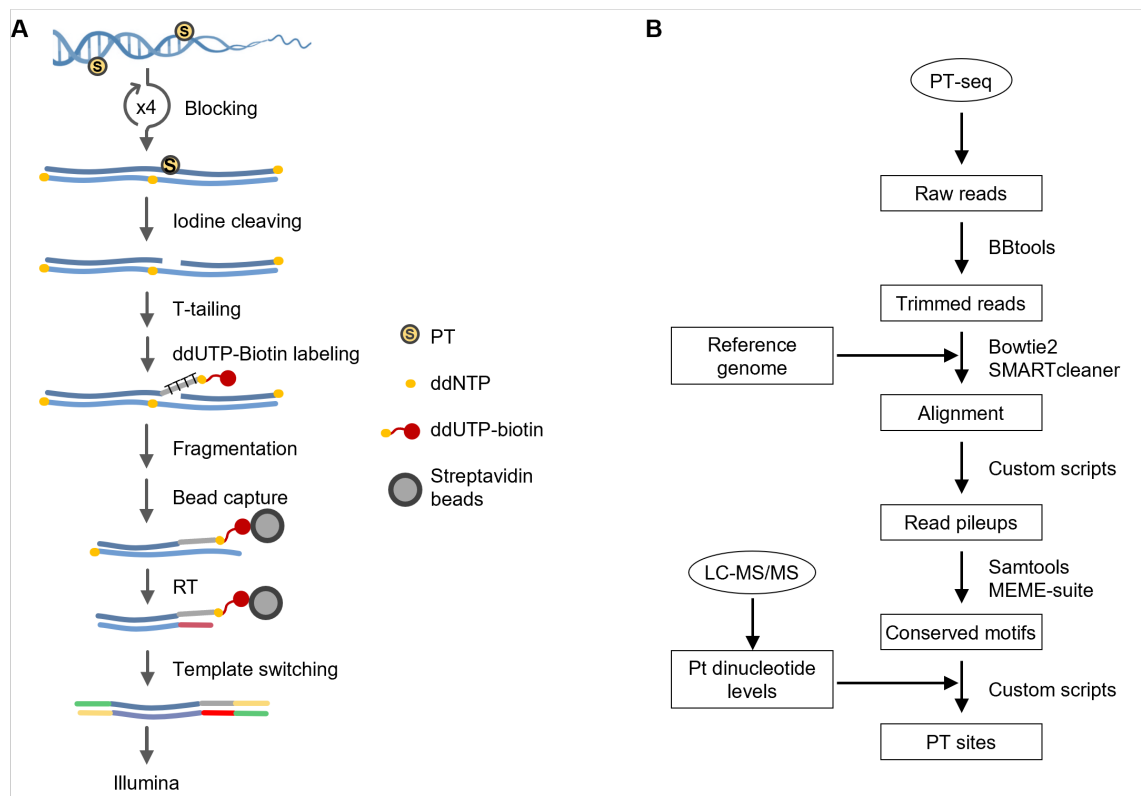


Figure S4. Workflow for PT-seq. **(A)** Using the published PT-seq protocol¹, pre-existing 3'-ends are capped by four cycles of denaturation and blocking using terminal transferase and ddNTP. After iodine cleavage at PT sites, newly introduced 3'-ends are labeled with polyT tails and biotin-ddUTP. Sonication is then used to fragment the DNA and biotinylated fragments are captured on streptavidin beads followed by library preparation using template switching and PCR technologies. **(B)** The workflow for PT-seq data analysis.

