

# Comparative genomics provides structural and functional insights into *Bacteroides* RNA biology

Gianluca Prezza <sup>1</sup> | Daniel Ryan<sup>1</sup> | Gohar Mädler<sup>2</sup> | Sarah Reichardt<sup>1</sup> | Lars Barquist<sup>1,3</sup> | Alexander J. Westermann <sup>1,2</sup>

<sup>1</sup>Helmholtz Institute for RNA-based Infection Research (HIRI), Helmholtz Centre for Infection Research (HZI), Würzburg, Germany

<sup>2</sup>Institute of Molecular Infection Biology (IMIB), University of Würzburg, Würzburg, Germany

<sup>3</sup>Faculty of Medicine, University of Würzburg, Würzburg, Germany

## Correspondence

Lars Barquist and Alexander J. Westermann, Helmholtz Institute for RNA-based Infection Research (HIRI), Helmholtz Centre for Infection Research (HZI), Würzburg, Germany.

Email: lars.barquist@helmholtz-hiri.de and alexander.westermann@uni-wuerzburg.de

## Funding information

Deutsche Forschungsgemeinschaft, Grant/Award Number: We6689/1-1

## Abstract

Bacteria employ noncoding RNA molecules for a wide range of biological processes, including scaffolding large molecular complexes, catalyzing chemical reactions, defending against phages, and controlling gene expression. Secondary structures, binding partners, and molecular mechanisms have been determined for numerous small noncoding RNAs (sRNAs) in model aerobic bacteria. However, technical hurdles have largely prevented analogous analyses in the anaerobic gut microbiota. While experimental techniques are being developed to investigate the sRNAs of gut commensals, computational tools and comparative genomics can provide immediate functional insight. Here, using *Bacteroides thetaiotaomicron* as a representative microbiota member, we illustrate how comparative genomics improves our understanding of RNA biology in an understudied gut bacterium. We investigate putative RNA-binding proteins and predict a *Bacteroides* cold-shock protein homolog to have an RNA-related function. We apply an in silico protocol incorporating both sequence and structural analysis to determine the consensus structures and conservation of nine *Bacteroides* noncoding RNA families. Using structure probing, we validate and refine these predictions and deposit them in the Rfam database. Through synteny analyses, we illustrate how genomic coconservation can serve as a predictor of sRNA function. Altogether, this work showcases the power of RNA informatics for investigating the RNA biology of anaerobic microbiota members.

## KEYWORDS

BT\_1884, cold-shock protein, GibS, RNA-binding proteins, secondary structure, 6S RNA

## 1 | INTRODUCTION

Small regulatory RNAs (sRNAs) are key mediators of bacterial gene expression, allowing microbes to rapidly adapt to changing environmental conditions and cope with diverse stresses (Holmqvist & Wagner, 2017). Their functionality is defined by primary sequence

and secondary structure. The former determines their ability to base-pair with complementary regions within target RNAs, whereas the latter influences a broad range of cellular properties, from half-life to recognition by RNA-binding proteins (RBPs) (Wagner & Romby, 2015). Comparative genomics has been a foundational approach for investigating both. The first genome-wide screens for

Gianluca Prezza and Daniel Ryan contributed equally to this study.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2021 The Authors. *Molecular Microbiology* published by John Wiley & Sons Ltd.

sRNAs were enabled by the availability of multiple enterobacterial genome sequences that allowed for the evaluation of sequence conservation (Argaman et al., 2001; Rivas et al., 2001; Wassarman et al., 2001). Conservation has also been used as evidence of function for particular regions of sRNAs (Papenfert et al., 2010) and appears to be a general marker of target interaction sites (Peer & Margalit, 2011; Richter & Backofen, 2012). The use of comparative genomics to determine secondary structure has an even longer history; for instance, manual comparisons of related sequences allowed the determination of highly accurate secondary structures for basic biomolecules such as tRNAs (Madison et al., 1966) and rRNAs (Woese et al., 1983) decades before experimentally determined structures were available. This basic approach, mathematically formalized and computationally automated, has driven a long series of noncoding RNA (ncRNA) discoveries in bacterial genomes. These include identification of 6S as an ancient housekeeping RNA in bacteria (Barrick et al., 2005), the discovery of an astonishingly diverse menagerie of metabolite-responsive riboswitches (McCown et al., 2017), studies of autoregulatory sequences in the untranslated regions of mRNAs encoding ribosomal proteins (Fu et al., 2013), and many more. As these approaches only require access to genomic sequences, they are particularly well suited to investigating the structure and function of noncoding RNAs in organisms that are difficult to cultivate or manipulate.

*Bacteroides* spp. are dominant colonizers of the mammalian large intestine (Wexler, 2007) and have emerged as model organisms for the microbiota (Wexler & Goodman, 2017). These obligate anaerobic, Gram-negative bacteria are metabolic specialists with the capacity to break down a variety of dietary fiber- and host mucus-derived polysaccharides, thereby facilitating nutrient absorption by the gut epithelium (Glowacki & Martens, 2020). Additionally, intestinal *Bacteroides* spp. protect their host from enteric infections by priming the development of the immune system and providing colonization resistance against pathogens (Buffie & Pamer, 2013; Hooper et al., 2000). In contrast to the situation in the model organisms of the Proteobacteria, Bacteroidetes RNA research is only in its infancy (Ryan et al., 2020b). For example, in contrast to other Gram-negative phyla where sRNA function is often tied to assisting RNA chaperones such as Hfq (Kavita et al., 2018) or FinO domain-containing proteins such as ProQ (Olejniczak & Storz, 2017), no global RBP is known in Bacteroidetes, though RRM-domain-containing proteins have recently been suggested as candidates (Adams et al., 2021).

