Nature Communications

https://doi.org/10.1038/s41467-026-68412-5

Article in Press

# Phosphorothioate DNA modification by BREX type 4 systems in the human gut microbiome

Received: 4 June 2025

Accepted: 2 January 2026

Cite this article as: Yuan, Y., DeMott, M.S., Byrne, S.R. *et al.* Phosphorothioate DNA modification by BREX type 4 systems in the human gut microbiome. *Nat Commun* (2026). https://doi.org/10.1038/s41467-026-68412-5

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

# Phosphorothioate DNA modification by BREX type 4 systems in the human gut microbiome

Yifeng Yuan[1], Michael S. DeMott[1,2], Shane R. Byrne[1,&], Katia Flores[3], Mathilde Poyet[1,4,5], Mathieu Groussin[1,5,6], Brittany Berdy[7,†], John Rusine Bahunde[5,8,9,#,¥], Catherine Girard[5,10,11,¥], Jenni Lehtimäki[5,12,¥], Audax Z. P. Mabulla[5,13,¥], Ivan Emil Mwikarago[5,8,14,¥], Yvonne Ayerki Nartey[5,15,¥], Le Thanh Tu Nguyen[1,5,¥], Charles A Onyekwere[5,16,¥], Lewis R. Roberts[5,17,¥], Jesse Shapiro[5,18,19,20,¥], Tommi Vatanen[5,7,21,22,23,24,¥], Laurie E. Comstock[3], Eric J. Alm[1,5,7,25,26], and Peter C. Dedon[1,2,26,*]

[1] Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.

[2] Center for Environmental Health Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.

[3] Department of Microbiology, Duchossois Family Institute, University of Chicago, Chicago, Illinois, USA.

[4] Institute of Experimental Medicine, Kiel University, Germany.

[5] Global Microbiome Conservancy (https://microbiomeconservancy.org/), Kiel University, Kiel, Germany

[6] Institute of Clinical and Molecular Biology, Kiel University, Kiel, Germany.

[7] Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA

[8] University of Rwanda College of Medicine and Health Sciences, Department of Laboratory Sciences, Kigali, Rwanda

[9] United States Pharmacopeia, Rockville, Maryland, USA

[10] University of Quebec at Chicoutimi, Saguenay, Quebec, Canada

[11] University of Laval, Quebec, Quebec, Canada.

[12] Finnish Environment Institute, Helsinki, Finland

[13] University of Dar es Salaam, Dar es Salaam Region, Tanzania

[14] Rwanda Food and Drug Authority, Department of Drugs and Device Registration, Division of Human Medicine and Device Registration, Kigali, Rwanda

[15] Department of Internal Medicine. School of Medical Sciences, University of Cape Coast, Cape Coast, Ghana

[16] Department of Medicine Lagos State University College of Medicine, Lagos, Nigeria
[17] Mayo Clinic College of Medicine and Science, Rochester, Minnesota, USA
[18] Department of Microbiology and Immunology, McGill University, Montreal, Quebec Canada
[19] McGill Genome Centre, McGill University, Montreal, Quebec Canada
[20] McGill Centre for Microbiome Research, Montreal, Quebec Canada
[21] Institute of Biotechnology, Helsinki Institute of Life Science, University of Helsinki, Helsinki, Finland
[22] Department of Microbiology, Faculty of Agriculture and Forestry, University of Helsinki, Helsinki, Finland
[23] Research Program for Clinical and Molecular Metabolism, Faculty of Medicine, University of Helsinki, Helsinki, Finland
[24] Liggins Institute, University of Auckland, Auckland, New Zealand
[25] Center for Microbiome Informatics and Therapeutics, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[26] Singapore-MIT Alliance for Research and Technology, Singapore

* Corresponding author: Peter Dedon, pcdedon@mit.edu
[&] Present address: Codomax Inc., Worcester, MA USA
[†] Present address: MGH Institute of Health Professions, Boston, Massachusetts, USA
[#] Present address: Community Options, Derwood, MD USA
[¥] These members of the Global Microbiome Conservancy are listed in alphabetical order

**Keywords:** epigenetics, microbiome, phosphorothioate, metagenomics, comparative genomics, mass spectrometry, next-generation sequencing

## Abstract

Among dozens of microbial DNA modifications regulating gene expression and host defense, phosphorothioation (PT) is the only known backbone modification, with sulfur inserted at a non-bridging oxygen by *dnd* and *ssp* gene families. Here we explored the distribution of PT genes in 13,663 human gut microbiome genomes, finding that 6.3% possessed *dnd* or *ssp* genes predominantly in Bacillota, Bacteroidota, and Pseudomonadota. This analysis revealed several previously undescribed PT synthesis systems, including type 4 Bacteriophage Exclusion (BREX) *brx* genes, which we genetically validated in *Bacteroides salyersiae.* Mass spectrometric analysis of DNA from 226 gut microbiome isolates possessing *dnd*, *ssp*, and *brx* genes revealed 8 PT dinucleotide settings confirmed in 6 consensus sequences by PT-specific DNA sequencing. Genomic analysis showed PT enrichment in rRNA genes and depletion at gene boundaries. These results illustrate the power of the microbiome for discovering prokaryotic epigenetics and the widespread distribution of oxidation-sensitive PTs in gut microbes.


## Introduction

Epigenetic modifications have been found in DNA from all domains of life. Among the variations in canonical A, T, C and G structure, such as $N^6$-methyl-adenine (6mA), $N^4$-methyl-cytosine (4mC), $C^5$-methyl-cytosine (5mC), and 7-deazaguanine derivatives[1, 2], phosphorothioates (PT) are the only known modification of the sugar-phosphate backbone, with a non-bridging oxygen replaced by sulfur in $R_P$ specific configuration[3-5] (**Fig. 1A**). Two PT-based restriction and modification (R-M) systems, the *dnd*[4, 6] and *ssp*[7, 8] gene clusters, have been observed in ~10% of bacteria and archaea[9, 10], with a third *tpd* system recently discovered in extremophilic bacteria[11]. Typically, the modification components are organized as a four-gene operon for *dndBCDE* and a three-gene operon for *sspBCD*, with an additional modification gene for *dndA* or *sspA* located adjacent to the operon, dispersed elsewhere in the genome, or replaced by a gene for a homolog such as *iscS*. However, the recently discovered tpdABC cluster

requires only a single gene, *tpdC*, for PT synthesis[11]. Like methylation-based R-M systems[8, 10, 12], DndACDE and SspABCD proteins catalyze PT modification on one or both strands of specific consensus sequences. For example, DndACDE confer double-stranded PTs at 5′-G$_{PS}$AAC-3′/5′-G$_{PS}$TTC-3′ sequences in *Escherichia coli* B7A and *Salmonella enterica* serovar Cerro 87, 5'-G$_{PS}$GCC-3′/5′-G$_{PS}$GCC-3′ in *Pseudomonas fluorescens* pf0-1 and *Streptomyces lividans* 1326, and 5′-G$_{PS}$ATC-3′/5′-G$_{PS}$ATC-3′ in *Hahella chejuensis* KCTC 2396[12-14]. SspABCD proteins, on the other hand, catalyze single-stranded 5′-C$_{PS}$CA-3′ in *Vibrio cyclitrophicus* FF75[7, 13]. The *dnd* and *ssp* systems share some similarities, such as encoding a homolog of phosphoadenosine phosphosulphate (PAPS) reductase (DndC, SspD)[7, 15], a homolog of cysteine desulfurase (DndA, SspA)[7, 15], and a P-loop containing ATPase (DndD, SspC)[7, 16] (**Fig. 1B**). The restriction counterparts DndFGH and SspFGH sense PT modifications by poorly understood mechanisms, with only 10-15% of consensus sequences modified with PTs[7, 13].

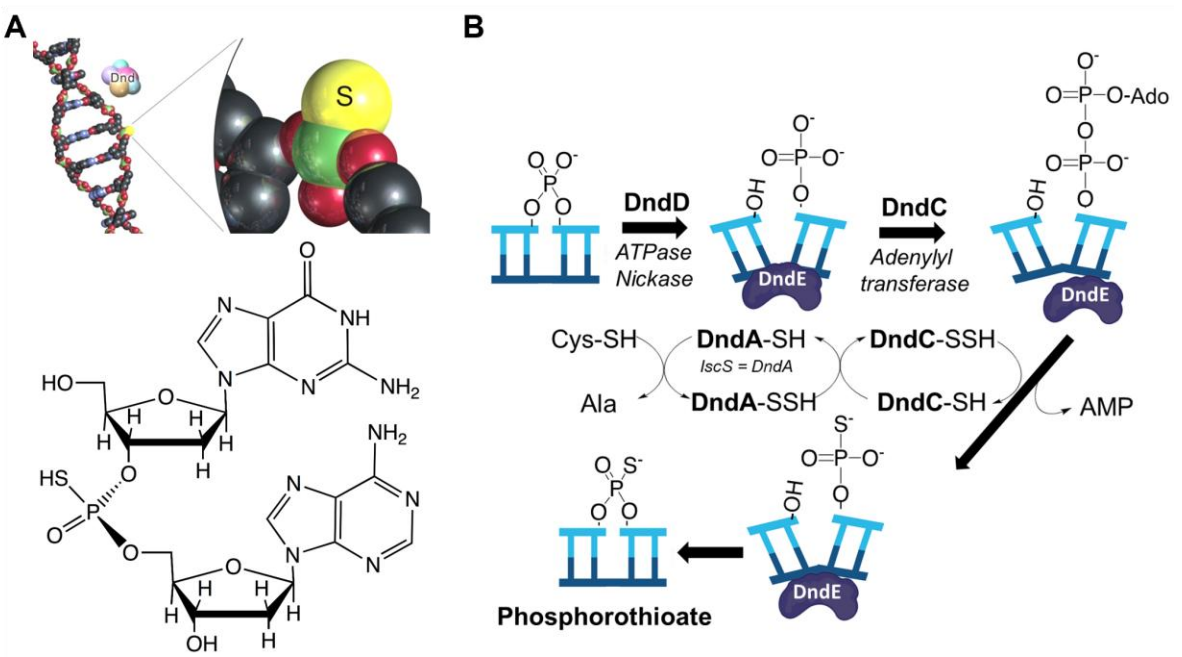


**Figure 1**. Microbial phosphorothioate (PT) DNA modifications. (**A**) Sulfur replaces a non-bridging phosphate oxygen in the DNA backbone in PT modifications. A limit nuclease digest of PT-containing DNA leaves PT-linked dinucleotides that can be identified and quantified by LC-MS. (**B**) Synthesis of PTs generally follows the biochemical steps performed by Dnd proteins.

Both the metabolic pathways enabling sulfur incorporation into PTs and the altered chemical properties of sulfur-modified DNA have implications for interactions of PT-containing microbes with human hosts. For example, the sulfur in PTs is both readily oxidized and nucleophilic, which is proposed to provide epigenetic regulation of transcription of redox homeostasis genes[12]. PTs also provide weak to modest protective effects in cells exposed to reactive oxygen and nitrogen species, such as peroxides [12, 17] and peroxynitrite[18]. Contrasting with this protection, PT-containing bacteria are 5-fold more sensitive to neutrophil-derived hypochlorous acid (HOCl) due to extensive DNA breaks at PTs [19]. These unusual chemical properties of PTs raise questions about how PT-containing microbes might behave in the healthy gut microbiome or be altered by inflammatory bowel disease (IBD) or other chronic inflammatory conditions[20-22]. Evidence for the presence of PT genes in bacterial strains associated with the human gut microbiome and PT dinucleotides in fecal DNA[23, 24] thus motivated us to systematically analyze PT genomics in the human gut microbiome.

Here we defined the landscape of PT-containing microbes in the human gut by performing a genomic analysis of 13,000 human microbiome genomes from the Broad Institute-OpenBiome Microbiome Library (BIO-ML)[25], the Global Microbiome Conservancy (GMbC)[26-28], and the Unified Human Gastrointestinal Genome (UHGG) collection[29]. Mass spectrometric analysis of PT dinucleotides in 226 of these isolates coupled with PT-specific next-generation sequencing (PT-seq)[30] led to the discovery of a previously undescribed PT system involving type 4 Bacteriophage Exclusion (BREX) genes *brxPCZL*. These results expand our knowledge about the diversity of PT epigenetics and lay the foundations for understanding the role of PT-containing microbes in human health and disease.