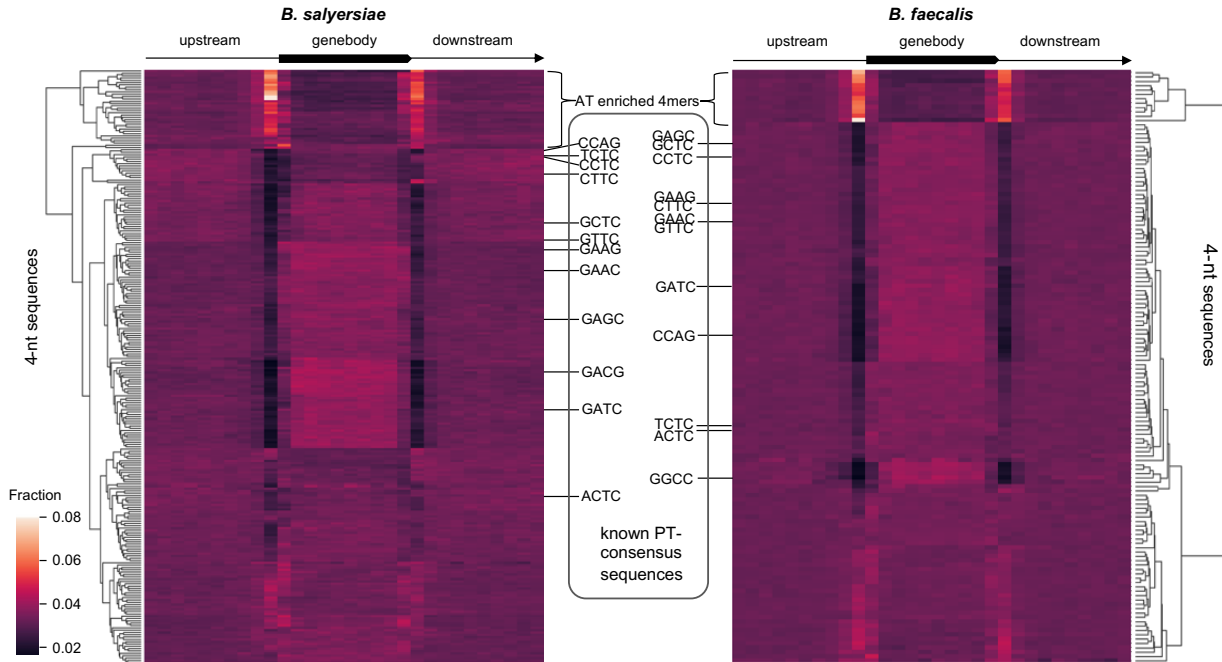


Figure S5. The frequencies of known PT- consensus sequences were low at boundaries of coding regions. The heat map of the distribution of 4- nucleotide sequence in all CDSs and their 1kb upstream and downstream regions in *B. salyersiae* (**left**) and *B. faecalis* (**right**). Each row represents the distribution of each of 256 4- nucleotide sequences from AAAA to TTTT in CDSs and their 1kb upstream and downstream regions. The distribution was colored according to the frequency of each 4- nucleotide sequence in each position. The known PT- consensus sequences, such as GAAC/GTTC in *E.coli*, GAGC/GCTC, and ACTC/CCTC/TCTC identified in this study are indicated by lines. The density and clustering were plotted using the heatmap package in R.