RNA sequencing (RNA-seq) was applied to *Bacteroides fragilis* (Cao et al., 2016) and *Bacteroides thetaiotaomicron* (Ryan et al., 2020a)—the two workhorses of *Bacteroides* research—and revealed hundreds of noncoding RNA candidates, yet their conservation, secondary structures, and functions have not been determined systematically. For example, our differential RNA-seq (Sharma & Vogel, 2014) screen identified 151 sRNAs in *B. thetaiotaomicron* type strain VPI-5482 including 124 intergenically encoded candidates (Ryan et al., 2020a). As of now, however, only two *trans*-encoded *B. thetaiotaomicron* sRNAs have been functionally characterized: GibS, which posttranscriptionally binds and controls the expression of

metabolic target mRNAs (Ryan et al., 2020a), and RteR, which likely acts as a cotranscriptional repressor of a transposon operon (Waters & Salyers, 2012). For GibS, we recently determined the secondary structure—composed of a single-stranded 5' portion harboring the seed sequence, two meta-stable hairpins in the central region, followed by the intrinsic terminator hairpin—by a combination of computational prediction and chemical/enzymatic validation (Ryan et al., 2020a). In the case of RteR, a structure consisting of a long 5' hairpin, an eight nt-long single-stranded stretch, followed by the 3' terminator was proposed based on minimum free energy (MFE) calculations (Waters & Salyers, 2012). It is, however, currently unclear to what degree the structures and working mechanisms of GibS and RteR might be representative for the bulk of sRNAs in this genus.

Here, we apply a suite of *in silico* analyses to improve our understanding of *Bacteroides* sRNA biology. We start with a computational search for *Bacteroides* proteins with known RNA-binding domains, identifying a conserved Bacteroidetes cold-shock protein (CSP) as a putative RBP in this phylum, and examine global properties of *Bacteroides* sRNAs in comparison with Proteobacteria and Firmicute sRNAs. We then predict structures of a number of previously identified intergenic *B. thetaiotaomicron* sRNAs and core ncRNAs, harnessing information about covarying base-pairing ribonucleotides. Using structure probing, we validate and refine the predicted consensus structures for selected ncRNAs, namely the ubiquitous 6S and 4.5S RNAs as well as two intergenic sRNAs (an RteR homolog and BTnc201). The curated alignments and consensus structures of *Bacteroides* RNA families have been deposited in the Rfam database (Kalvari et al., 2018). Finally, based on the assumption that adjacently encoded genes and operons might be functionally connected, we perform synteny analyses to predict biological pathways that individual sRNAs might be involved in. Altogether, this study illustrates the power of *in silico* approaches to infer structural and functional features of sRNAs in human-relevant bacterial species for which the experimental toolkit is still in the making.

## 2 | RESULTS

### 2.1 | RBP candidates in *Bacteroides*

Bacterial species of the Proteobacterium phylum—including the genera *Escherichia* and *Salmonella*—have long served us as model Gram-negative organisms including for RNA research (Hör et al., 2020). In stark contrast, RNA biology is poorly understood in the distantly related genus *Bacteroides* (Figure 1a) (Ryan et al., 2020b). Given the profound impact global RBPs exert on proteobacterial sRNA function, we first performed an *in silico* search for *Bacteroides* proteins harboring known RNA-binding domains using Pfam release 32 (El-Gebali et al., 2019). Neither an Hfq nor a ProQ homolog could be identified (Figure 1b). Likewise, CsrA/RsmA—an otherwise highly conserved translational regulatory RBP (Romeo & Babitzke, 2018)—is absent from the Bacteroidetes. Instead, the analysis yielded *B. thetaiotaomicron* proteins with domains found occasionally in other bacterial RBPs,

namely putative K homology (KH), RNA recognition motif (RRM), or cold-shock domains (CSDs). Specifically, BT\_2563, BT\_4417, BT\_3835, BT\_2721, and BT\_3403 contain KH domains (Pfam IDs: PF00013, PF07650, PF13184), BT\_0784, BT\_1887, and BT\_3840 contain each an RRM-1 domain (PF00076), whereas BT\_1884 harbors a putative CSD (PF00313) (Figure 1c). Based on their homology to known transcriptional regulators (BT\_3403 is a homolog of the transcription termination/antitermination factor NusA), ribosomal proteins and biogenesis factors (BT\_3835 is a putative GTP-binding protein with a role in ribosome biogenesis; BT\_2721 is the 30S ribosomal protein S3), or ribonucleases (BT\_2563 is a putative polyribonucleotide nucleotidyltransferase; BT\_4417 is a putative ribonuclease Y), the *B. thetaiotaomicron* KH-containing proteins appear unlikely to have a global RNA chaperone function. In contrast, the three RRM-1 proteins and the CSP could indeed act as regulatory RBPs in *Bacteroides*.

## 2.2 | Cold-shock protein BT\_1884 has an RNA-related function

Closer inspection of these candidates sparked our interest in the CSD-containing protein BT\_1884, as the eggNOG-mapper tool for functional annotation (Huerta-Cepas et al., 2017) suggested it to be a likely homolog of CspC/E (COG1278)—known global RBPs in Enterobacteriaceae with hundreds of mRNA and sRNA ligands (Michaux et al., 2017). We generated a *B. thetaiotaomicron* deletion mutant devoid of this protein (strain  $\Delta BT_{1884}$ ), its complementation in *trans* under its native promoter (strain  $BT_{1884}^+$ ), and a strain overexpressing *BT\_1884* from a strong constitutive phage promoter (strain  $BT_{1884}^{++}$ ) (Supplementary Figure S1a,b and Supplementary Table S1). Note that strain  $BT_{1884}^+$  harbors *BT\_1884* as a standalone gene, rather than as part of an operon as in the wild type (Supplementary Figure S1a). As a likely consequence, we detected *BT\_1884* mRNA at elevated levels (~100-fold) in this *trans*-complemented strain compared with the wild type (Supplementary Figure S1b). Based on observations made in Proteobacteria (Chao & Vogel, 2010; Romeo & Babitzke, 2018), where global RBP deletion or overexpression may result in pleiotropic effects, including altered growth rates, we recorded growth curves of the *B. thetaiotaomicron* strains with varying *BT\_1884* levels. However, at least when cultured in rich media at 37°C, depletion or constitutive expression of *BT\_1884* had no obvious effect on bacterial growth (Supplementary Figure S1c).