## Results

### Discovery of previously uncharacterized PT modification systems by analyzing genome neighborhoods of *dndC* and *sspD* genes

Given the power of physical clustering analyses to identify gene functions in bacteria, we first performed a comprehensive gene neighborhood analysis[31] of *dnd* and *ssp* genes to find new PT modification systems. Here we used Enzyme Function Initiative (EFI) tools[31] to first search Uniprot for DndC and SspD homologs in sequence similarity networks (SSN) followed by the EFI Genome Neighborhood Tool to identify the genomic contexts of the SSNs. DndC and SspD possess a PAPS reductase domain with pyrophosphatase activity essential for PT biosynthesis (**Fig. 1B**), an activity shared by sulfur-inserting RNA modification enzymes ThiI[32], MnmA[33], and Ncs6[33], and TtcA[34]. As a second criterion for PT synthesis, we required that DndC and SspD neighborhoods also possess a P-loop-containing NTPase gene. Both DndD and SspC possess P-loop-containing ATPase activity essential for PT synthesis (**Fig. 1B**). Using this approach, we retrieved 3,120 DndC homologs and 2,132 SspD homologs from Uniprot by BLAST (E-value cutoff $10^{-5}$) and searched genome neighborhoods encoding both a PAPS reductase domain and a P-loop NTPase (**Supplementary Data 1, 2**).

The resulting gene neighborhoods revealed several previously undescribed putative PT-modifying gene candidates. Each circle or node in **Figure 2** depicts a neighborhood containing *dndC* (Fig. 1A) and *sspD* (Fig. 1B), with clustering of nodes based on SSNs. Here it is clear that most of the *dndC*-containing gene neighborhoods fall within the largest cluster and possess additional *dnd* genes consistent with canonical Dnd-based PT synthesis (**Fig. 2A**, nodes with red outlines). However, several of the smaller clusters contained neighborhoods with a *dndC* gene, a P-loop NTPase, and other known defense genes, suggesting PT-based RM systems. In one instance, *dndC* and the NTPase gene lie near genes for a Bacteriophage Exclusion (BREX) type 4 system[35] (**Fig. 2A**, yellow nodes). Five of six BREX systems possess a BrxC/PglY ATPase, a PglX DNA methyltransferase, and BrxZ (PglZ) phosphatase. In BREX type 4 clusters containing *brxP*, the PglX adenine methyltransferase is replaced with the DndC S-inserting PAPS reductase domain protein[35]. For example, in *Methanothermobacter sp.* (**Fig. 2A** yellow node, black outline), *brxP* (*dndC* homolog) is adjacent to a *brxC* ATPase gene, a *brxZ* phosphatase gene, and a Lon-like protease domain-containing *brxL* gene typically found in BREX type 1 and 4 systems. In *Fischerella thermalis* (**Fig. 2A,** yellow

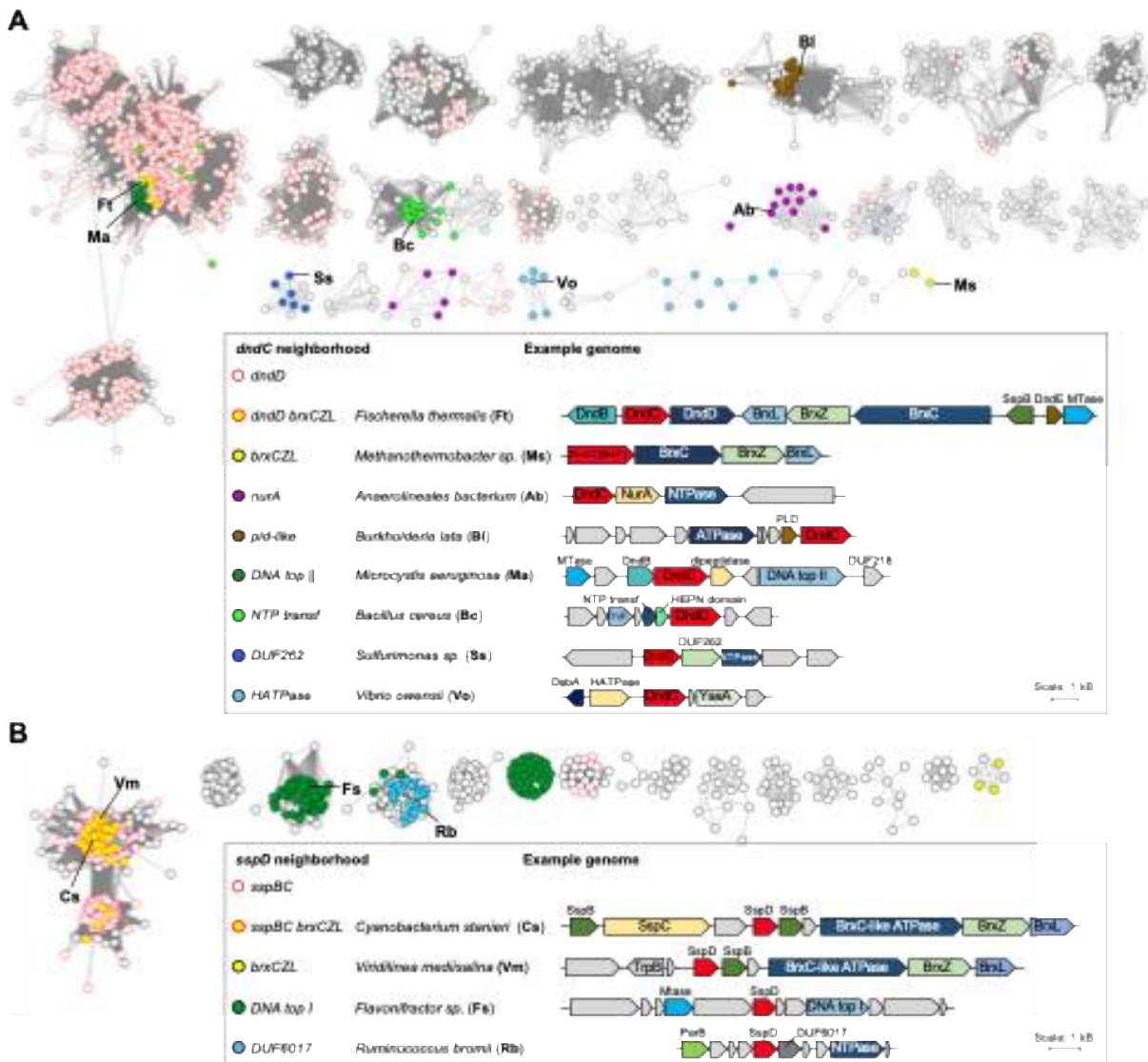


**Figure 2. Gene neighborhoods analyses based on sequence similarity networks (SSNs) of DndC and SspD proteins essential for PT synthesis.** Sequence similarity network (SSN) analyses were performed for the 3,120 closest homologues of DndC (**A**) and 2,132 homologues of SspD (**B**) in the UniProt database. Each node (circle) in the network cluster represents one DndC or SspD protein. An edge is drawn between two nodes that have a BLAST E-value cutoff of $\leq 10^{-100}$ (alignment score of 100) in the DndC SSN or $10^{-50}$ (alignment score of 50) in the SspD SSN. The color-filled nodes in panels **A** and **B** denote specific genome neighborhood structures, with representative bacterial species and their neighborhood structures shown in the rectangular insets. Two-letter labels on the SSN clusters denote bacterial species in the insets. The node outlines are colored red when *dndD* and *dndC* are present in the same genome neighborhood and similarly for *sspD* and *sspBC*. For better visualization, single nodes and clusters with only a few nodes were omitted. Abbreviations: HEPN, higher eukaryotes and prokaryotes nucleotide-binding; MTase, methyltransferase; NTPase, nucleoside triphosphatase; NTP transf, Nucleotidyltransferase; top, topoisomerase.

node, red outline), the full set of *dndBCDE* genes cluster with *brxL*, *brxZ* phosphatase, *sspB* nickase, and a methyltransferase. It is common to find several different defense systems in the same region due to horizontal gene transfer and genetic decay. The systems in these islands can function independently, so the methyltransferases located next to the candidate PT genes may be functional independently of *dnd* or BREX type 4 systems. The synthesis of PTs in the DndC-containing BREX type 4 system was subsequently validated genetically and by LC-MS, as discussed shortly.

In a second putative PT defense system, a small cluster including *Bacillus cereus* (**Fig. 2A**, light green node) pairs *dndC* with a putative minimal nucleotidyltransferase (MNT) and a higher eukaryotes and prokaryotes nucleotide-binding (HEPN) protein. This gene neighborhood resembles a Class II MNT-HEPN toxin-antitoxin (TA) system[36], in which the HEPN protein is a RNase toxin, but its activity is neutralized by adenylylation by the MNT antitoxin[37, 38]. The HEPNs in the *dndC* clusters lack the RNase domain, but the neighboring MNT, ThiF, has the conserved motif $GSX_{10}DXD$ of an adenylyl transferase. Coupled with the adjacent NTPase, the DndC encoded in this Class II MNT-HEPN neighborhood could confer a three-component PT-based stress response system.

Finally, the *dndC* genes clustered with P-loop protein genes in the genome neighborhoods in small clusters such as *Sulfurimonas sp.* (**Fig. 2A**, blue node), *Anaerolineales bacterium* (**Fig. 2A**, pink node), *Burkholderia lata* (**Fig. 2A**, brown node). A DUF262 gene was found in the former genome neighborhood that is found in the SspE restriction component of Ssp-mediated PT systems and is involved in the PT-sensing anti-phage activity[7].

The *sspD* gene neighborhoods were less complicated than those involving *dndC* (**Fig. 2B**). Again, *sspD* genes were associated with BREX defense system genes, as in the small cluster containing *Cyanobacterium stanieri* (**Fig. 2B**, yellow node). This reinforces the idea that BREX type 4 genes represent a PT-based defense system.

To compare the distribution of *brx* genes to *dnd* and *ssp* families in prokaryotes, we performed a BLASTp search with protein sequences for DndABCDE, SspABCDE and BrxPCZL as queries in 6,616 representative genomes from the Bacterial and Viral Bioinformatics Resource Center (BV-BRC; as of January 2021)[39]. Here we defined the minimal genes necessary for a functional PT synthesis system as *dndCD*, *sspBCD*, *brxPCZL* based on the following observations: (1) DndA/SspA are often replaced by cysteine desulfurase such as IscS; (2) DndB is a non-essential regulator; (3) DndE protein is too short to search rigorously; (4) SspE is not required for PT modification; and (5) *brxPCZL* are the core 4 genes that define a BREX type 4 system, with the *brxR* regulator not essential. Based on gene locus information for each protein hit, we found the *dndCD*, *sspBCD*, and *brxPCZL* gene clusters present in 4.3%, 3.0%, and 0.6%, respectively, of the BV-BRC genomes (**Supplementary Data 3**). Notably, both *dndCD* and *sspBCD* gene clusters co-occurred in the same genomes of some Cyanobacteria, such as *Gloeocapsa* sp., *Coleofasciculus chthonoplastes*, and *Scytonema hofmanni*, among others (**Supplementary Data 3**). This complements the gene neighborhood analysis, which revealed that *dndC* clusters containing *brx* genes, MNT-HEPN genes, and DUF3696 genes are located near *dndBCD* operons in Cyanobacteria such as *Fischerella, Nodosilinea,* and *Pseudanabaena* (**Fig. 2A**, red-outlined yellow, green, and blue nodes in the main cluster). Similarly, 7 of 40 genomes with *sspBCD* operons located near *brxCZL* genes occurred in some Cyanobacteria genomes (**Fig. 2B,** red-outlined yellow nodes in the main cluster). These observations complement the observations of Lin *et al*.[9] and suggest the co-evolution of the variety of PT-based epigenetic systems in the ancient photosynthetic Cyanobacteriota phylum.[40, 41]