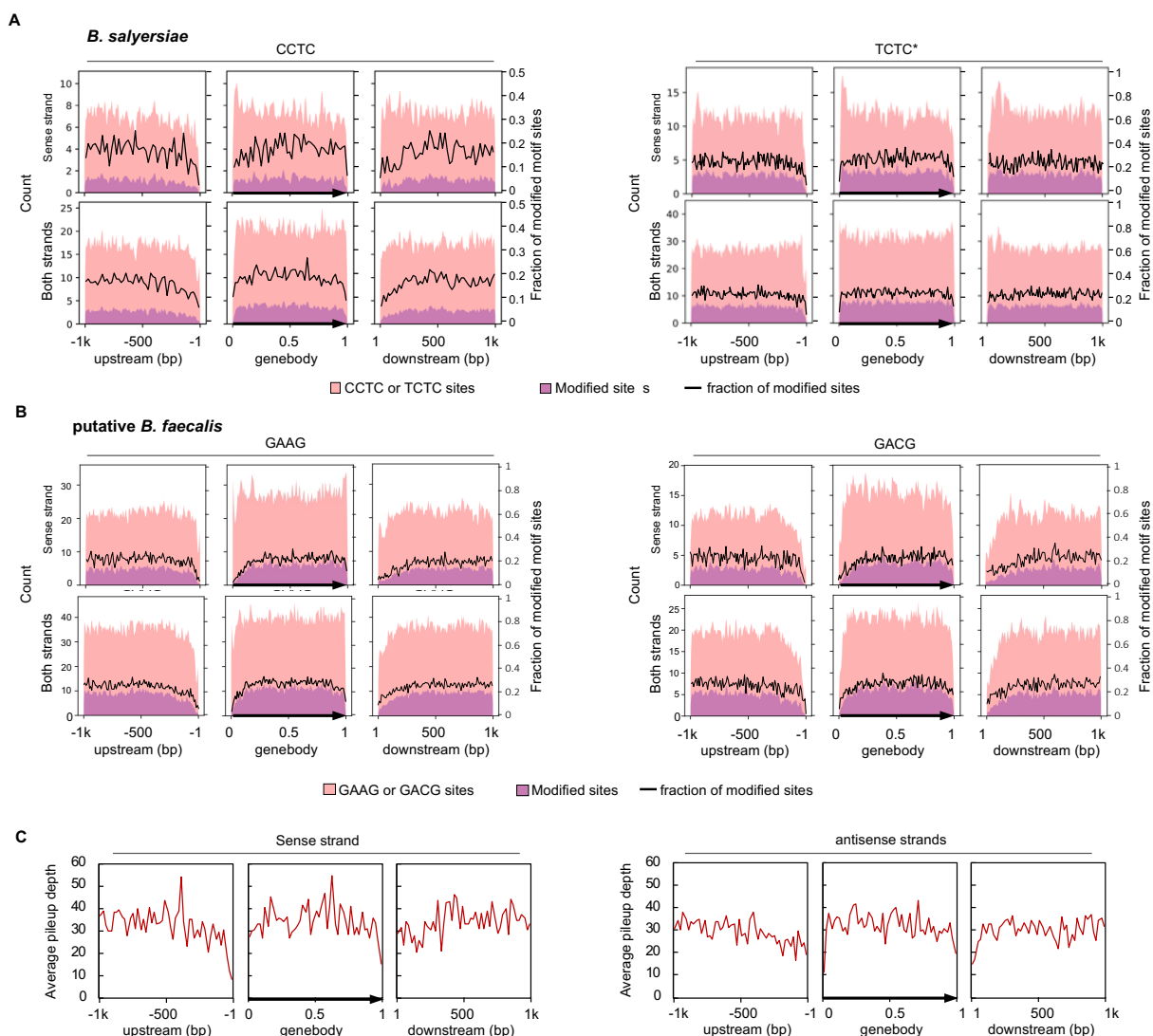


Figure S6. The density of PT-modified consensus sequences in CDSs and the boundaries of CDSs. The number of total consensus sequences in CDSs and their 1kb upstream and downstream regions in *B. salyersiae* (**A**) and putative *B. faecalis* (**B**), respectively. The number of consensus sequences in the sense strand (upper row) or both strands (lower row) are indicated in pink. The numbers of PT-modified consensus sequences are indicated in purple. The fraction of PT-modified sites is indicated by the black line. The asterisk denotes that a less stringent cutoff of depth (50) was applied for better sensitivity to plot a smooth curve for $T_{PS}CCTC$ sites. (**C**) The average depth of read pileups that start with at least 50 reads in *B. salyersiae*. There is no significant difference between sense strand and antisense strand.