Next, we isolated total RNA from the same *B. thetaiotaomicron* strains in the stationary phase. Assuming that sRNA-binding proteins may affect the abundance of their ligands in the bacterial cell, we profiled the steady-state levels of 11 randomly selected sRNAs by northern blot (Figure 1d and Supplementary Figure S1d). None of the selected sRNAs showed altered levels upon the deletion of *BT\_1884*, not surprising in light of the low expression of endogenous *BT\_1884* mRNA in the wild-type strain (Supplementary Figure S1b; Ryan et al., 2020a). In contrast, two sRNAs were slightly (approximately two-fold) yet significantly upregulated in the *BT\_1884* overexpression

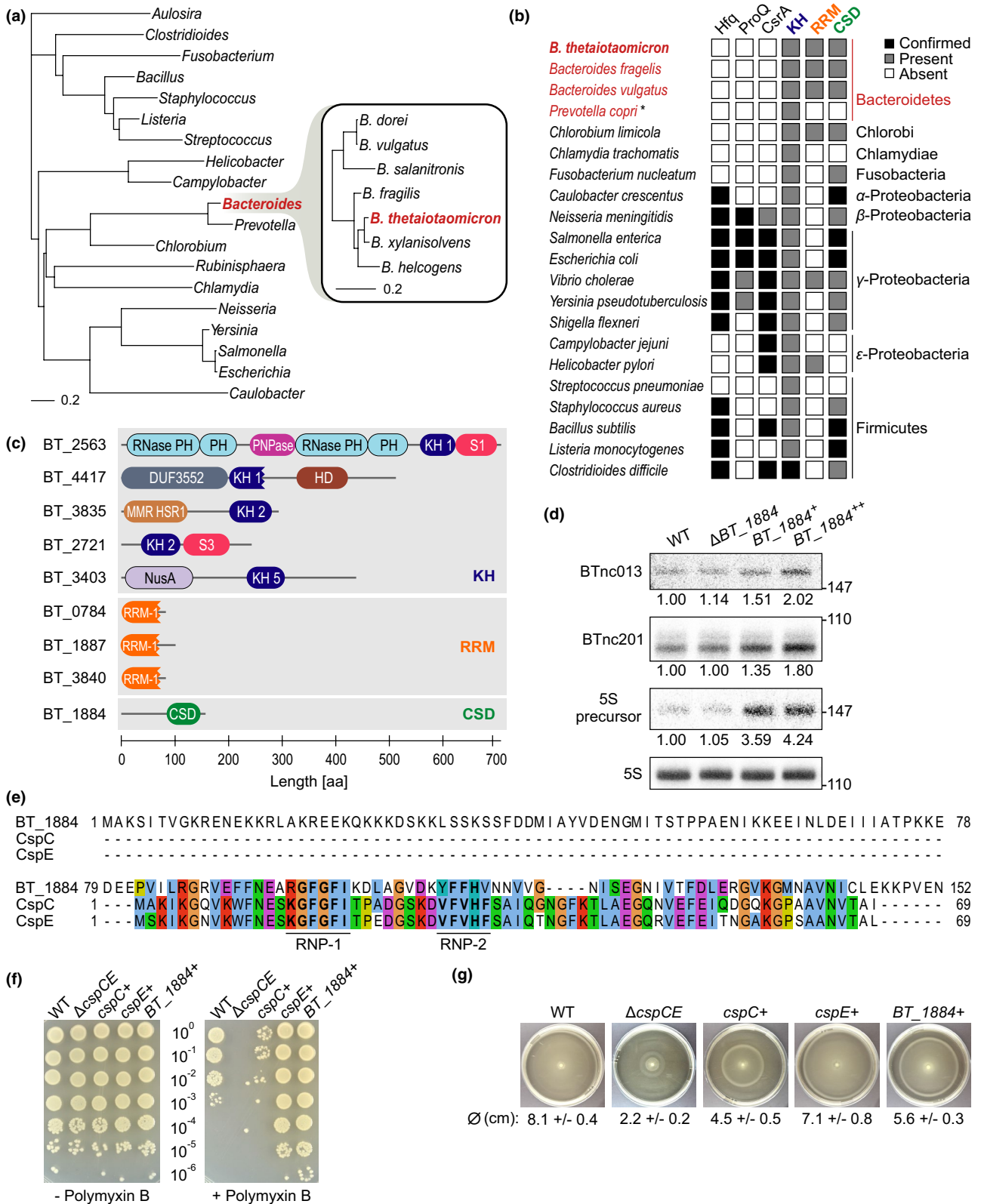
background compared with the wild-type strain, namely 5' UTR-derived BTnc013 and intergenic BTnc201 (Figure 1d). Overexpression of *BT\_1884* did not show an effect on the steady-state levels of the remaining nine sRNAs (Supplementary Figure S1d). Mature 5S ribosomal RNA (rRNA) was used as our loading control on the northern blots. Interestingly, whereas the levels of mature 5S (~110 nt) were not affected by the *BT\_1884* status, a putative precursor 5S (~150 nt) accumulated in a *BT\_1884*-dependent manner (Figure 1d; Supplementary Figure S1e). In *Escherichia coli*, RNA helicase RhIE is implicated in rRNA processing and ribosome maturation (Jain, 2008) and the *B. thetaiotaomicron* RhIE homolog (BT\_1885) is encoded upstream of *BT\_1884* within the same operon (Supplementary Figure S1a). However, we are confident that the observed effect on 5S rRNA processing is due to altered *BT\_1884* levels and independent of RhIE, because we excluded major polar effects in the *B. thetaiotaomicron* mutant strains (Supplementary Figure S1b).

Proteobacterial CSPs may facilitate RNA melting (Phadtare & Severinov, 2005), promote transcriptional antitermination (Bae et al., 2000), or stabilize their RNA ligands by protecting them from nucleolytic decay (Michaux et al., 2017). Here, through rifampicin run-out experiments (Chen et al., 2015), we excluded RNA stability as a major factor contributing to the observed changes in steady-state transcript levels across the different *BT\_1884* backgrounds (Supplementary Figure S1f).