**BREX type 4 systems are homologous to Ssp proteins and catalyze PT synthesis**

The genome neighborhood analyses revealed strong associations between the PAPS domain and NTPase genes essential for PT synthesis and *brx* genes from the BREX type 4 family. This association was strengthened by the homology analysis (HHpred[42]) of BREX family and SSP proteins shown in **Figure 3A**, which reiterates the hallmark *brxPCZL* genes in BREX type 4 and the lack of the PT-defining PAPS domain and

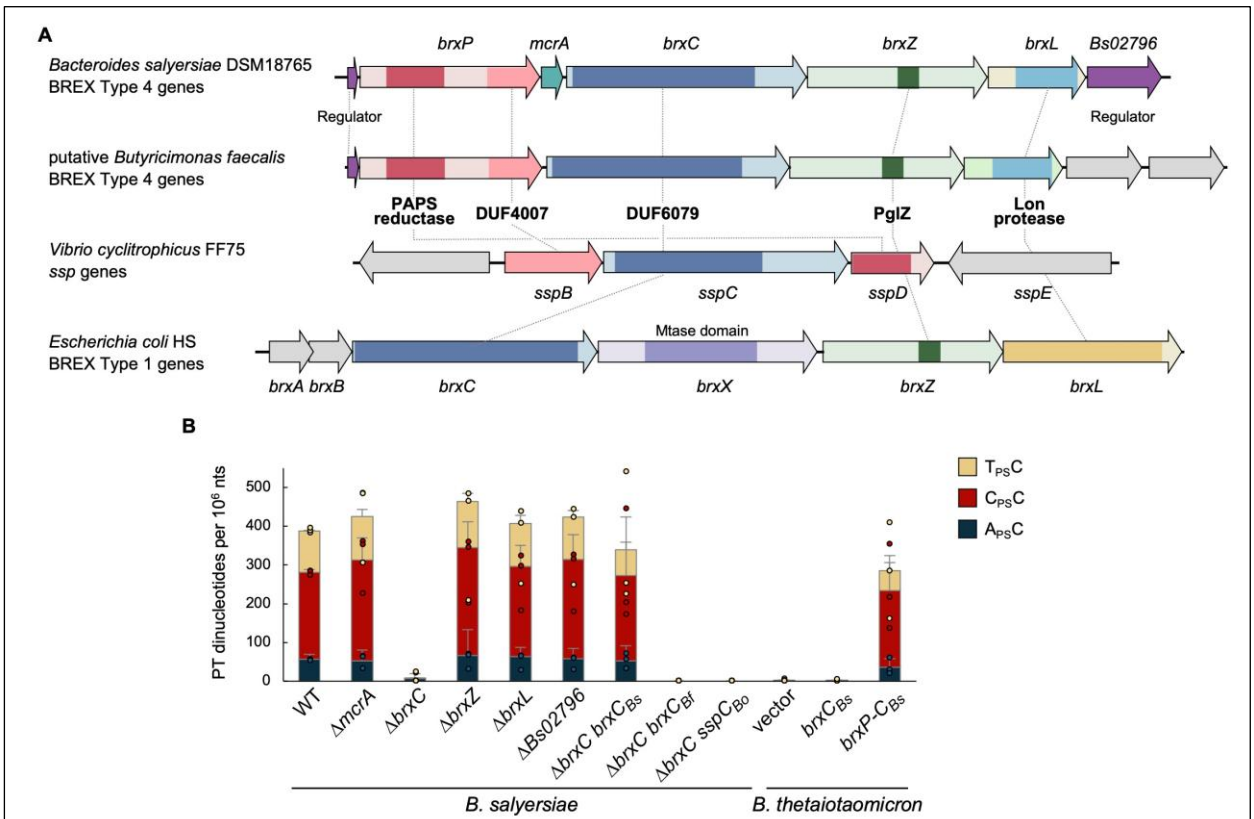


**Figure 3. Homology analysis of BREX type 4 systems and evidence of PT synthesis.** (**A**) The genomic organization of the BREX type 4 gene systems in *Bacteroides salyersiae* DSM18765 and putative *Butyricimonas faecalis* (GMbC ID 5893AJ_0218_015_F2). *Vibrio cyclitrophicus* FF75 is included to demonstrate a typical *ssp* gene system and *Escherichia coli* HS to demonstrate a typical BREX type 1 system. Protein domains are color-coded and labeled. (**B**) The levels of PT dinucleotides in engineered *B. salyersiae* and *Bacteroides thetaiotaomicron* strains. *ΔbrxC, ΔbrxC brxC*$_{Bf}$, *ΔbrxC* p*sspC*$_{Bo,}$ vector, and *brxC*$_{Bs}$ differ significantly from wild-type (WT) based on a two-tailed Student's *t*-test with three biological replicates ($p < 0.05$). Data represent mean values ± SD. Bs: *B. salyersiae*, Bf: *Butyricimonas faecalis*, Bo: *Bacteroides*

NTPase genes in BREX type 1. For example, BrxP possesses both the PAPS reductase domain found in SspD and the DUF4007 domain found in SspB (**Fig. 3A**), which effectively renders BrxP as a SspD-SspB fusion protein. Similarly, BrxC is essentially an SspC homolog, both possessing the same DUF6079 domain (**Fig. 3A**). Based on these similarities, we tested the PT synthesis activity of proteins encoded by genes in the BREX type 4 family. The first step was to identify PT modifications in bacterial strains known to possess BREX type 4 genes. Here we quantified PTs as PT-linked dinucleotides by liquid chromatography-coupled triple-quadrupole mass spectrometry (LC-MS/MS) in a limited digest with nuclease P1 and phosphatase dephosphorylation[43]. In *Bacteroides salyersiae* DSM18765, which harbors a 7-gene *brx*

operon but no *dnd* or *ssp* genes (**Fig. 3A**), we detected the PT dinucleotides $A_{PS}C$, $C_{PS}C$, and $T_{PS}C$ at 56, 228, and 98 per $10^6$ nt, respectively (**Fig. 3B, Supplementary Fig. S1A, Supplementary Data 4**). As shown in **Supplementary Data 4**, several other bacterial isolates with a 4-5 gene complement of *brx* genes (*brxCLPRZ, brxCLPZ*) and insufficient sets of *dnd* or *ssp* genes also yielded $C_{PS}C$ (*Prevotalla sp*.), $A_{PS}A$ and $A_{PS}C$ (putative *Butyricimonas*), $T_{PS}C$ (*Parabacteroides*), and $A_{PS}A$ (*Bacteroides*). These results clearly distinguished the BREX type 4 gene systems from *dnd* and *ssp* families.

Building on this associative evidence, we validated PT synthesis activity by creating in-frame deletion mutants of *brx* genes in *B. salyersiae*. As shown in **Figure 3B**, loss of *brxC* abolished PTs, with PTs restored by complementation with *in trans* expression of *B. salyersiae brxC* (Δ*brxC* $\underline{brxC_{Bs}}$). Based on the similarity between BrxP and SspD, we attempted to create a *brxP* deletion mutant with *B. salyersiae* but were unable to obtain this mutant, suggesting that its deletion may be deleterious. As an alternative, we reconstructed the PT pathway from *B. salyersiae* by cloning $brxP_{Bs}$, $mcrA_{Bs}$ (a PT-dependent restriction enzyme in *Streptomyces coelicolor*)[44], and $brxC_{Bs}$ together in an expression plasmid or $brxC_{Bs}$ alone in the plasmid. These plasmids were transferred into *Bacteroides thetaiotaomicron* VPI, which lacks PT genes but possesses a SufS cysteine desulfurase homolog of DndA to provide a presumably complete system for PT modification (SufS, BrxP, BrxC) and restriction (McrA). The combination of $brxP_{Bs}$, $mcrA_{Bs}$ and $brxC_{Bs}$ resulted in the $T_{PS}C$, $C_{PS}C$, and $A_{PS}C$ dinucleotides in the same proportion as wild-type *B. salyersiae* $brxC_{Bs}$ alone was not able to confer PTs, which is consistent with the essentiality of BrxP in PT synthesis. Furthermore, *mcrA*, *brxZ, brxL* or an unknown *Bs02795* gene were not required for PT synthesis as shown in corresponding mutants (**Fig. 3B**). To test the generality of the PT synthesis by BrxCP, determine if *brx*C and *ssp*C are interchangeable, and, since *B. salyersiae* and *B. faecalis* have different PT motifs, assess if BrxC is the component that determines the target motif, we heterologously expressed $brxC_{Bf}$ from *B. faecalis* or $sspC_{Bo}$ (*brxC* equivalent) from the human gut isolate *B. ovatus* CL03T12C18 *in trans* in the *B. salyersaie brxC* mutant. PTs were not detected in either strain (**Fig. 3B**). These results suggest species-specific complexities of the BREX system components requiring further

investigation. We conclude that, in presence of a cysteine desulfurase gene (*sufS*), *brxP* and *brxC* are the minimal set of BREX genes needed for PT synthesis.

**The distribution of PT modifying systems in the human gut microbiome**

Based on the observation of PT-containing microbes in the mouse and human gut microbiome[23, 24], we wondered about the presence of *brx*-based PT systems in gut bacteria and their relationship to *dnd* and *ssp* systems. To quantify the distribution of the PT systems in human gut microbiome, we searched for PT gene clusters in 13,663 human gut microbial genomes from the BIO-ML[25] and the GMbC[26-28] collection, and from isolate sequences of human gut microbes and metagenome-assembled genomes (MAGs) in the Unified Human Gastrointestinal Genome (UHGG) collection[29]. As with the BV-BRC searches noted earlier, we performed a BLASTp search of proteins DndABCDE, SspABCDE and BrxPCZL as queries in these 13,663 genomes (**Supplementary Data 5**). The essential gene sets *dndCD*, *sspBCD*, *brxPCZL* were found to be present in 2.7%, 3.6%, and 1.4% of the human gut microbiome genomes, respectively (**Fig. 4**). This represents a 1.6- and 1.2-fold reduction in *dnd* and *ssp* systems and a 2.3-fold enrichment in *brx* systems in the gut microbiome compared to the general BV-BRC genomes. The distributions of the three gene clusters at the genome level were nearly exclusive with few exceptions (**Supplementary Data 5**). Among the most prevalent phyla and orders in the human gut microbiome, the PT-modifying gene clusters were mainly found in Bacteroidota (Bacteroidales), Bacillota (Clostridiales), Pseudomonadota (Enterobacterales), and, to a lesser extent, Actinomycetota (Coriobacteriales) (**Fig. 4**). These observations raised the question of the predictive power of the genomic analyses.