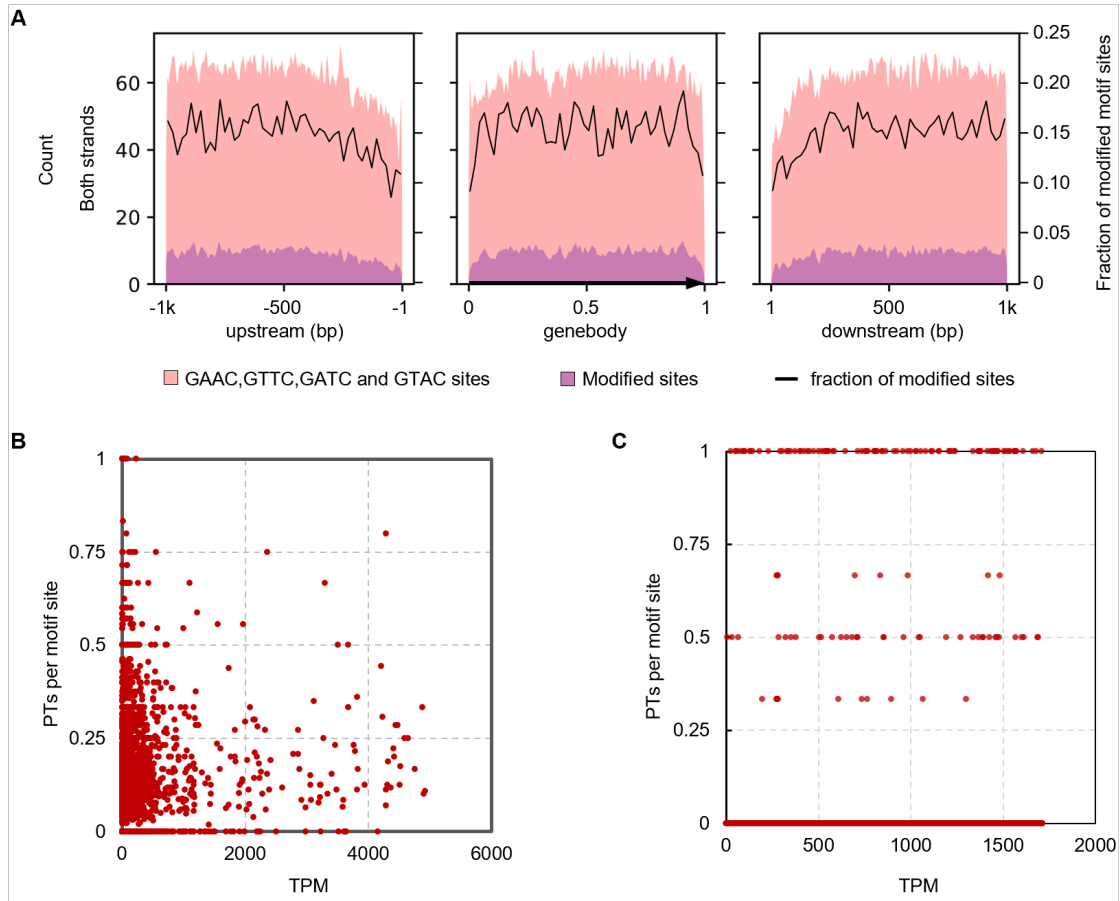


Figure S7. The distribution of PTs in *E. coli* BW25113. **(A)** The number of GAAC, GTTC, GATC and GTAC consensus sequences in both strands of CDSs and their 1kb upstream and downstream regions in *E. coli* BW25113. The number of total GAAC, GTTC, GATC and GTAC sites are indicated in pink. The PT-modified sites are plotted in purple. The fraction of PT-modified motif sites is indicated by the black line. **(B)** The plot of TPM (Transcripts Per Million) and PTs per GAAC, GTTC, GATC and GTAC motifs present in transcripts in *E. coli* BW25113. Data shown here are for 3227 transcripts that contain at least one motif site and 5 to 5000 TPM of RNA-seq. **(C)** A plot of TPM and PTs per GAAC, GTTC, GATC and GTAC motifs within 50 bp upstream of start codons in transcripts in *E. coli* BW25113. Transcripts that contain at least one motif site within 50 bp upstream of start codons are shown.