*Salmonella* deletion mutants devoid of both CspC and CspE exhibit severe phenotypes, ranging from an increased sensitivity to peroxide, bile salts, and antimicrobial peptides to defects in biofilm formation, motility, and pathogenesis (Michaux et al., 2017). These phenotypes are likely a consequence of aberrant transcript levels for CspC/E ligands in the double mutant background and could be reverted when complementing with either CspC or CspE, establishing functional redundancy between these two *Salmonella* CSPs (Michaux et al., 2017). Compared with *Salmonella* CspC and -E, *Bacteroides* *BT\_1884* harbors an N-terminal extension yet shares the two RNA-binding motifs (RNP-1 and RNP-2; Figure 1e). We therefore asked whether *BT\_1884* might have a similarly global effect and would complement a *Salmonella*  $\Delta cspC/E$  double mutant. Reading out the survival rates when bacteria were challenged with the antimicrobial peptide polymyxin B, we found that heterologous *BT\_1884* expression overcomplemented polymyxin B sensitivity of *Salmonella*  $\Delta cspC/E$  (Figure 1f). Swimming assays further revealed that the *Bacteroides* CSP homolog can efficiently cross-complement the motility defect of *Salmonella*  $\Delta cspC/E$  (Figure 1g). Taken together, these observations suggest the CSP protein *BT\_1884* may indeed have a global RNA-related function and should encourage future studies of *Bacteroides* RBPs.

## 2.3 | Global characteristics of *B. thetaiotaomicron* noncoding RNAs

By genome-wide RNA-seq analysis, we previously identified 124 intergenic sRNAs in *B. thetaiotaomicron* (Ryan et al., 2020a). As



the *B. thetaiotaomicron* genome has some unique qualities with respect to bacterial models of RNA biology, being both AT-rich (42% GC content) and lacking known RBPs such as FinO-like proteins that tend to bind highly structured targets (Bauriedl et al., 2020; Gonzalez et al., 2017; Holmqvist et al., 2018; Pandey et al., 2020;

Stein et al., 2020), we asked if this was reflected in the properties of *Bacteroides* ncRNAs as a whole. First, we determined their average length (Figure 2a). Second, we calculated a measure of RNA "structuredness" as the MFE derived from the ViennaRNA RNAfold tool (Lorenz et al., 2011) corrected for the MFE distribution of random

**FIGURE 1** *Bacteroides* RNA-binding protein candidates. (a) Phylogenetic tree with zoom-in on *Bacteroides* species. The horizontal scale indicates the number of substitutions per site between the representative genomes of the selected genera/species. (b) Pfam search for RNA-binding domain (RBD)-containing proteins in Bacteroidetes and bacterial landmark species of other phyla. White square: not detected by our search; gray square: detected by our search; black square: detected and previously confirmed to be a (global) RNA binder. CSD, cold-shock domain; KH, K homology; RRM, RNA recognition motif. Asterisk: There is no completed genome available yet for *Prevotella copri*, so the hits are not guaranteed to be complete. (c) RBD proteins in *Bacteroides thetaiotaomicron* and their domain structure as determined using the Pfam sequence search function. (d), Steady-state levels of *B. thetaiotaomicron* sRNAs BTnc013 and BTnc201 and of a putative 5S rRNA precursor (see Supplementary Figure S1e) in different BT\_1884 backgrounds. Northern blots were performed with total RNA samples extracted from wild-type,  $\Delta$ BT\_1884, BT\_1884<sup>+</sup> (trans-complemented), or BT\_1884<sup>++</sup> (overexpression) *B. thetaiotaomicron* strains grown in TYG medium to stationary phase and were probed with RNA-specific, radiolabeled oligonucleotides. Mature 5S rRNA serves as loading control. Values to the right refer to the size of the corresponding marker bands (in nt). One representative out of three biological replicate experiments is depicted and quantification over all replicates is given at the bottom. See Supplementary Figure S1d for the full set of sRNAs selected for probing. (e) Amino acid alignment of *Bacteroides* BT\_1884 and *Salmonella* CspC and -E. Conserved residues are colored according to the Clustal X color scheme (blue: hydrophobic, red: positively charged, magenta: negatively charged, green: polar, pink: cysteine, orange: glycine, yellow: proline, cyan: aromatic). The two RNA-binding motifs (RNP-1 and RNP-2) are highlighted in bold font. Figure generated with Jalview (Waterhouse et al., 2009). (f) Polymyxin B sensitivity assay. Before and after polymyxin challenge (2  $\mu$ g/ml for 1 hr), serial dilutions of the indicated *Salmonella* Typhimurium strains were spotted on LB agar to assess survival. Shown is a representative result of three replicates. (g) Swimming assay of the indicated *Salmonella* strains. The diameters of swimming areas are given at the bottom (average of three replicates  $\pm$ SD)

sequences of the same dinucleotide composition as the background genome (Figure 2b). Both average length and structuredness were then compared with ncRNAs from model Proteobacteria and Firmicutes species (Dugar et al., 2013; Howden et al., 2013; Irnov et al., 2010; Koo et al., 2011; Kroger et al., 2013; Sharma et al., 2010; Shinhara et al., 2011; Slager et al., 2018; Wurtzel et al., 2012). Overall, neither of these parameters differed markedly between ncRNAs in *B. thetaiotaomicron* and the nine unrelated species, indicating that the absence of the classical RNA chaperones and low GC content of *Bacteroides* do not appear to be reflected in these general ncRNA properties.