**Validating the predicted PT modifications in gut microbiome isolates**

Given the presence of genes essential for PT synthesis in ~1,000 out of 13,663 gut microbiome isolates, we next used LC-MS to validate the presence of PT dinucleotides in DNA in a collection of 226 bacterial isolates possessing *dndCD*, *sspBCD*, *brxPCZL* gene neighborhoods or individual PT synthesis genes not predicted to lead to PTs

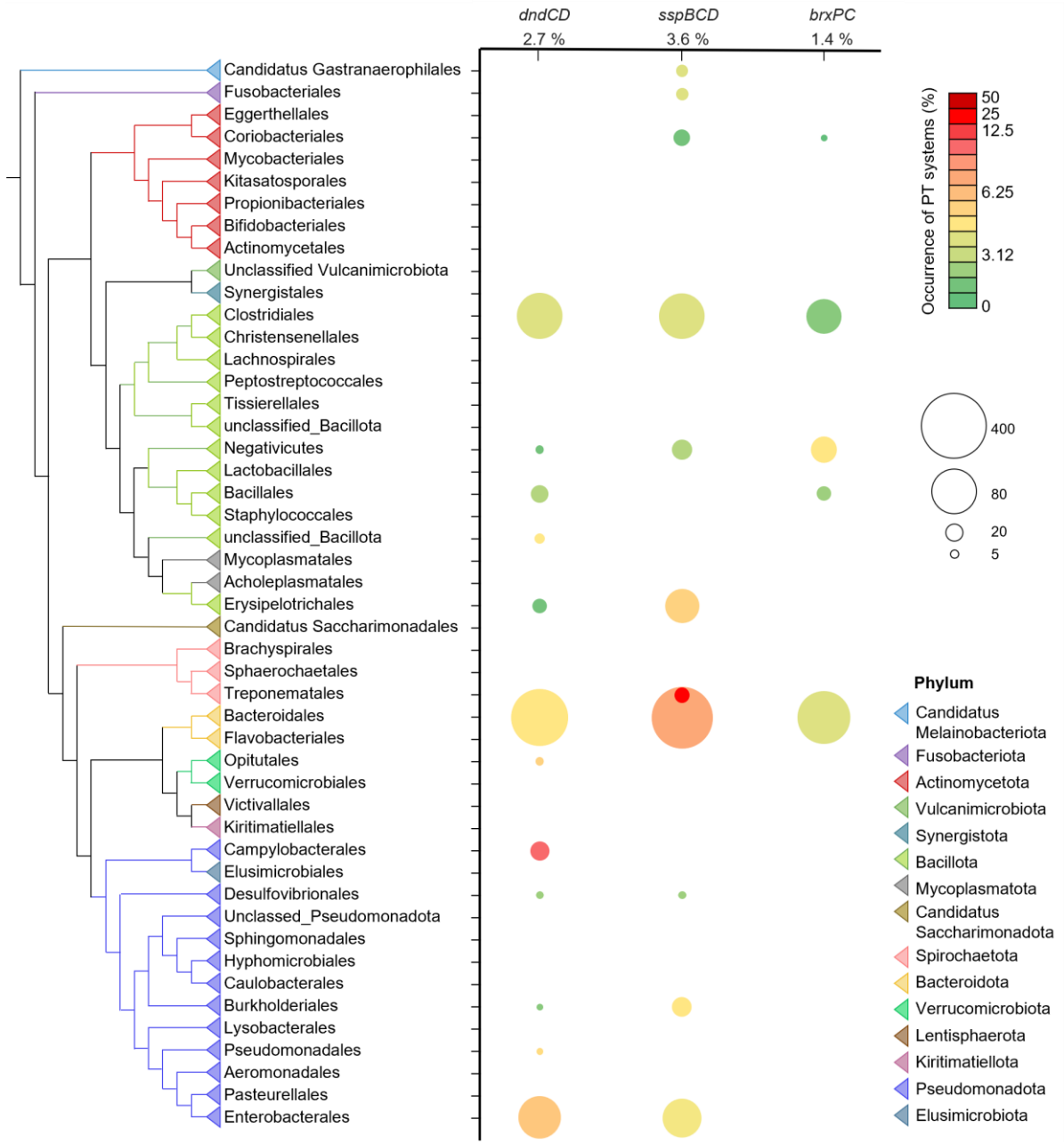


**Figure 4**. **Phylogenetic distribution of *dndCD, sspBCD, and brxPC* genes in human gut microbiome genomes.** The reference phylogeny was reconstructed from the concatenated alignment of 10 ribosomal proteins. The colors of the triangles in the tree show the taxa at the order level. The occurrence of gene clusters was quantitatively presented by the colors and size of circles. For better visualization, orders containing less than 5 genomes are hidden. Source data are provided in **Supplementary Data 5**.

(**Supplementary Data 4**). In total, we identified 8 PT dinucleotides, including $A_{PS}A$, $A_{PS}C$, $C_{PS}A$, $C_{PS}C$, $G_{PS}A$, $G_{PS}C$, $G_{PS}T$, and $T_{PS}C$, and did not detect an additional two dinucleotides ($C_{PS}T$, $G_{PS}G$) that we found in human fecal DNA samples (**Supplementary Data 4**)[24]. Previous studies in three bacterial species with *dnd* and *ssp* genes[4, 43, 45] revealed $G_{PS}A$ and $G_{PS}T$ at >500 per $10^6$ nt and $C_{PS}C$ at >2000 per $10^6$ nt, with $C_{PS}A$, $A_{PS}A$, $A_{PS}C$, and $T_{PS}C$ detected much lower levels of 1-6 per $10^6$ nt[4, 43, 45]. The minor PT dinucleotides represent low-affinity binding sites for the DNA shape-selective Dnd proteins[45]. In sharp contrast, however, these minor sites in gut microbiome bacteria represent major PT modification sites, as high as 1,800 per $10^6$ nt in bacteria with *brxPCZL* genes (**Supplementary Data 4**).

The $A_{PS}A$, $A_{PS}C$, $C_{PS}A$, $C_{PS}C$, $G_{PS}A$, $G_{PS}C$, $G_{PS}T$, and $T_{PS}C$ dinucleotides were distributed unevenly among the isolates based on the type of PT modification system, with four groups emerging from the analyses (**Fig. 4**). The first group of 39 isolates was characterized by $G_{PS}A$ and the predominance of *dnd* genes. The largest portion of Group 1 (32 isolates) possessed both $G_{PS}A$ and $G_{PS}T$ and harbored *dndCD* genes with or without *dndBE*, including isolates from Bacteroidales and Clostridiales orders. Five isolates from both Bacteroidales and Clostridiales harboring the minimal *dndCD* gene set showed $G_{PS}A$ and $G_{PS}C$. One isolate that possessed a solitary *brxP* but no *dnd* genes and possessed both $G_{PS}A$ and $G_{PS}T$. Given the presence of different dinucleotides for the *ssp* and *brx* gene families, this latter observation could be explained by genome sequencing errors, or a sample mix up during regrowth, either resolved by re-sequencing. A single *Blautia* species harboring *dnd* genes showed only $G_{PS}A$ (**Supplementary Data 4**).

The second group of 34 isolates was characterized by the $C_{PS}C$ dinucleotide and *ssp* genes, with all but 2 harboring *sspBCD* ±*sspE* (**Fig. 5**). One isolate harbored *dndCD* genes that are not proximal to the *sspBCD* operon and showed LC-MS evidence of low levels of $G_{PS}A$ and $G_{PS}T$ at the detection limit of ~2 PTs per $10^6$ nt (**Supplementary Fig. S2**), though these signals await high-resolution mass spectrometric validation. One isolate in this group lacks *sspBC* genes and one lacks all essential *sspBCD* genes.

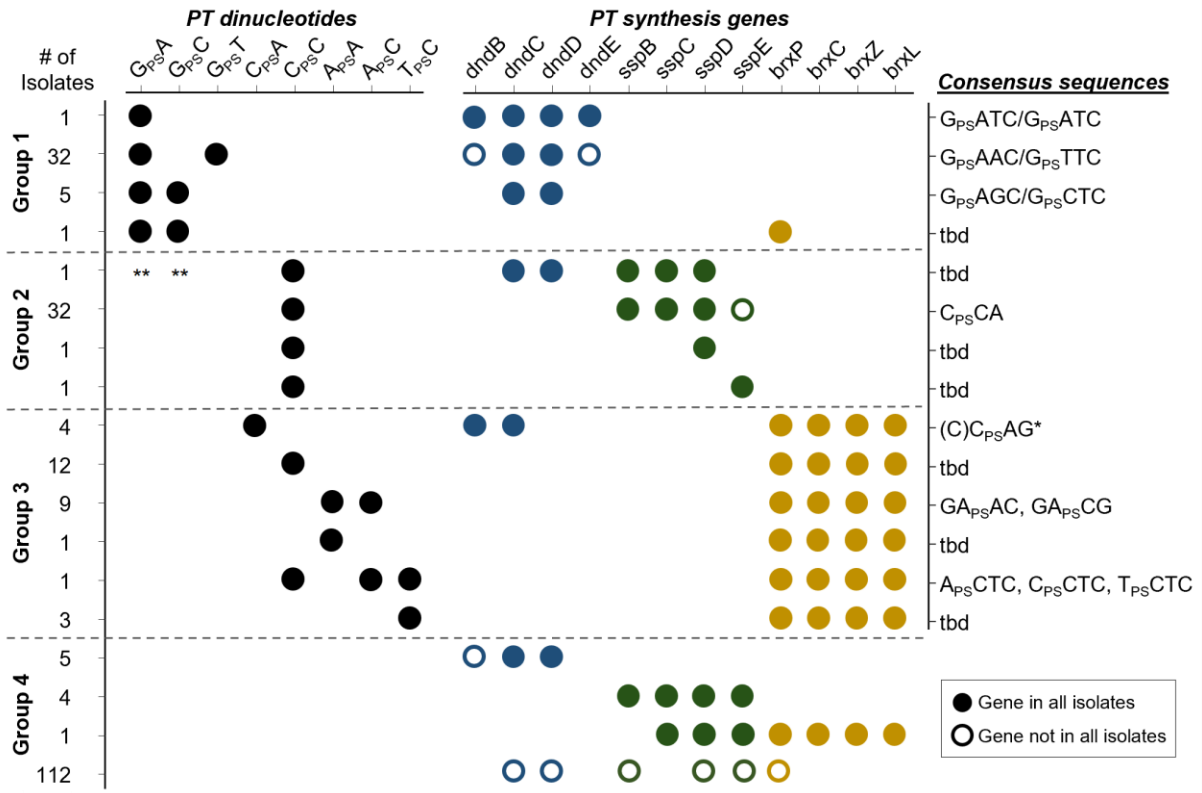


**Figure 5**. **PT dinucleotides and PT consensus sequences in human gut microbiome isolates.** The numbers of isolates were analyzed are listed on left. The circles represent PT dinucleotides quantified by LC-MS (left) and the presence of corresponding genes (right). The PT-modified consensus motifs were characterized in one representative isolate from each group using PT-seq. tbd, to be determined. *, the $C_{PS}AG$ motif was characterized using metagenomics PT-seq in another on-going study. **, $G_{PS}A$ and $G_{PS}T$ were detected using LC-MS/MS but the exact mass was not verified by high-resolution mass spectrometry.

Again, the consistency of dinucleotide distributions suggests that *sspBCD* or *brxPCZL* genes are present in the genomes and perhaps missed due to insufficient genome sequencing coverage.

The third group of 30 isolates carried the BREX type 4 genes *brxPCZL* and the most diverse sets of PT dinucleotides. Four isolates of *Parabacteroides* contain $C_{PS}A$, 12 isolates of *Prevotella* sp. contain $C_{PS}C$, 3 isolates of *Parabacteroides* contain $T_{PS}C$, and 9 isolates of putative *Butyricimonas faecalis* contain $A_{PS}A$ and $A_{PS}C$ (Fig. S1B). The *B. salyersiae* strain examined earlier, which contains $C_{PS}C$, $T_{PS}C$, and $A_{PS}C$ (**Fig. 5, Supplementary Data 4**).

The fourth group of 122 bacteria lacked reliable LC-MS signals for PT dinucleotides. The majority (83) lacked the minimal sets of *dndCD*, *sspBCD* or *brxPC* gene clusters. For example, *Bacteroides dorei* CL03T12C01 harbors a PAPS reductase-encoding gene next to an ATPase without an adjacent *dndD* (**Supplementary Fig. S3**). Several isolates harbor *dndD* and *dptH* next to a methylase and a restriction enzyme but lack *dndC* (**Supplementary Fig. S3**). The absence of PT dinucleotides in these isolates agrees with the requirement for a PAPS reductase domain-containing protein and an NTPase. However, several (10) carry the minimal set of *dndCD* or *sspBCD* genes (**Fig. 4**), so PT dinucleotides were expected. It is possible but unlikely that the lack of PT dinucleotides detected in the isolates is due to a symbiont phenotype in the gut microbiome, with loss of one of the PT synthesis genes accounted for by uptake of PT intermediates from other bacteria. The PT modification pathway does not require unique small-molecule intermediates and the modification genes are generally not interchangeable among different bacteria, so it is unlikely that gut microbes can lose one or more PT modification genes and exist as a “PT symbiont” in the gut community. Again, we cannot rule out errors in the reference genomes due to genome sequencing and sample handling.



Although based on a limited number of samples, these observations suggest the hypothesis that (1) Dnd proteins produce G$_{PS}$A, G$_{PS}$C, and G$_{PS}$T, (2) Ssp proteins produce C$_{PS}$C, and (3) Brx proteins produce everything except G$_{PS}$A, G$_{PS}$C, and G$_{PS}$T. Understanding the basis for these differences requires knowledge about the longer consensus sequences containing the PT dinucleotides in the gut microbes.