Supplementary Tables

Supplementary Table S1. Bacterial strains and plasmids used in the present studies. Gene names are noted in *italics* and their locus_tags are noted as “HMPREFXXX_XXXXX”.

Strain	source
<i>B. salyersaie</i> DSM 18765	DSMZ repository
<i>B. salyersaie</i> DSM 18765 $\Delta mcrA$ (HMPREF1532_02792)	this study
<i>B. salyersaie</i> DSM 18765 $\Delta brxC$ (HMPREF1532_02793)	this study
<i>B. salyersaie</i> DSM 18765 $\Delta brxC$ with pKF48	this study
<i>B. salyersaie</i> DSM 18765 $\Delta brxC$ with pKF58	this study
<i>B. salyersaie</i> DSM 18765 $\Delta brxC$ with pKF59	this study
<i>B. salyersaie</i> DSM 18765 $\Delta brxZ$ (HMPREF1532_02794)	this study
<i>B. salyersaie</i> DSM 18765 $\Delta brxL$ (HMPREF1532_02795)	this study
<i>B. salyersaie</i> DSM 18765 $\Delta Bs02796$ (HMPREF1532_02796)	this study
<i>Bacteroides thetaiotaomicron</i> VPI 5482	Zitomersky <i>et al.</i> ²
<i>Bacteroides thetaiotaomicron</i> VPI 5482 with pKF48	this study
<i>Bacteroides thetaiotaomicron</i> VPI 5482 with pKF49	this study
Plasmid	source
pLGB13	Garcia-Bayona <i>et al.</i> ³
pFD340	Smith <i>et al.</i> ⁴
pKF58 [<i>brxC</i> (DPCBNLOC_02819) from <i>Butyrivimonas faecalis</i> (sample ID 5893AJ_0218_015_A8 in Supplementary Data Table 4) in pFD340]	this study
pKF59 [<i>sspC</i> (HMPREF1070_00305) from <i>Bacteroides ovatus</i> (sample ID CL03T12C18 in Supplementary Data Table 4) in pFD340]	this study
pKF49 (HMPREF1532_02791 - 02793 in pFD340)	this study
pKF48 (HMPREF1532_02793 in pFD340)	this study

Supplementary Table S2. Primers used in the present studies. Gene names are noted in *italics* and their locus_tags are noted as “HMPREFXXX_XXXXX”.