We used iterative sequence homology searches over a database of genomes from class Bacteroidia with the nucleotide profile hidden Markov model (profile HMM)-based nhmmer (Wheeler & Eddy, 2013) to identify 22 intergenic *B. thetaiotaomicron* ncRNAs as conserved within and occasionally beyond *Bacteroides* spp. (Ryan et al., 2020a; Figure 2c). The most broadly conserved ncRNAs, as in most bacteria, are housekeeping transcripts (Jose et al., 2019). These include the transfer-messenger RNA (tmRNA) that rescues stalled ribosomes; the M1 RNA subunit of RNase P; the 4.5S RNA component of the signal recognition particle (SRP) that guides nascent integral membrane proteins to the translocation pore; and the 6S RNA, which sequesters RNA polymerase holoenzyme to globally modulate transcriptional activity. While our putative *B. thetaiotaomicron* tmRNA (RF00023) and M1 RNA (RF00010) sequences were recognized by covariance models in the Rfam database, our 4.5S sequence was not. Reverse BLAST searches using *Bacteroides* sequences from Rfam hit this sequence in *B. thetaiotaomicron* (see Methods), and nhmmer searches of the RNACentral database (Consortium, 2021) found hits to sequences annotated as the RNA component of SRP by the European Nucleotide Archive (Harrison et al., 2021). Other Bacteroidia sequences identified by our HMM were recognized by the Rfam bacterial small SRP covariance model (RF00169), further indicating the *B. thetaiotaomicron* sequence is likely a divergent member of this family. The *B. thetaiotaomicron* 6S sequence on

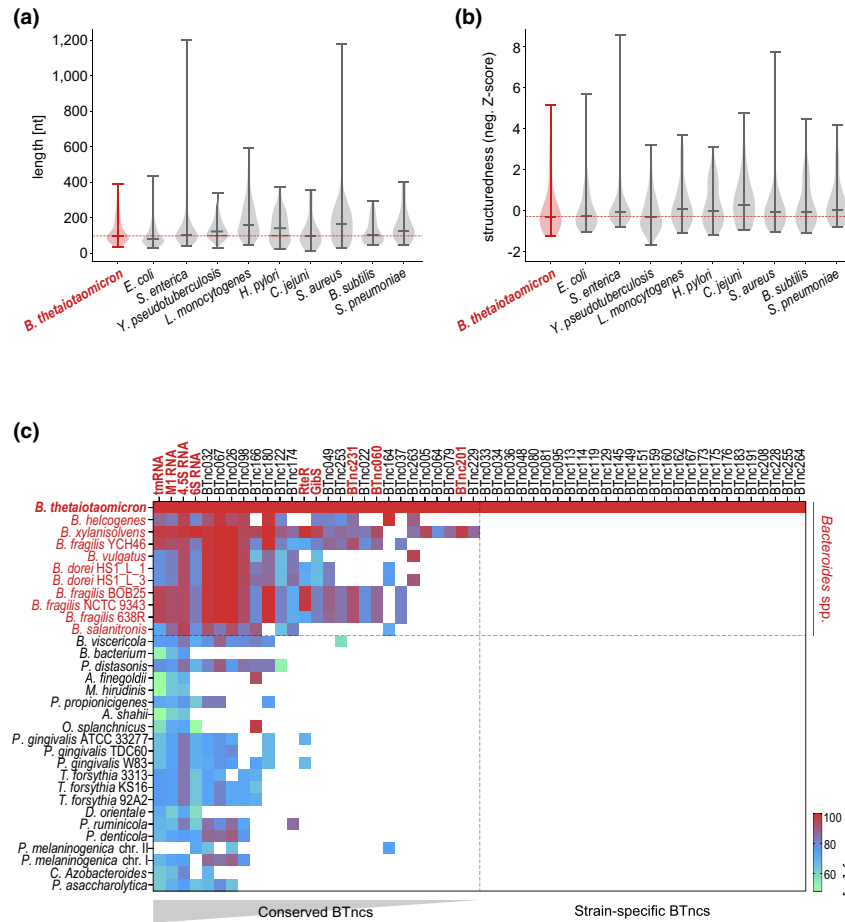
the other hand was recognized by Rfam, but as a member of the Bacteroidales-1 family (RF01693) of computationally discovered (Weinberg et al., 2010) experimentally uncharacterized sequence motifs. However, this sequence had formerly been informatically proposed as a likely 6S molecule (Wehner et al., 2014), and we previously showed that it produces the expected short product RNAs (pRNAs) (Ryan et al., 2020a).

Beyond core bacterial ncRNAs, the only two *Bacteroides* trans-encoded sRNAs characterized so far are RteR (Waters & Salyers, 2012) and GibS (Ryan et al., 2020a). RteR was not included in our original sRNA annotation (Ryan et al., 2020a). However, we here predicted a putative functional RteR homolog within the conjugative transposon CTn1 (see Methods for details), which we refer to as RteR throughout this study. Both GibS and RteR belong to a cluster of 12 partially conserved sRNAs present in most *Bacteroides* spp. but not beyond the genus level. While the uncharacterized sRNA candidates BTnc026, BTnc032, BTnc067, BTnc098, BTnc166, and BTnc180 are strongly conserved, the majority of sRNA candidates are narrowly conserved, with five candidates (e.g., BTnc201) detected in very few species other than *B. thetaiotaomicron* and the remaining 75 sRNAs being strain specific. This pattern of relatively narrow conservation is similar to that observed for sRNAs in other bacterial phyla (Lindgreen et al., 2014; Peer & Margalit, 2014).

## 2.4 | Integrated structure and sequence analysis of *B. thetaiotaomicron* core ncRNAs

In the following, we sought to determine the secondary structures of *B. thetaiotaomicron* ncRNAs. To this end, we focused on the well-conserved housekeeping RNAs and the partially conserved intergenic sRNAs, as this allowed for the incorporation of sequence conservation and single-nucleotide covariation information to strengthen the computational predictions. We based our approach on a published protocol to construct RNA families (Barquist