**Novel PT consensus sequences in gut microbiome isolates**

Here we defined novel larger PT consensus sequences, especially for the BREX type 4 system, in gut microbiome isolates using an innovative and highly sensitive NGS sequencing technique, PT-seq[30]. The idea here is that the PT dinucleotides such as G$_{PS}$A and G$_{PS}$T, are found in the larger consensus sequences of G$_{PS}$ATC and G$_{PS}$AAC-3'/5'-G$_{PS}$TTC in several types of bacteria[5]. Application of PT-seq (**Supplementary Fig.**

**S4**) to the BREX type 4-containing human gut microbiome isolate *B. salyersiae* DSM18765, earlier found by LC-MS/MS to possess $A_{PS}C$, $C_{PS}C$, and $T_{PS}C$ at 62, 228, and 98 per $10^6$ nt (**Supplementary Data 4**), showed 5459 PT sites: 56 at $A_{PS}CTC$, 3881 at $C_{PS}CTC$, and 1513 at $T_{PS}CTC$ sites, respectively (**Fig. 6A**, **Supplementary Data 6**). As observed with other *dnd* and *ssp* systems, the Brx proteins modified only a portion of the 27274 ACTC (0.2%), 22698 CCTC (17%), and 36290 TCTC (4%) total sites available, respectively. The observation of PTs at 3 NCTC sites, with a strong preference for CCTC, is consistent with previous observations that Dnd proteins select their target sequences based on DNA shape rather than precise binding contacts[45]. We appreciate the discrepancy between LC-MS/MS and PT-seq data for the quantity of $A_{PS}C$ in the $A_{PS}CTC$ context in *B. salyersiae* DSM18765. We previously observed that

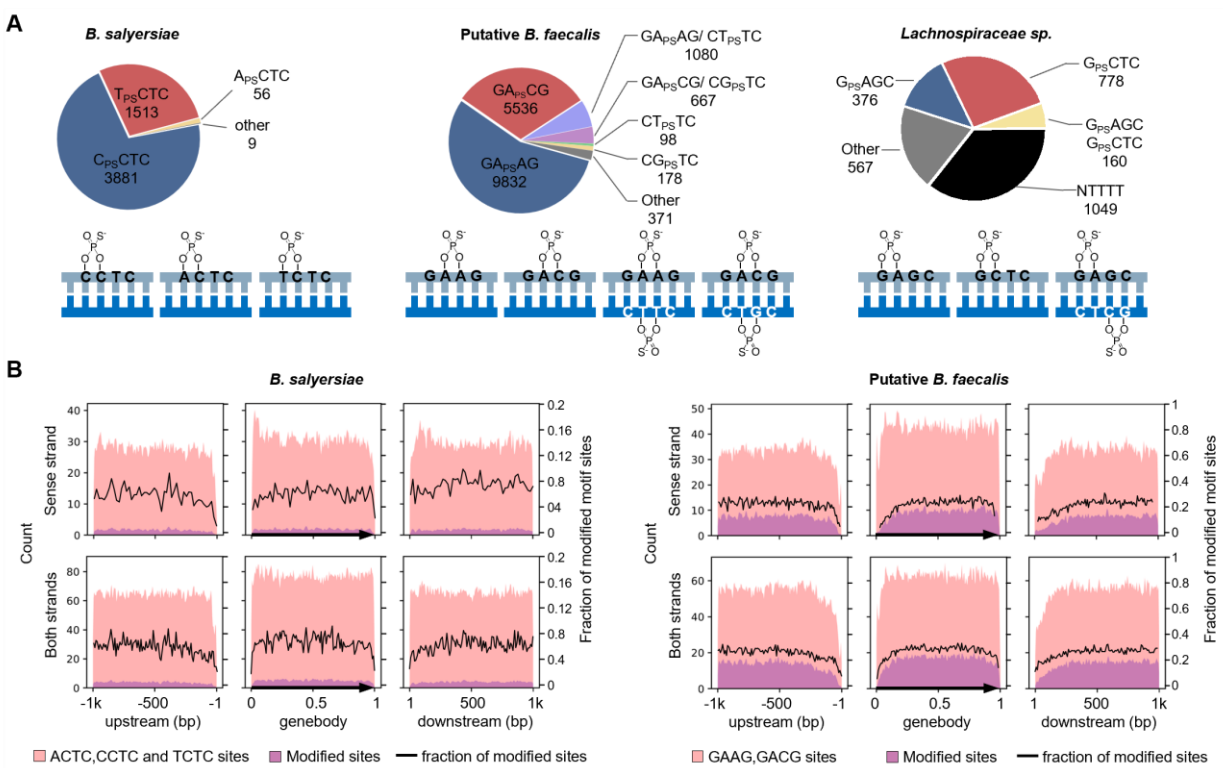


**Figure 6**. **Biogeographical maps of PTs in bacteria with BREX type 4 systems.** (**A**) Pie charts depict the number of single- or bi-stranded PT-modified consensus sequences. The structural diagrams depict these motifs. (**B**) Analysis of the distribution of PT consensus sequences and modified sites in 1kb upstream and downstream regions in *B. salyersiae* (**left**) and putative *B. faecalis* (**right**). The number of total motif sites in the sense strand (**upper**) or both strands (**lower**) are represented in pink. The PT modified sites are represented in purple. The fraction of PT-modified motif sites is represented by the black line.

the efficiency of iodine-induced cleavage of PT sites to strand breaks in the PT-seq method varies by 2-fold for different dinucleotides[30] and that the extension efficiency of the terminal transferase used for T-tailing depends on the nucleobase sequence of the initiators, where A-ending initiators are generally less favored[46-48]. Together, these could account for some portion of the discrepancy between the LC-MS and PT-seq data for ApsC. However, in another BREX type 4-containing putative *B. faecalis* (GMbC ID 5893AJ_0218_015_F2), PT-seq revealed 9832 GA$_{PS}$AG and 5536 GA$_{PS}$CG (**Supplementary Data 7**). These sites agreed with the A$_{PS}$A and A$_{PS}$C dinucleotides detected by LC-MS/MS (**Supplementary Data 4**). It thus remains to be determined if the A$_{PS}$CTC motif is uniquely problematic for PT-seq. Finally, PT-seq also detected sites with bistranded PTs: 1080 GA$_{PS}$AG/CT$_{PS}$TC and 667 GA$_{PS}$CG/GC$_{PS}$TC. However, we could not detect reliable signals for the corresponding T$_{PS}$T and T$_{PS}$G dinucleotides by LC-MS/MS, which may be due to their low abundance and to insensitive MS detection of T-containing nucleotides.

Application of the original PT-seq protocol without biotin labeling and capturing to a *dnd* system-containing human gut microbiome isolate, *Lachnospiraceae sp.* (GMbC ID 2807EA_1118_063_H5), revealed 1,474 G$_{PS}$AGC and G$_{PS}$CTC sites occurring among the 16,683 possible GAGC/GCTC sites, with 1,154 modified on one strand and 160 modified on both strands (**Fig. 6A, Supplementary Data 8**). A large fraction of pileup sites occurring at NTTT and other motifs represents artifacts of poly(dT) tailing at single strand nicks by terminal transferase.

**The uneven genomic distribution of PTs**

In addition to defining the consensus sequences for BREX type 4 and other PT synthesis systems, PT-seq also revealed that PTs are nonrandomly distributed across bacterial genomes, which agrees with single-molecule real-time (SMRT) sequencing maps in *Escherichia coli*[13]. For example, the percentage of gene classes and intergenic regions containing PTs in *Lachnospiraceae sp., B. salyersiae, and B. faecalis* varied significantly: 6% of intergenic regions, 13% of tRNA genes, and 58% of coding sequences (CDS) (**Supplementary Data 9**). The percentage of consensus sequences

modified with PTs, which is roughly 10% in genomes studied to date[12, 13], also varied in individual bacterial genomes, ranging from 1% in tRNA genes in *Lachnospiraceae sp.* to 40% in rRNA genes in *B. faecalis* (**Supplementary Data 9**). Interestingly, the consensus sequences recognized by PT synthesis proteins were underrepresented at the boundaries of CDS (**Supplementary Fig. S5**). We further assessed the distribution of PTs within and between CDSs by quantifying PT-modified and unmodified consensus sequences in the sense strand of gene bodies and in regions 1 Kbp upstream and downstream of all CDSs. For both *B. faecalis* and *B. salyersiae,* this analysis revealed that the percentage of PT-modified consensus sequences was lower at the boundaries of coding regions (**Fig. 5SB**, black line; **Supplementary Fig. S6**). Published PT-seq data for *E. coli* BW25113, which has a bistranded G$_{PS}$AAC/G$_{PS}$TTC PT consensus, also showed this biased distribution of PTs (**Supplementary Fig. S7A**). Thus, both the consensus sequences for PTs and the proportion modified with PT were underrepresented on either side of CDSs. While PTs have been proposed to interfere with transcription[13], the percentage of PT-modified consensus sequences did not correlate with the level of transcription activity (**Supplementary Fig. S7B**), nor did the number of PTs in possible promoter regions (**Supplementary Fig. S7C**). The basis for these biased distributions remains to be defined.



## Discussion

Here we applied systems-level informatic, mass spectrometric, and sequencing tools to discover and characterize new PT modification systems in bacteria, with a focus on gut microbes. A rigorous neighborhood analysis of the key PT synthesis genes in *dnd* and *ssp* systems led to the discovery of several potential previously undescribed PT modification systems, including MNT-HEPN, DUF262, and BREX type 4 gene gene families. Genetic and bioanalytical validation of the BREX type 4 system revealed the essentiality of *brxPC* genes for PT synthesis in bacteria with *brxPCZL* clusters. With the discovery of the *dnd* system in 2005[3], the *ssp* system in 2020[7], and the *tdp* system in 2025[11], the *brxPCZL* cluster represents the fourth established PT modification system. While the BREX type 4 system has been shown to have anti-phage activity[35], such activity for the variant type 4 systems observed here awaits investigation.

The extent of distribution of the *dnd*, *ssp*, and *brx* gene clusters was assessed first among 6,616 representative prokaryotic genomes and then among 13,663 gut microbiome isolates. The difference in the frequency of *dndCD*, *sspBCD*, *brxPCZL* gene clusters in the BV-BRC general set of bacteria (4.3%, 3.0%, and 0.6%, respectively) and the gut microbiome isolates (2.7%, 3.6%, and 1.4%, respectively) suggests a bias for the BREX type 4 systems in the gut environment. The total of 7.7% of gut microbiome isolates possessing PT-synthesis genes is consistent with the estimate of 5-10% of stool microbes possessing PT modifications, as we observed using mass spectrometric measurement of PT dinucleotides in human fecal DNA[24].

With the discovery of BREX type 4 as the third PT synthesis gene family, one feature of PT modifications appears to be universal for PT epigenetics: partial modification of available consensus sequences in individual genomes. In agreement with previous sequencing studies with bacteria containing Dnd proteins[13, 49], PT-seq analysis revealed that the BREX type 4 proteins also only partially modify their four-nucleotide consensus sequences, ranging from 0.2% at ACTC to 17% at CCTC in *B. salyersiae* (**Fig. 5A, Supplementary Data 6**). Partial modification of all available sequence motifs, hemi-modification within a sequence motif, cell replication-dependent shifts in the locations of modified motifs, and active maintenance of a constant density of PT modifications have all been observed in PT systems identified to date[5, 13, 19, 45], despite the presence of active restriction components[13]. While some portion of the hemi-modification and partial modification could result from naturally occurring oxidation and repair or replacement[19], the unusual behavior of all PT systems to date appears to be inherent to the mechanisms of modification target selection and surveillance for restriction. The BREX type 4 system also highlights another feature of PT systems: sequence-selectivity of the *dnd, ssp*, *tdp*, and *brx* gene families. Previous[7, 8, 50] and present studies (**Fig. 4**) consistently show that *ssp* systems catalyze C$_{PS}$C dinucleotides mainly in CCA motifs, while the *dnd* systems insert PTs mainly in GN motifs: G$_{PS}$A, G$_{PS}$C, G$_{PS}$G[43], and G$_{PS}$T dinucleotides. The latter occurs mainly as bistranded modifications of GNNC motifs. To date, a *tdp* consensus of G$_{PS}$ATC has been observed in a few bacteria[11]. We previously

showed that Dnd proteins select modification sites based on DNA shape, with GAAC, GTTC, and GATC all sharing similar shapes and all being modified with PTs at G$_{PS}$A and G$_{PS}$T in *S. enterica*[45]. The fact that *ssp* is selective for CCA suggests a similar shape selectivity. BREX type 4-dependent PT systems are more complicated, which may reflect the hybrid nature of the PT-synthesizing gene families in different bacteria. The five types of PT nucleotides produced by Brx proteins (C$_{PS}$A, C$_{PS}$C, A$_{PS}$C, A$_{PS}$A, T$_{PS}$C) share C$_{PS}$C with *ssp* systems but lack the G$_{PS}$N unique to *dnd* systems (**Fig. 5A**). This may be partly explained by the sequence similarities between SspBCD and BrxPC proteins. The shape selectivity argument is again supported with *brxPCZL* in *B. salyersiae* DSM18765, with the ACTC, CCTC, and TCTC sites in an NCTC motif all modified to differing extents (**Fig. 5A**) and in *B. faecalis* with GAAG and GACG motifs. However, the limited number of strains analyzed in the present studies and published work may have missed *dnd*, *ssp,* or *brx* systems that confer other dinucleotide patterns.