Purpose	Direction	Sequence
Construction of <i>B. salyersaie</i> DSM 18765 $\Delta mcrA$ (HMPREF1532_02792)	left flank forward	taagattagcattatgagtgggggtgagacatcgggtattaag
	left flank reverse	ttgaaattcagttcattacgctgtcaatattttaag
	right flank forward	cgtaatgaactgaattcaattttatacttgacaatgg
	right flank reverse	cgaattcctgcagcccggggggtgacggatgagaaatc
Construction of <i>B. salyersaie</i> DSM $\Delta brxC$ (HMPREF1532_02793)	left flank forward	taagattagcattatgagtgggcttgactaaagacgc
	left flank reverse	atgattcgagcgttgccattgtcaagtataaaattg
	right flank forward	aatggcaacgctcgaatcatcaaataaataagaag
	right flank reverse	cgaattcctgcagcccggggggtgtttcgcaatcaaacc
Construction of <i>B. salyersaie</i> DSM $\Delta brxZ$ (HMPREF1532_02794)	left flank forward	taagattagcattatgagtgccaccatacgggctgtcg
	left flank reverse	tagcaccggtgtgaacgactgatacataggctac
	right flank forward	gtcgttaacacgggtgctaccgaagatg
	right flank reverse	cgaattcctgcagcccggggggtgactcacgaaagtatgagg
Construction of <i>B. salyersaie</i> DSM $\Delta brxL$ (HMPREF1532_02795)	left flank forward	taagattagcattatgagtgcgctaattctgcatcttc
	left flank reverse	cttcataccacgaatcttatctgataagg
	right flank forward	taagattcgtggtatggaagaaaatgagttattttaagg
	right flank reverse	cgaattcctgcagcccggggggtgctctgtccagtaatg
Construction of <i>B. salyersaie</i> DSM Δ HMPREF1532_02796	left flank forward	taagattagcattatgagtgccttataggcagttgatg
	left flank reverse	attacccccaaagcttatccaaattttccatag
	right flank forward	ggataagctttgggggtaatttaactgtctggtaac
	right flank reverse	cgaattcctgcagcccgggggagcgctccaccggtaccattc
Clone HMPREF1532_02793 into pFD340	forward	aatcagaattgactctagaggaatttcaattttatacttgacaatg
	reverse	attcgagctcgggtacccgggctatttatttgatgattcgagaattttatc
Clone HMPREF1532_02791-02793 into pFD340	forward	aatcagaattgactctagaggcagaactgctcaatgttg
	reverse	attcgagctcgggtacccgggctatttatttgatgattcgagaattttatc
Clone <i>brx Bf</i> (DPCBNLOC_02819) into pFD340	forward	aatcagaattgactctagagaatagtaatcagtattatcatgg
	reverse	attcgagctcgggtacccgggttaaagacggttgaaaaaatctg
Clone <i>sspC Bo</i> (HMPREF1070_00305) into pFD340	forward	aatcagaattgactctagagggtgctgaatagatattataacaac
	reverse	attcgagctcgggtacccgggtatcgcaatttttcgttc

Supplementary Table S3. Mass spectrometry detection parameters for PT-linked dinucleotides (“d(XpsX”).

Nucleosides	Precursor Ion	Product Ion	Retention Time (min)	Collision Energy	Cell Accelerator Voltage
d(ApsA)	581.1	136.1	12.53	40	3
d(ApsC)	557.1	136.1	11.01	32	3
	557.1	112.1	11.01	24	3
d(ApsG)	597.1	152.1	9.7	48	5
	597.1	136.1	9.7	36	1
d(ApsT)	572.1	136.1	14.75	32	1
	572.1	127.1	14.75	48	5
d(CpsA)	557.1	136.1	8.29	32	1
	557.1	112.1	8.29	28	3
d(CpsC)	533.1	112.1	4.6	16	5
d(CpsG)	573.1	152.1	5.11	32	1
	573.1	112.1	5.11	32	5
d(CpsT)	548.1	127.1	8.79	36	5
	548.1	112.1	8.79	28	5
d(GpsA)	597.1	152.1	9.1	28	3
	597.1	136.1	9.1	32	1
d(GpsC)	573.1	152.1	7.64	32	5
	573.1	112.1	7.64	40	3
d(GpsG)	613.1	152.1	7.14	36	3
d(GpsT)	588.1	152.1	10.64	32	5
	588.1	127.1	10.64	40	3
d(TpsA)	572.1	136.1	12.92	28	3
	572.1	127.1	12.92	44	3
d(TpsC)	548.1	127.1	10.19	44	1
	548.1	112.1	10.19	24	5
d(TpsG)	588.1	152.1	9.2	32	5
	588.1	127.1	9.2	48	1
d(TpsT)	563.1	127.1	14.43	40	5
dA	252.1	136.1	5.23	12	1
dC	228.1	112.1	1.45	16	1
dG	268.1	152.1	2.71	4	3
dT	243.1	127.1	3.22	12	1

Supplementary Table S4. PT dinucleotide high-resolution MS detection parameters used in the present studies.