**FIGURE 2** Global ncRNA properties in *Bacteroides thetaiotaomicron*. Average length (a) and structuredness (negative Z score; defined as minimum free energy [MFE] divided by the MFE distribution of random sequences of the same dinucleotide composition as the background genome sequence) (b) of the intergenic ncRNA complement of *B. thetaiotaomicron* type strain VPI-5482 ( $n = 124$ ) compared with that of the nine indicated model species: *Escherichia coli* ( $n = 209$ ; Shinhara et al., 2011), *Salmonella enterica* serovar Typhimurium ( $n = 280$ ; Kroger et al., 2013), *Yersinia pseudotuberculosis* ( $n = 148$ ; Koo et al., 2011), *Listeria monocytogenes* ( $n = 113$ ; Wurtzel et al., 2012), *Helicobacter pylori* ( $n = 54$ ; Sharma et al., 2010), *Campylobacter jejuni* ( $n = 35$ ; Dugar et al., 2013), *Staphylococcus aureus* ( $n = 116$ ; Howden et al., 2013), *Bacillus subtilis* ( $n = 65$ ; Irnov et al., 2010), *Streptococcus pneumoniae* ( $n = 39$ ; Slager et al., 2018). The center line of each violin plot indicates the median and the upper and lower lines the maximum and minimum values, respectively. The dashed red line denotes the respective *B. thetaiotaomicron* median value. (c) *Bacteroides* ncRNA conservation. Bacteroidetes genome references are given to the left and intergenic sRNAs of *B. thetaiotaomicron* (their “BTnc” ID according to Ryan et al., 2020a or, where available, their trivial name) are indicated at the top, sorted by conservation from left to right (gray triangle). We refer to ncRNAs as “conserved” when they are present in more than one strain and as “strain-specific” when they are only found in *B. thetaiotaomicron*. The color code denotes nucleotide identity as indicated at the lower right. The heat map represents an adaptation of Supplementary Figure S3a in Ryan et al. (2020a), wherein we manually added 4.5S RNA (identified in this study; see Methods for details), tmRNA (Ryan et al., 2020a), and a presumed CTn1-encoded homolog (see Methods) of RteR (Waters & Salyers, 2012)

et al., 2016), loosely based on procedures initially developed for the Rfam database (Gardner et al., 2009). Using sequences gathered by our previous iterative HMMER searches, we created sequence alignments with consensus secondary structure predictions using the Webserver for Aligning structural RNAs (WAR) (Torarinsson & Lindgreen, 2008). WAR comprises 14 different alignment and structure prediction methods that variously consider MFE, residue conservation, and covariation and can provide a maximum consistency alignment and structure integrating individual predictions with T-Coffee (Notredame et al., 2000). The consensus structures were then manually inspected and adapted where needed (Supplementary Table S2).

As a proof-of-concept, we began with the core conserved ncRNAs 6S and 4.5S, which were not already captured by alignments in the Rfam database. Secondary structure is central to the function of 6S RNA as an RNA polymerase sponge that mimics the transcription bubble in chromosomal DNA (Barrick et al., 2005; Wassarman, 2018). Indeed, WAR proposed a consensus structure for this 190 nt-long RNA similar to the 6S structure in Proteobacteria, with the major difference that the central bulge region was substantially larger in Bacteroidetes 6S (~70 nt) than in its proteobacterial counterparts (~20–30 nt) (Supplementary Figure S2a). The conformation of 4.5S RNA is best defined in *E. coli*, where it adopts a structure containing a number of internal loops connected by helical regions

and a conserved, functionally important GGAA tetraloop (Jagath et al., 2001; Larsen & Zwieb, 1991). However, our WAR predictions suggested *Bacteroides* 4.5S RNA to instead fold into a conformation consisting of a stem-loop with four internal bulges—three small ones and a large one (comprising 56 nt) near the tip—and a conserved terminal loop (Supplementary Figure S2b).

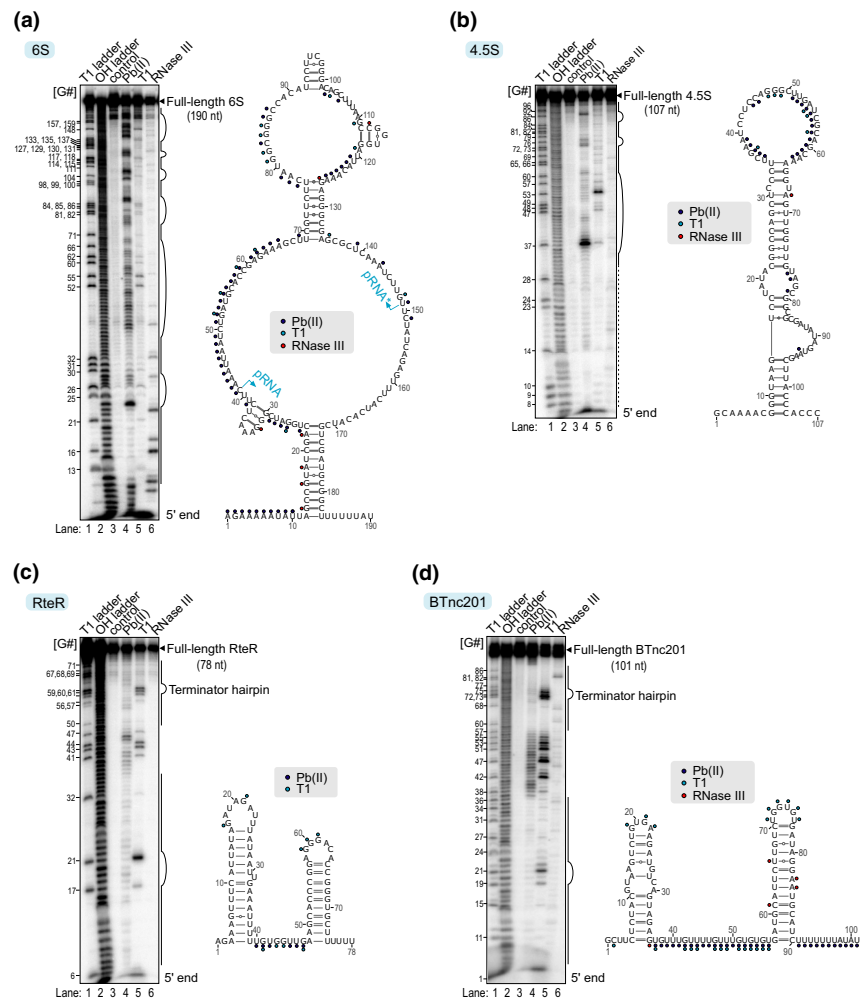
In order to improve the predicted consensus structures, we performed *in vitro* chemical and enzymatic probing of 6S and 4.5S from *B. thetaiotaomicron* VPI-5482. Structure probing validated the extensive central bulge region of *Bacteroides* 6S RNA (Figure 3a). Of note, the synthesis regions of product RNAs (pRNA, pRNA\*) that result from the reactivation of RNA polymerase and transcription using 14–20 nt sequences within 6S RNA as templates (Wassarman & Saecker, 2006) fall within the central bulge in the same fashion as in proteobacterial 6S (blue arrows in Figure 3a). We revised our computational prediction of the 6S secondary structure so that it was compatible with both probing data and base-pair conservation in our alignment (Figure 4a). We evaluated it using R-scape (Rivas et al., 2017), a statistical test for base covariation in secondary structures, and found evidence for evolutionary selection for base-pairing in three of the four predicted stem structures (asterisks in Figure 4a). In contrast, single sequence probing of 4.5S RNA was less conclusive and the manually inferred structure quite different from that inferred