Our studies revealed the complexity of interactions among the three PT systems, with widespread combinations of the various components of all three systems and coexistence of two or more gene clusters in the same organism (**Fig. 1**, **Supplementary Data 3**). These evolutionary co-occurrences likely reflect a fitness advantage, such as that observed for the presence of Dnd and Ssp together, which provides complementary and synergistic protection against temperate and lytic phages as well as phage induction[50]. The co-occurrence of combinations of all three established PT systems and two putative systems in Cyanobacteria supports the hypothesis of an evolutionary divergence that may originate from a sulfur-based metabolism in ancient Cyanobacteria ancestor[9]. We found that 7 strains harboring the *dndBCD-brxCZL* type gene islands, while 7 out of 40 strains harboring the *sspBCDB-brxCZL* type (or similar) gene islands are Cyanobacteria and 3 of 40 are Deinococcota. Phylogenetic analyses suggest that Deinococcota and Cyanobacteria are both relatively close to the root of the bacterial tree of life[51], which agrees with the argument that *dnd* systems originated in ancient Cyanobacteria after the Great Oxygenation Event[9]. Similarly, BREX systems have also been found in branches near the root of the bacterial tree and deep branches[35]. All four PT systems contain a gene encoding a PAPS reductase domain

protein that has been proposed to involve in the initial sulfur mobilization step[5]. The observation of a bias toward *brx*-based PT systems and the unique biochemical properties of PTs raise questions about the role of PT epigenetics in the human gut microbiome in health and disease.

## Methods

**Bacterial strains and growth conditions.** Bacteria revived from BIO-ML and GMbC were cultured as previously decribed[25]. *Parabacteroides* spp. and *Bacteroides* spp., *E. faecalis*, and BIO-ML isolates were grown on ASF plates (Becton Dickinson) and BHIS plates[52], BHI plates (Becton Dickinson), and Brucella agar, (Catalog # AS-141, Anaerobe Systems, Morgan Hill, CA), respectively. Culture manipulations were performed at 37 °C in an anaerobic chamber with an atmosphere of 80% nitrogen, 5% carbon dioxide, and 5% hydrogen. Growth on agar plates was harvested with phosphate buffered saline (Catalog # 10010-023, Gibco, Paisley, PA). The bacterial suspensions were pelleted by centrifugation at 4,000 g for 15 min at ambient temperature. The supernatant was removed, and the pellets were stored at -20 °C.

Bacterial stains and plasmids used and generated for BREX 4 system deletion and reconstruction are listed in **Supplementary Table S1**. Bacteroidales strains were grown in basal liquid medium[53] or on BHIS plates[52]. Antibiotics used for selection include erythromycin (10 µg/mL), gentamicin (200 µg/mL), anhydrotetracycline (50 ng/mL). *E. coli* S17 λ pir was grown in LB broth or plates with carbenicillin (100 µg/mL) added for selection.

**Creation of deletion mutants and complementing clones.** Internal non-polar deletion mutants of *brx* genes, including *mcrA* (HMPREF1532_02792), *brxC* (HMPREF1532_02793), *brxZ* (HMPREF1532_02794), *brxL* (HMPREF1532_02795), and *Bs02796* (HMPREF1532_02796), were constructed by amplifying DNA upstream and downstream of each gene using the synthetic primers listed in **Supplementary Table S2**. These flanking pieces were cloned into BamHI-digested pLBG13[52] using

NEBuilder (New England BioLabs) and transformed into *E. coli* S17 λ pir. PCR confirmed plasmids were sequenced (Plasmidsaurus) to confirm correct error-free PCR amplification. The correct construct was conjugally transferred from *E. coli* into *B. salyersiae* and cointegrates were selected on gentamycin/erythromycin. Double recombination cross-outs were selected on BHIS plates with anhydrotetracycline (aTC, 50 ng/ml) and screened via PCR for mutant genotype.

Genes expressed *in trans* in *B. thetaiotaomicron* and *Bacteroides* Δ*brxC* were PCR amplified and cloned into BamHI-digested pFD340[54] using NEBuilder v2.5. Transformants were PCR screened and the plasmid was sequenced. Plasmids were conjugally transferred from *E. coli* to *B. salyersaie* and transconjugants were selected on gentamycin/erythromycin plates.

**Sequence Similarity Networks.** Sequence similarity networks (SSNs) were generated by submitting the sequences of *E.coli* B7A DndC (GeneBank AIF62362.1) and *V. cyclitrophicus* FF75 SspD (NCBI Refseq WP_016789110.1) to the EFI-EST webtool v2021_03[31] using the BLAST option (E-value cutoff $10^{-5}$, maximum number of sequences retrieved 5000). The initial SSN was generated with an alignment score cutoff set such that each connection (edge) represents a sequence identity of approximately 40%. Sequences that share 100% sequence identity were grouped into a single node. More stringent SSNs were created by increasing the alignment score cutoff in small increments (usually by 5-10). This process was continued until each cluster is estimated to be isofunctional, in which the nodes represent enzymes that catalyze the same reaction. Isofunctionality was determined by mapping the known *dnd* clusters and following their movements as the alignment cutoff increased until the *dnd* clusters fell into subclusters. The network were visualized in alignment score weighted Prefuse Force-Directed Layout using Cytoscape v3.9[55]. The resulted SSNs were submitted to EFI-EST webtool[31] for genome neighborhood analysis with default setting. Nodes were highlighted in SSNs if the neighboring Pfam families, with maximal median distance of no more than 4 and minimal co-occurrence rate larger than 0.2, have NTPase activity, NTP binding, or DNA binding activities.

**Sequence analyses.** For sequence analyses, the BLAST tools v2.9[56] (with E-value cutoff of $10^{-10}$ and query coverage of 25%) and HHpred v20200717[42], and the resources of BV-BRC vbeta[39] were routinely used. In total, the data from 20279 bacterial and archaeal genomes were retrieved from the BIO-ML[25], GMbC[26-28], Unified Human Gastrointestinal Genome (UHGG)[29], and BV-BRC representative bacteria as collected in Jan 2021. The queries used in this study were listed in **Supplementary Data 10**, including IscS (GeneBank AIF64277.1), DndB (GeneBank AIF62361.1), DndC (GeneBank AIF62362.1), DndD (GeneBank AIF62363.1), DndE (GeneBank AIF62364.1) in *E.coli* B7A and SspA (Refseq WP_016789103.1), SspB (Refseq WP_022570853.1), SspC (Refseq WP_016789109.1), SspD (Refseq WP_016789110.1), SspE (Refseq WP_016789111.1), and BrxP (GMbC tag OIFBNFKG_02855), BrxC (GMbC tag OIFBNFKG_02856), BrxZ (GMbC tag OIFBNFKG_02857), BrxL (GMbC tag OIFBNFKG_02858). The genes *dndCD*, *sspBCD*, and *brxPCZL* were considered to be present when all genes were adjacent.

**Phylogeny tree.** We first built a concatenated alignment of 10 nearly universal and single-copy ribosomal protein families. We used Diamond v0.8.22[57] (with parameters blastx -more-sensitive -e 0.000001 -id 35 -query-cover 80) to BLAST all proteomes in out collection against the RiboDB database v1.4.1[58] of bacterial ribosomal protein genes. We excluded proteins bL17, bS16, bS21, uL22, uS3 and uS4, as they were not sufficiently distributed across all genomes. In each RiboDB gene family, we excluded genomes that contained gene duplicates. Then, we aligned all protein families individually with MUSCLE v5.1[59] (with default setting). We filtered out misaligned sites using BMGE v1.12[60] (with parameters -t AA -g 0.95 -m BLOSUM30) and concatenated all individual alignments using Seaview v4.7[61]. The phylogenomic tree was reconstructed using FastTree v2.1.10[62] (with parameters -lg -gamma) and visualized and modified in iTOL v5[63].

**DNA isolation.** The cells were resuspended in 500 µL of phosphate buffered saline upon receiving and re-pelleted by centrifugation at 6,000 g for 5 min at 4 °C. Genomic

DNA was isolated using E.Z.N.A Bacterial DNA kit (Catalog # D3350-02), with the bead beating option (speed of 4 m/s, 45 s “on”, 1 min rest, 3 cycles). DNA was eluted with 150-200 µL of RNase free water and stored at -80 °C.

**Digestion of DNA for LC-MS/MS analysis of PT dinucleotides**. DNA (20 µg, 78 µL) was incubated with Nuclease P1 (1.5 U, 3 µL, US Biological) in 30 mM ammonium acetate pH 5.3 and 0.5 mM $ZnCl_2$ (90 µL total reaction volume) for 2 h at 55 °C. The reaction mixture was diluted with Tris-HCl (100 mM final concentration, pH 8.0, 9 µL) and incubated with calf intestinal alkaline phosphatase (51 U, 3 µL, Sigma) for 2 h at 37 °C. Enzymes were removed by passing the mixture through a VWR 10 kDa spin filter with centrifugation at 12,000 g for 12 min. The solution was lyophilized to dryness and resuspended in $H_2O$ (50 µL).

**LC-MS/MS analysis of DNA PT dinucleotides.** Synthetic PT DNA dinucleotides or nuclease P1 hydrolyzed DNA were analyzed by LC-MS/MS on an Agilent 1290 series HPLC system equipped with a Synergi Fusion RP column (2.5 µm particle size, 100 Å pore size, 100 mm length, 2 mm inner diameter) and a DAD. The HPLC was coupled to an Agilent 6490 triple quadrupole mass spectrometer. The column was eluted at 0.35 mL/min at 35 °C with a linear gradient of 3-9% acetonitrile in 97% solvent A (5 mM ammonium acetate pH 5.3) over 15 min. The column was rinsed with 95% acetonitrile in solvent A for 1 min, and then the initial conditions were regenerated by rinsing the column with 97% solvent A for 3 min. Canonical deoxyribonucleosides that eluted from the column were quantified by their 260 nm absorbance with the DAD. PT-containing dinucleotides were identified and quantified by tandem quadrupole mass spectrometry with electrospray ionization operated with the following parameters: N2 temperature, 200 °C; N2 flow rate, 14 L/min; nebulizer pressure, 20 psi; capillary voltage, 1800 V; and fragmentor voltage, 380 V. For product identification, the mass spectrometer was operated in positive ion multiple reaction monitoring mode using the conditions tabulated in **Supplementary Table S3**. No statistical method was used to predetermine sample size. Comparisons between *Bacteroides* WT and engineered strains were based on three biological replicates analyzed using a t-test. For the remaining species,

only one sample was analyzed owing to limited material availability and the most recent technical replicate was included in this study.