Nucleosides	Calculated Exact Mass	Retention Time (min)
d(ApsA)	581.14388	12.41
d(ApsC)	557.13264	10.92
d(ApsG)	597.13879	9.61
d(ApsT)	573.13231	14.77
d(CpsA)	557.13264	8.27
d(CpsC)	533.12141	5.47
d(CpsG)	573.12756	5.31
d(CpsT)	548.12108	8.81
d(GpsA)	597.13879	9.61
d(GpsC)	573.12756	7.65
d(GpsG)	613.13371	7.2
d(GpsT)	588.12722	11.54
d(TpsA)	573.13231	12.88
d(TpsC)	548.12108	10.14
d(TpsG)	588.12722	9.18
d(TpsT)	563.12074	14.43
dA	252.10912	5.23
dC	228.09788	1.45
dG	268.10403	2.71
dT	243.09755	3.22

Supplementary Table S5. PT dinucleotides (A_{PS}C, C_{PS}C, T_{PS}C) per 10⁶ nucleotides in the *Bacteroides* wild-type and engineered strains.

Strain	replicate 1	replicate 2	replicate 3	mean	SD
WT	383.0	387.5	394.6	388.4	5.9
ΔmcrA	485.0	483.7	305.5	424.7	103.3
ΔbrxC	0.0	23.4	0.9	8.1	13.3
ΔbrxZ	483.4	463.8	207.6	384.9	153.9
ΔbrxL	437.9	407.1	251.5	365.5	99.9
ΔBS02796	443.5	423.3	248.1	371.7	107.4
ΔbrxC brxC Bs	225.0	540.4	252.9	339.5	174.6
ΔbrxC brxC Bf	0.0	0.0	0.0	0.0	0.0
ΔbrxC sspC Bo	0.0	0.0	0.0	0.0	0.0
vector	5.3	1.7	0.7	2.5	2.4
brxC Bs	0.6	0.2	3.9	1.6	2.1
brxC-P Bs	161.6	408.7	284.0	284.8	123.5

Supplementary Table S6. Statistical comparison of PT dinucleotide levels between the wild-type and engineered strains. The data represent only significant comparisons (two-tailed, effect size = Cohen's d, $p < 0.05$).