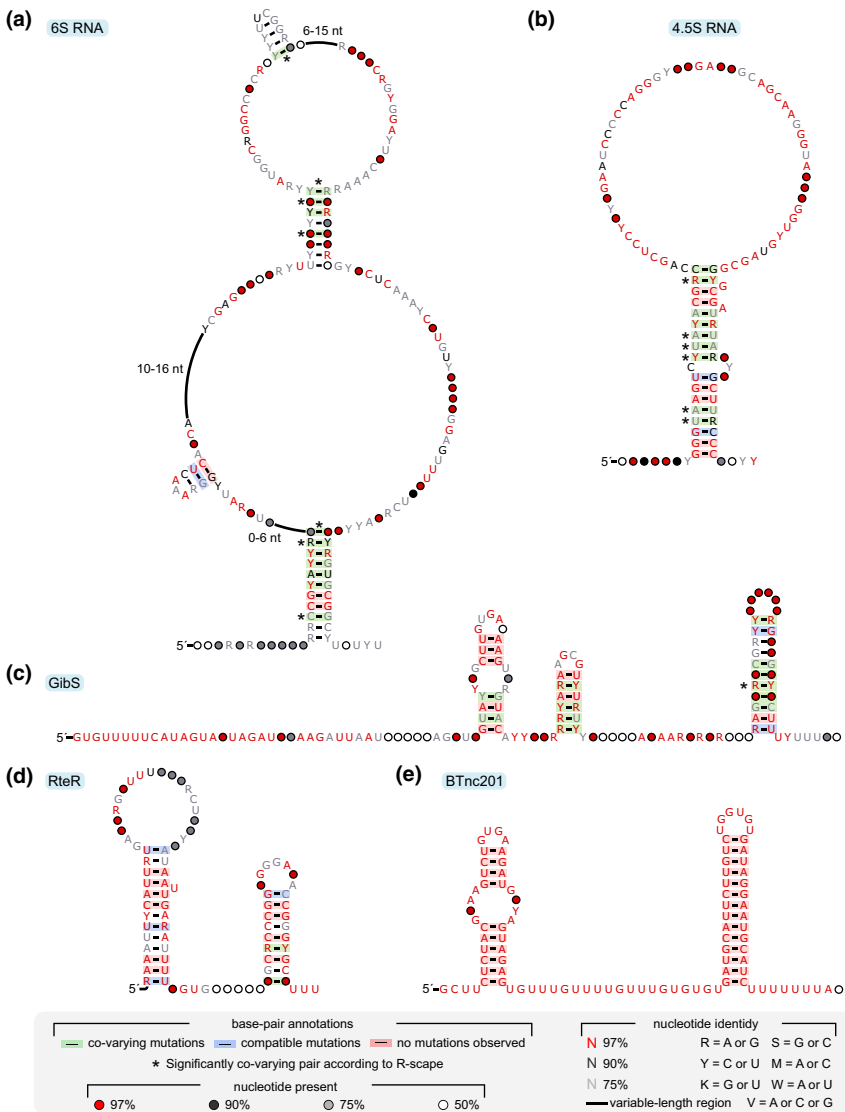
from sequence alignment (compare Figure 3b; Supplementary Figure S2b). However, unlike the 6S structure, which was supported by numerous cleavage events, few unambiguous cleavage events were detected in 4.5S (Figure 3b). We resolved this by using these unambiguous cleavage events as structural constraints in RNAalifold (Bernhart et al., 2008), resulting in a single long stem-loop structure well supported by covariation (Figure 4b).

## 2.5 | Secondary structure determination of *trans*-encoded sRNAs in *B. thetaiotaomicron*

The GibS and RteR sRNAs are conserved within *Bacteroides* spp. and are currently the only two functionally characterized *trans*-encoded sRNAs in this genus (Ryan et al., 2020a; Waters & Salyers, 2012). The WAR-derived GibS structure prediction (Supplementary Figure S2c) was similar to that previously determined by *in vitro* probing (Ryan et al., 2020a) and required only minor manual editing for consistency (Figure 4c). The secondary structure of RteR had previously been inferred from Mfold MFE predictions using the *B. thetaiotaomicron* CTnDOT sequence (Waters & Salyers, 2012), including a 5' hairpin and a Rho-independent terminator at the 3' end, interspaced by an eight nt-long single-stranded stretch. Both computational

**FIGURE 3** Structure probing of selected *Bacteroides thetaiotaomicron* ncRNAs. *In vitro* structure probing of 5' end-labeled 6S RNA (a), 4.5S RNA (b), CTn1-encoded RteR (c), and BTnc201 (d). T1 and OH ladders refer to partial digestion under denaturing conditions with nuclease T1 (Lane 1: cleaves unpaired G residues) or alkali (Lane 2: cleaves at all positions), respectively. "Control" (Lane 3) refers to the untreated RNA substrate. Lanes 4–6 reveal cleavages induced by lead (II) acetate (cleaves single-stranded nucleotides), RNase T1, or RNase III (cleaves extended double-stranded regions), respectively. For each RNA, one representative gel picture (from two replicates) is depicted to the left and the secondary structure, as deduced from the band pattern, is drawn to the right with individual cleavages indicated by the colored dots. The light blue arrows in Panel a refer to the transcription start sites for 6S product RNAs (pRNA, pRNA\*) as inferred from Ryan et al. (2020a)