**High-resolution mass spectrometry.** Synthetic PT dinucleotides (2 pmol per 10 µL injection) or Nuclease P1 hydrolyzed RNA (4 µg per 10 µL injection) were analyzed on a Dionex Ultimate 3000 UHPLC system equipped with a Synergi Fusion RP column (2.5 µm particle size, 100 Å pore size, 100 mm length, 2 mm inner diameter). The HPLC was coupled to a Thermo Fisher Q Exactive Hybrid Quadrupole-Orbitrap mass spectrometer. The column was eluted at 0.35 mL/min at 35 °C with a linear gradient of 3-9% acetonitrile in 97% solvent A (5 mM ammonium acetate pH 5.3) over 15 min. The column was rinsed with 95% acetonitrile in solvent A for 1 min, and then the initial conditions were regenerated by rinsing the column with 97% solvent A for 3 min. High resolution mass spectra for the PT- containing dinucleotides were obtained by hybrid quadrupole-Orbitrap mass spectrometry with the following parameters: sheath gas flow rate, 50 L/min; aux gas flow rate, 15 L/min; sweep gas flow rate, 3 L/min; spray voltage, 4.20 kV; and capillary temperature, 275 °C. For product identification, the mass spectrometer was operated in positive ion targeted single ion monitoring mode using the conditions tabulated in **Supplementary Table S4**.

**PT-seq library preparation.** PT modifications were mapped in the genomes of *Lachnospiraceae sp.*, *B. faecalis,* and *B. salyersiae* by PT-seq as described elsewhere[29] (**Supplementary Fig. S4**). DNA (10 µg) was diluted to 30.5 µL in filtered $H_2O$ and subjected to four blocking cycles each consisting of the following: **a)** DNA was heated at 94 °C for 2 min to denature and immediately cooled on ice for 2 min. **b)** rSAP (1 µL) and rCutSmart buffer (3.5 µL) were added, the reaction was incubated for 30 min at 37 °C, and the phosphatase was heat deactivated for 10 min at 65 °C. **c)** Four ddNTPs (1 µL, 2 mM each), TdT reaction buffer (1.5 µL), $CoCl_2$ (5 µl, 0.25 mM), terminal transferase (1 µL, 20 units), and filtered $H_2O$ (3.5 µL) were added and the reaction incubated for 60 min at 37 ˚C. Fresh reagents were added to each subsequent cycle at specific steps: rSAP (1 µL) and rCutSmart buffer (0.3 µL) were added at **Step b**, while ddNTPs (0.3 µL each), Tdt reaction buffer (0.3 µL), $CoCl_2$ (0.3 µL), and terminal transferase (1 µL) were

added at **Step c**. After all cycles were complete, excess unincorporated ddNTPs were removed by treatment with a DNA Clean & Concentrator kit (Zymo #11-304C). For iodine cleavage of PTs, the blocked DNA (32 µL) was incubated with 500 mM Tris-HCl pH 9.0 (4 µL) and iodine solution (4 µL, 5 mM) (Fluka #318981-100) for 10 min at 65 °C and cooled to 4 °C at a speed of 0.1 °C/s. The reaction was then split into two 20 µL aliquots and each filtered using a single DyeEX column (QIAGEN #63206) to remove salts and iodine, followed by recombining eluents. Filtered DNA was again denatured by heating at 94 °C for 2 min and then cooling on ice for 2 min. Thereafter, the reaction was adjusted with rCutSmart buffer (5 µL), rSAP (1 µL), and filtered $H_2O$ to a final volume of 50 µL, and incubated for 30 min at 37 ˚C to remove 3'-terminal phosphates arising from iodine cleavage. The phosphatase enzyme was then inactivated for 10 min at 65 ˚C. The reaction was further adjusted with dTTP (1 µL, 1 mM) (NEB #N0443S), TdT reaction buffer (1 µL), $CoCl_2$ (6 µl, 0.25 mM), filtered $H_2O$ (1 µL), and terminal transferase (1 µL, 20 units) and incubated for 45 min at 37 °C to add dT-tails to the 3'-terminal ends created specifically by iodine at PT sites. The reaction was split into three 20 µL aliquots, excess dTTP was removed using three DyeEX columns, and then eluents were recombined. The reaction (60 µL) was then adjusted with TdT reaction buffer (7.8 µL), $CoCl_2$ (7.8 µL), ddUTP-biotin (1 µL, 1 mM) (Jena Bioscience #NU-1619-BIOX-S), filtered $H_2O$ (0.4 µL), and terminal transferase (1 µL) and incubated for 60 min at 37 °C to terminate the dT-tails with a conjugated biotin moiety. After cleaning with four DyeEX columns, and recombining eluents, the processed DNA was diluted in filtered $H_2O$ (500 µL) and fragmented by probe sonication as described above. Thereafter, DNA fragments were mixed with Hydrophilic Streptavidin Magnetic Beads (5 µL) (NEB #S1421S) and binding buffer (500 µL) (5 mM Tris-HCl, pH 7.5, 1 M NaCl, 0.5 mM EDTA) and incubated for 60 min on a Nutator at room temperature. The beads were subjected to a pull-down process using a magnetic separation rack (Dynal #MPC-S) and washed three times with binding buffer (100 µL). After discarding the supernatant, beads were resuspended in filtered $H_2O$ (20 µL. Finally, the enriched DNA that was captured on the beads was directly subjected to Illumina library preparation using a SMART ChIP-seq kit (Takara #634865) following the manufacturer's protocol. The final PCR step was performed using Illumina primers provided in the kit and 12

cycles were run for amplification. The PCR product, with unique sequencing barcodes, was submitted to the MIT BioMicro Center for next-generation sequencing. Paired-end sequencing (150-bp) was performed on an Illumina MiSeq instrument, using a Custom Read2 sequencing primer that was supplied in the kit (poly(dA) followed by universal read 2 primer).

**Data analysis.** As illustrated in the workflow (**Supplementary Fig. S4B**), the data analysis started with trimming adapters. Adapters were removed using bbduk from BBtools v35.85 (sourceforge.net/projects/bbmap/), (with parameters ktrim=r k=18 hdist=2 hdist2=1 rcomp=f mink=8 qtrim=r trimq=30 for R1, ktrim=r k=18 mink=8 hdist=1 rcomp=f qtrim=r trimq=30 for R2). T-tails were removed using bbduk with parameters ktrim=r k=15 hdist=1 rcomp=f mink=8 for R1 and ktrim=l k=15 hdist=1 rcomp=f mink=8 for R2. PhiX were removed using bbduk with parameters k=31 hdist=1. Trimmed reads were aligned to the corresponding genome using Bowtie2 v2.4.5 with setting “sensitive”. The bam files were cleaned using SMARTcleaner v1.0[64] and split into two strands using samtools v1.19.2[65]. The coverage of each strand was calculated separately using bedtools v2.30.0[66] with the genomecov -d -5 option. Then, custom scripts were used for pileup calling. Briefly, all read start positions were recorded for both strands separately. The read pileup depth at each position was represented by the number of read starting (5'-end) at each position. The 13 nt sequences centered at positions with depth ≥1 were retrieved using samtools[65]. The consensus motifs were analyzed with incrementing depth, typically from 1 to 100 with step of 10, using MEME v5.3.3[67] with parameters -dna -objfun classic -nmotifs 5 -mod zoops -evt 0.05 -minw 3 -maxw 6 -markov_order 0 -nostatus -oc. The read starting sites mapping within less than 3 bps collapsed into the centermost consensus motif sites. Regions where starting positions mapped within 4 bp on opposite strands and within reverse complementary sequences were considered as double stranded modifications.

**Data availability**

Sequencing data have been deposited in NCBI SRA database under BioProject ID PRJNA1006039 (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA1006039). Raw

LC-MS data have been deposited to the PRIDE archive with accession PXD059148 (https://www.ebi.ac.uk/pride/archive/projects/PXD059148).

**Code availability**

Custom scripts for processing the sequencing data are described in Methods and are available at https://github.com/dedonlab/PTseq_data_analysis.git[68].

**References**


1. Sanchez-Romero, M.A. & Casadesus, J. The bacterial epigenome. *Nat Rev Microbiol* **18**, 7-20 (2020).
2. Thiaville, J.J. et al. Novel genomic island modifies DNA with 7-deazaguanine derivatives. *Proc Natl Acad Sci U S A* **113**, E1452-1459 (2016).
3. Zhou, X. et al. A novel DNA modification by sulphur. *Mol Microbiol* **57**, 1428-1438 (2005).
4. Wang, L. et al. Phosphorothioation of DNA in bacteria by dnd genes. *Nat Chem Biol* **3**, 709-710 (2007).
5. Wang, L., Jiang, S., Deng, Z., Dedon, P.C. & Chen, S. DNA phosphorothioate modification—a new multi-functional epigenetic system in bacteria. *FEMS Microbiol Rev* **43**, 109-122 (2019).
6. Xu, T., Yao, F., Zhou, X., Deng, Z. & You, D. A novel host-specific restriction system associated with DNA backbone S-modification in Salmonella. *Nucleic Acids Res* **38**, 7133-7141 (2010).
7. Xiong, X. et al. SspABCD-SspE is a phosphorothioation-sensing bacterial defence system with broad anti-phage activities. *Nat Microbiol* **5**, 917-928 (2020).

8. Wang, S. et al. SspABCD-SspFGH Constitutes a New Type of DNA Phosphorothioate-Based Bacterial Defense System. *mBio* **12** (2021).
9. Jian, H. et al. The origin and impeded dissemination of the DNA phosphorothioation system in prokaryotes. *Nat Commun* **12**, 6382 (2021).
10. Xiong, L. et al. A new type of DNA phosphorothioation-based antiviral system in archaea. *Nat Commun* **10**, 1688 (2019).
11. An, T. et al. A DNA phosphorothioation pathway via adenylated intermediate modulates Tdp machinery. *Nat Chem Biol* **21**, 1160-1170 (2025).
12. Tong, T. et al. Occurrence, evolution, and functions of DNA phosphorothioate epigenetics in bacteria. *Proc Natl Acad Sci U S A* **115**, E2988-E2996 (2018).
13. Cao, B. et al. Genomic mapping of phosphorothioates reveals partial modification of short consensus sequences. *Nat Commun* **5**, 3951 (2014).
14. Chen, C. et al. Convergence of DNA methylation and phosphorothioation epigenetics in bacterial genomes. *Proc Natl Acad Sci U S A* **114**, 4501-4506 (2017).
15. You, D.L., Wang, L.R., Yao, F., Zhou, X.F. & Deng, Z.X. A novel DNA modification by sulfur: DndA is a NifS-like cysteine desulfurase capable of assembling DndC as an iron-sulfur cluster protein in Streptomyces lividans. *Biochemistry-Us* **46**, 6126-6133 (2007).
16. Yao, F., Xu, T., Zhou, X., Deng, Z. & You, D. Functional analysis of spfD gene involved in DNA phosphorothioation in Pseudomonas fluorescens Pf0-1. *FEBS Lett* **583**, 729-733 (2009).

17. Dai, D. et al. DNA phosphorothioate modification plays a role in peroxides resistance in *Streptomyces lividans*. *Frontiers in Microbiology* **7**, 1-13 (2016).

18. Huang, Q. et al. Defense Mechanism of Phosphorothioated DNA under Peroxynitrite-Mediated Oxidative Stress. *ACS Chem Biol* **15**, 2558-2567 (2020).

19. Kellner, S. et al. Oxidation of phosphorothioate DNA modifications leads to lethal genomic instability. *Nat Chem Biol* **13**, 888-894 (2017).

20. Bhattacharyya, A., Chattopadhyay, R., Mitra, S. & Crowe, S.E. Oxidative stress: an essential factor in the pathogenesis of gastrointestinal mucosal diseases. *Physiol Rev* **94**, 329-354 (2014).

21. Zhu, S. et al. Development of Methods Derived from Iodine-Induced Specific Cleavage for Identification and Quantitation of DNA Phosphorothioate Modifications. *Biomolecules* **10** (2020).

22. Mangerich, A. et al. Infection-induced colitis in mice causes dynamic and tissue-specific changes in stress response and DNA damage leading to colon cancer. *Proc Natl Acad Sci U S A* **109**, E1820-1829 (2012).

23. Sun, Y. et al. DNA Phosphorothioate Modifications Are Widely Distributed in the Human Microbiome. *Biomolecules* **10** (2020).

24. Byrne, S.R. et al. Temporal dynamics and metagenomics of phosphorothioate epigenomes in the human gut microbiome. *Microbiome* **13**, 81 (2025).