Group A	Group B	t	df	p	Cohen d	CI low 95	CI high 95
WT	<i>ΔbrxC</i>	45.4535	4	1.40E-06	37.1126	357.0387	403.4946
WT	<i>ΔbrxC brxC Bf</i>	115.0187	4	3.43E-08	93.9124	378.9919	397.7415
WT	<i>ΔbrxC sspC Bo</i>	115.0187	4	3.43E-08	93.9124	378.9919	397.7415
WT	vector	105.5809	4	4.83E-08	86.2065	375.6547	395.9453
WT	<i>brxC Bs</i>	108.2173	4	4.37E-08	88.3590	376.8762	396.7238
<i>ΔmcrA</i>	<i>ΔbrxC</i>	6.9315	4	2.27E-03	5.6595	249.7489	583.5177
<i>ΔmcrA</i>	<i>ΔbrxC brxC Bf</i>	7.1243	4	2.05E-03	5.8169	259.2077	590.2590
<i>ΔmcrA</i>	<i>ΔbrxC sspC Bo</i>	7.1243	4	2.05E-03	5.8169	259.2077	590.2590
<i>ΔmcrA</i>	vector	7.0793	4	2.10E-03	5.7802	256.5956	587.7378
<i>ΔmcrA</i>	<i>brxC Bs</i>	7.0966	4	2.08E-03	5.7944	257.6090	588.7243
<i>ΔbrxC</i>	<i>ΔbrxZ</i>	-4.2257	4	1.34E-02	-3.4503	-624.4260	-129.2410
<i>ΔbrxC</i>	<i>ΔbrxL</i>	-6.1414	4	3.56E-03	-5.0145	-518.9750	-195.8250
<i>ΔbrxC</i>	<i>ΔBS02796</i>	-5.8154	4	4.35E-03	-4.7483	-537.0940	-189.9730
<i>ΔbrxC</i>	<i>ΔbrxC brxC Bs</i>	-3.2774	4	3.06E-02	-2.6760	-612.0200	-50.6464
<i>ΔbrxC</i>	<i>brxC-P Bs</i>	-3.8564	4	1.82E-02	-3.1487	-475.8550	-77.4785
<i>ΔbrxZ</i>	<i>ΔbrxC brxC Bf</i>	4.3325	4	1.23E-02	3.5375	138.2545	631.6122
<i>ΔbrxZ</i>	<i>ΔbrxC sspC Bo</i>	4.3325	4	1.23E-02	3.5375	138.2545	631.6122
<i>ΔbrxZ</i>	vector	4.3031	4	1.26E-02	3.5135	135.6573	629.0760
<i>ΔbrxZ</i>	<i>brxC Bs</i>	4.3145	4	1.25E-02	3.5228	136.6663	630.0670
<i>ΔbrxL</i>	<i>ΔbrxC brxC Bf</i>	6.3357	4	3.18E-03	5.1730	205.3289	525.6711
<i>ΔbrxL</i>	<i>ΔbrxC sspC Bo</i>	6.3357	4	3.18E-03	5.1730	205.3289	525.6711
<i>ΔbrxL</i>	vector	6.2893	4	3.27E-03	5.1352	202.7153	523.1514
<i>ΔbrxL</i>	<i>brxC Bs</i>	6.3072	4	3.23E-03	5.1498	203.7291	524.1375
<i>ΔBS02796</i>	<i>ΔbrxC brxC Bf</i>	5.9901	4	3.91E-03	4.8909	199.3790	543.8876
<i>ΔBS02796</i>	<i>ΔbrxC sspC Bo</i>	5.9901	4	3.91E-03	4.8909	199.3790	543.8876
<i>ΔBS02796</i>	vector	5.9472	4	4.01E-03	4.8559	196.7687	541.3646
<i>ΔBS02796</i>	<i>brxC Bs</i>	5.9638	4	3.97E-03	4.8694	197.7816	542.3517
<i>ΔbrxC brxC Bs</i>	<i>ΔbrxC brxC Bf</i>	3.3672	4	2.81E-02	2.7493	59.5521	619.3145
<i>ΔbrxC brxC Bs</i>	<i>ΔbrxC sspC Bo</i>	3.3672	4	2.81E-02	2.7493	59.5521	619.3145
<i>ΔbrxC brxC Bs</i>	vector	3.3414	4	2.88E-02	2.7283	56.9586	616.7747
<i>ΔbrxC brxC Bs</i>	<i>brxC Bs</i>	3.3514	4	2.85E-02	2.7364	57.9665	617.7668
<i>ΔbrxC brxC Bf</i>	<i>brxC-P Bs</i>	-3.9921	4	1.62E-02	-3.2595	-482.8180	-86.7155
<i>ΔbrxC sspC Bo</i>	<i>brxC-P Bs</i>	-3.9921	4	1.62E-02	-3.2595	-482.8180	-86.7155
vector	<i>brxC-P Bs</i>	-3.9554	4	1.67E-02	-3.2295	-480.2890	-84.1108
<i>brxC Bs</i>	<i>brxC-P Bs</i>	-3.9696	4	1.65E-02	-3.2412	-481.2780	-85.1221

References for Supplementary information

¹Yuan, Y., DeMott, M., Byrne, S., GMbC & Dedon, P. PT-seq for highly sensitive metagenomic mapping of phosphorothioate DNA modifications. *bioRxiv* **2024.06.03.597111** (2024).

²Zitomersky NL, Coyne MJ, Comstock LE. Longitudinal analysis of the prevalence, maintenance, and IgA response to species of the order Bacteroidales in the human gut. *Infect Immun* **79**(5):2012-20 (2011).

³García-Bayona L, Comstock LE. Streamlined Genetic Manipulation of Diverse *Bacteroides* and *Parabacteroides* Isolates from the Human Gut Microbiota. *mBio* **10**(4), e01762-19 (2019).

⁴Smith, C. J., Parker, A. Rogers, M. B. Plasmid transformation of *Bacteroides* spp. by electroporation. *Plasmid* **24**(2), 100–109 (1990).