25. Poyet, M. et al. A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. *Nat Med* **25**, 1442-1452 (2019).

26. Groussin, M. et al. Elevated rates of horizontal gene transfer in the industrialized human microbiome. *Cell* **184**, 2053-2067 e2018 (2021).

27. Poyet, M. et al. Industrialization drives convergent microbial and physiological shifts in the human metaorganism. *bioRxiv*, 2025.2010.2020.683358 (2025).

28. Rühlemann, M. et al. Convergent genomic responses of human gut bacteria to variations in industrialization. *bioRxiv*, 2025.2010.2020.683395 (2025).

29. Almeida, A. et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat Biotechnol* **39**, 105-114 (2021).

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

31. Oberg, N., Zallot, R. & Gerlt, J.A. EFI-EST, EFI-GNT, and EFI-CGFP: Enzyme Function Initiative (EFI) Web Resource for Genomic Enzymology Tools. *J Mol Biol* **435**, 168018 (2023).

32. Kambampati, R. & Lauhon, C.T. Evidence for the transfer of sulfane sulfur from IscS to ThiI during the in vitro biosynthesis of 4-thiouridine in Escherichia coli tRNA. *J Biol Chem* **275**, 10727-10730 (2000).

33. Shigi, N. Biosynthesis and functions of sulfur modifications in tRNA. *Front Genet* **5**, 67 (2014).

34. Bouvier, D. et al. TtcA a new tRNA-thioltransferase with an Fe-S cluster. *Nucleic Acids Res* **42**, 7960-7970 (2014).

35. Goldfarb, T. et al. BREX is a novel phage resistance system widespread in microbial genomes. *EMBO J* **34**, 169-183 (2015).

36. Yao, J. et al. Identification and characterization of a HEPN-MNT family type II toxin-antitoxin in Shewanella oneidensis. *Microb Biotechnol* **8**, 961-973 (2015).

37. Songailiene, I. et al. HEPN-MNT Toxin-Antitoxin System: The HEPN Ribonuclease Is Neutralized by OligoAMPylation. *Mol Cell* **80**, 955-970 e957 (2020).

38. Yao, J. et al. Novel polyadenylylation-dependent neutralization mechanism of the HEPN/MNT toxin/antitoxin system. *Nucleic Acids Res* **48**, 11054-11067 (2020).

39. Olson, R.D. et al. Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR. *Nucleic Acids Res* **51**, D678-D689 (2023).

40. Schirrmeister, B.E., Gugger, M. & Donoghue, P.C. Cyanobacteria and the Great Oxidation Event: evidence from genes and fossils. *Palaeontology* **58**, 769-785 (2015).

41. Luo, G. et al. Rapid oxygenation of Earth's atmosphere 2.33 billion years ago. *Sci Adv* **2**, e1600134 (2016).

42. Zimmermann, L. et al. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *J Mol Biol* **430**, 2237-2243 (2018).

43. Wang, L. et al. DNA phosphorothioation is widespread and quantized in bacterial genomes. *Proc Natl Acad Sci U S A* **108**, 2963-2968 (2011).

44. Liu, G. et al. Cleavage of phosphorothioated DNA and methylated DNA by the type IV restriction endonuclease ScoMcrA. *PLoS Genet* **6**, e1001253 (2010).

45. Wu, X. et al. Epigenetic competition reveals density-dependent regulation and target site plasticity of phosphorothioate epigenetics in bacteria. *Proc Natl Acad Sci U S A* **117**, 14322-14330 (2020).

46. Schaudy, E., Lietard, J. & Somoza, M.M. Sequence Preference and Initiator Promiscuity for De Novo DNA Synthesis by Terminal Deoxynucleotidyl Transferase. *ACS Synth Biol* **10**, 1750-1760 (2021).

47. Zhang, A. et al. Solid-phase enzyme catalysis of DNA end repair and 3' A-tailing reduces GC-bias in next-generation sequencing of human genomic DNA. *Sci Rep* **8**, 15887 (2018).

48. Forget, S.M. et al. Evolving a terminal deoxynucleotidyl transferase for commercial enzymatic DNA synthesis. *Nucleic Acids Res* **53** (2025).

49. Cao, B. et al. Nick-seq for single-nucleotide resolution genomic maps of DNA modifications and damage. *Nucleic Acids Res* (2020).

50. Jiang, S. et al. A DNA phosphorothioation-based Dnd defense system provides resistance against various phages and is compatible with the Ssp defense system. *mBio* **14**, e0093323 (2023).

51. Zhu, Q. et al. Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat Commun* **10**, 5477 (2019).

52. Garcia-Bayona, L. et al. Nanaerobic growth enables direct visualization of dynamic cellular processes in human gut symbionts. *Proc Natl Acad Sci U S A* **117**, 24484-24493 (2020).

53. Pantosti, A., Tzianabos, A.O., Onderdonk, A.B. & Kasper, D.L. Immunochemical characterization of two surface polysaccharides of Bacteroides fragilis. *Infect Immun* **59**, 2075-2082 (1991).

54. Smith, C.J. & Callihan, D.R. Analysis of rRNA restriction fragment length polymorphisms from Bacteroides spp. and Bacteroides fragilis isolates associated with diarrhea in humans and animals. *J Clin Microbiol* **30**, 806-812 (1992).

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

56. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410 (1990).

57. Buchfink, B., Reuter, K. & Drost, H.G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* **18**, 366-368 (2021).

58. Jauffrit, F. et al. RiboDB Database: A Comprehensive Resource for Prokaryotic Systematics. *Mol Biol Evol* **33**, 2170-2172 (2016).

59. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-1797 (2004).

60. Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol Biol* **10**, 210 (2010).

61. Gouy, M., Guindon, S. & Gascuel, O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol* **27**, 221-224 (2010).

62. Price, M.N., Dehal, P.S. & Arkin, A.P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* **26**, 1641-1650 (2009).

63. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res* **49**, W293-W296 (2021).

64. Zhao, D. & Zheng, D. SMARTcleaner: identify and clean off-target signals in SMART ChIP-seq analysis. *BMC Bioinformatics* **19**, 544 (2018).

65. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

66. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).

67. Bailey, T.L., Johnson, J., Grant, C.E. & Noble, W.S. The MEME Suite. *Nucleic Acids Res* **43**, W39-W49 (2015).

68. Yuan, Y., DeMott, M.S., & Dedon, P.C. Phosphorothioate DNA modification by BREX type 4 systems in the human gut microbiome, PTseq_data_analysis, https://doi.org/10.5281/zenodo.17890169 (2025).

**Acknowledgements**

We thank Susan Weir and Katya Moniz at the Openbiome for providing us with bacterial isolates. This work was supported by National Institutes of Health (R01 ES031576, P.C.D., E.J.A.; R01AI093771, L.C.), by a NIEHS Training Grant in Environmental Toxicology T32-ES007020 (S.R.B), and the Duchossois Family Institute (L.C.), and by funding for the GMbC from the MIT Center for Microbiome Therapeutics and the Neil and Anna Rasmussen Family Foundation.

**Author contributions**

Conceptualization, P.C.D. and E.J.A.; methodology, P.C.D., E.J.A., Y.Y., and L.E.C.; data acquisition, S.B., Y.Y., L.E.C., M.P., M.G., K.F., Y.Y., J.R.B., C.G., J.L., A.Z.P.M., I.E.M., Y.A.N., L.T.T.N., C.A.O., L.R.R., J.S., T.V.; data analysis P.C.D., E.J.A., Y.Y., and M.S.D.; writing original draft, Y.Y., M.S.D., P.C.D., and S.R.B. GMbC represents authors affiliated with the Global Microbiome Conservancy.

**Competing Interests Statement**

The authors declare no competing interests.

**Figure Legends**

**Figure 1. Microbial phosphorothioate (PT) DNA modifications**. (**A**) Sulfur replaces a non-bridging phosphate oxygen in the DNA backbone in PT modifications. A limit nuclease digest of PT-containing DNA leaves PT-linked dinucleotides that can be identified and quantified by LC-MS. (**B**) Synthesis of PTs generally follows the biochemical steps performed by Dnd proteins.

**Figure 2. Gene neighborhoods analyses based on sequence similarity networks (SSNs) of DndC and SspD proteins essential for PT synthesis.** Sequence similarity networks (SSN) analysis was performed for the 3,120 closest homologues of DndC (A) and 2,132 homologues of SspD (B) in the UniProt database. Each node (circle) in the network represents one DndC or SspD protein. An edge is drawn between two nodes that have a BLAST E-value cutoff of $\leq 10^{-100}$ (alignment score of 100) in the DndC SSN or $10^{-50}$ (alignment score of 50) in the SspD SSN. Some nodes are filled with colors according to their genome neighborhood structure, with a representative species shown and indicated in the SSNs with a two-letter label. The node outlines are colored red when *dndD* is present in the genome neighborhood of *dndC* and similarly for *sspD* when *sspBC* are present in the neighborhood. For better visualization, single nodes and clusters with only a few nodes were hidden. Abbreviations: HEPN, higher eukaryotes

and prokaryotes nucleotide-binding; MTase, methyltransferase; NTPase, nucleoside triphosphatase; NTP transf, Nucleotidyltransferase; top, topoisomerase.

**Figure 3. Homology analysis of BREX type 4 systems and evidence of PT synthesis.** (**A**) The genomic organization of the BREX type 4 gene systems in *Bacteroides salyersiae* and putative *Butyricimonas faecalis*. *Vibrio cyclitrophicus* FF75 is included to demonstrate a typical *ssp* gene system and *Escherichia coli* HS to demonstrate a typical BREX type 1 system. Protein domains are color-coded and labeled. (**B**) The levels of PT dinucleotides in engineered *B. salyersiae* and *Bacteroides thetaiotaomicron* strains. *ΔbrxC, ΔbrxC* $brxC_{Bf}$, *ΔbrxC* $sspC_{Bo}$, vector, and $brxC_{Bs}$ all differ significantly from wild-type (WT) based on a two-tailed Student's *t*-test with three biological replicates ($p < 0.05$). Data are presented as mean values +/- SD. Bs: *B. salyersiae*, Bf: *Butyricimonas faecalis*, Bo: *Bacteroides ovatus*. Statistic data are provided in **Supplementary Data 15 and 16**.

**Figure 4. Phylogenetic distribution of *dndCD, sspBCD, and brxPC* genes in human gut microbiome genomes.** The reference phylogeny was reconstructed from the concatenated alignment of 10 ribosomal proteins. The colors of the triangles in the tree show the taxa at the order level. The occurrence of gene clusters was quantitatively presented by the colors and size of circles. For better visualization, orders containing less than 5 genomes were hidden. Source data are provided in **Supplementary Table S5**.

**Figure 5**. **PT dinucleotides and PT consensus sequences in human gut microbiome isolates.** The numbers of isolates were analyzed are listed on left. The circles represent PT dinucleotides quantified by LC-MS (left) and the presence of corresponding genes (right). The PT-modified consensus motifs were characterized in one representative isolate from each group using PT-seq. t.b.d., to be determined. *, the $C_{PS}AG$ motif was characterized using metagenomics PT-seq in another on-going study. **, $G_{PS}A$ and $G_{PS}T$ were detected using LC-MS QQQ but the exact mass was not verified by high-resolution mass spectrometry.

**Figure 6**. **Biogeographical maps of PTs in bacteria with BREX type 4 systems.** (**A**) Pie charts depict the number of single- or bi-stranded PT-modified consensus sequences. The structural diagrams depict these motifs. (**B**) Analysis of the distribution of PT consensus sequences and modified sites in 1kb upstream and downstream regions in *B. salyersiae* (**left**) and putative *B. faecalis* (**right**). The number of total motif sites in the sense strand (**upper**) or both strands (**lower**) are represented in pink. The PT modified sites are represented in purple. The fraction of PT-modified motif sites is represented by the black line.

Here, the authors combine bioinformatic, mass spectrometric, and sequencing-based mapping tools to identify previously undescribed phosphorothioate epigenetic systems widespread in the human gut microbiome.

**Peer Review Information:** *Nature Communications* thanks Guang Liu, Richard Morgan, and Peter Weigele for their contribution to the peer review of this work. A peer review file is available.