**FIGURE 4** Curated models of *Bacteroides thetaiotaomicron* ncRNA sequence conservation and structure. Examples of curated structures for selected *Bacteroides* sRNA families, namely 6S RNA (a), 4.5S RNA (b), GibS (c), RteR (d), and BTnc201 (e). RNA structures were visualized using the R2R software (Weinberg & Breaker, 2011) and statistically significant base covariations according to R-scape (Rivas et al., 2017) are labeled by asterisks. See Supplementary Figure S2a–e for the corresponding original predictions, prior to manual curation

predictions based on the alignment of *Bacteroides* RteR homologs using WAR (Supplementary Figure S2d) and our structural probing (Figure 3c) yielded structures largely similar to previous predictions as reflected in our final model (Figure 4d). Finally, BTnc201 is an example of an uncharacterized, narrowly conserved sRNA that we previously identified and validated by northern blot as a highly abundant, 101 nt-long transcript (Ryan et al., 2020a) and found above to exhibit a BT<sub>1884</sub>-dependent expression pattern (Figure 1d). Present in only a few *Bacteroides* genomes, covariation information for this sRNA is not available due to the high level of sequence conservation within these strains. Our structural probing results (Figure 3d) largely agreed with the predicted structure of BTnc201 based purely on MFE (Supplementary Figure S2e): a bulged 5' hairpin, a 17 nt-long single-stranded region, followed by a strong intrinsic terminator, and a run of uridines. The final structure for BTnc201 required only minor revision at the 5' end (Figure 4e).

Altogether, the results of applying our computational protocol (Barquist et al., 2016) have been largely corroborated by experimental probing, with the obvious exception of 4.5S. We have

further constructed RNA families for an additional four uncharacterized *Bacteroides* sRNAs (Supplementary Figure S2f–i): BTnc005, BTnc049, BTnc060, and BTnc231. We note, however, that due to their narrow conservation, the structures of these latter sRNAs are not supported by extensive covariation, rendering the predicted conformations less reliable. Upon manual curation (Supplementary Table S2), all these families have been deposited in the Rfam database (Supplementary Table S3).

## 2.6 | Synteny and target co-occurrence analyses for conserved sRNAs

Gene neighborhood can be predictive of cellular processes and biological functions of an uncharacterized genetic element. With regard to sRNAs, examples from *Salmonella* illustrate that a transcription factor regulating the expression of a given sRNA (e.g., HilD activating InvR sRNA; Pfeiffer et al., 2007) or the gene for a bona fide sRNA target (e.g., *rtsA* mRNA and PinT sRNA; Kim et al., 2019) can



be encoded in close proximity to the respective sRNA gene. In fact, we recently made a similar observation for *B. thetaiotaomicron* sRNA GibS, as one of its targets, *BT\_0771*, is encoded directly downstream of the *gibS* gene itself (Ryan et al., 2020a).

Inspired by a previous study on the proteobacterial sRNA SgrS (Horler & Vanderpool, 2009), we performed gene synteny analyses for some of the most conserved (Figure 2c), yet unstudied, *Bacteroides* sRNAs. BTnc060, for example, was a candidate sRNA with a predicted size of ~220 nt (Ryan et al., 2020a). Here, northern blotting revealed the existence of two stable BTnc060 variants, a longer isoform (matching the predicted 220 nt) and a shorter one (~80 nt; Supplementary Figure S1d). The BTnc060 sequence from *B. thetaiotaomicron* is partially conserved (~90% nucleotide identity) in *B. xylanisolvans* and different strains of *B. fragilis* (Figure 2c). In these genomes, the BTnc060-flanking loci encode genes for RNA metabolism (RNases BN and P1), NADH nitroreductase, riboflavin synthase, and cytochrome genes downstream as well as a phosphate ABC transporter cluster upstream of the sRNA gene (Supplementary Figure S3a). A second *B. thetaiotaomicron* sRNA, BTnc231, is also encoded in the genomes of *B. helcogenes*, *B. xylanisolvans*, and *B. fragilis* (~90% nucleotide identity; Figure 2c). Previously, we validated BTnc231 on northern blots as a ~180 nt-long sRNA (Ryan et al., 2020a). Synteny analysis now revealed this sRNA to be encoded on a locus in between a DNA/RNA helicase gene and a catabolic gene set (Supplementary Figure S3b). The biological implications of these associations are currently not clear and warrant further investigation.

We also included GibS in our analysis since—despite being partially characterized (Ryan et al., 2020a)—its physiological role is still poorly understood. In the *B. thetaiotaomicron* type strain VPI-5483, where this sRNA was first discovered (Ryan et al., 2020a), GibS is encoded in between a putative para-aminobenzoate synthase cluster (*BT\_0763–68*) and a glycogen biosynthesis operon (*BT\_0769–71*), with the latter being a direct target of GibS. This genomic organization is partially conserved in a fraction of *Bacteroides* spp., including diverse *B. fragilis* strains and *B. xylanisolvans*; *B. helcogenes* encodes the glycogen biosynthesis gene cluster downstream of GibS but lacks the upstream para-aminobenzoate synthase (Figure 5a). Genomic colocalization of GibS with metabolic gene clusters may reflect its known function as an “RNA switch” that adapts *Bacteroides* metabolism to altered nutrient availability (Ryan et al., 2020a). Since the individual GibS homologs are often found in genomic vicinity to tRNA genes (orange in Figure 5a), the sRNA might be spread via horizontal transfer (Williams, 2002) across, and potentially even within, individual *Bacteroides* genomes. In support of the latter, *B. dorei* and *B. vulgatus* possess a second *gibS* copy in their chromosome, in between an efflux pump operon and a long-chain fatty acid ligase gene and its cognate transcriptional regulator (Figure 5a). This could imply further roles of GibS, for example, in the control of the *Bacteroides* cell envelope and transmembrane transport processes—a common theme among proteobacterial sRNAs (Guillier et al., 2006).

In an attempt to assess the completeness of our previous GibS target identification screen through transcriptomic profiling (Ryan et al., 2020a), we here performed an *in silico* co-occurrence analysis

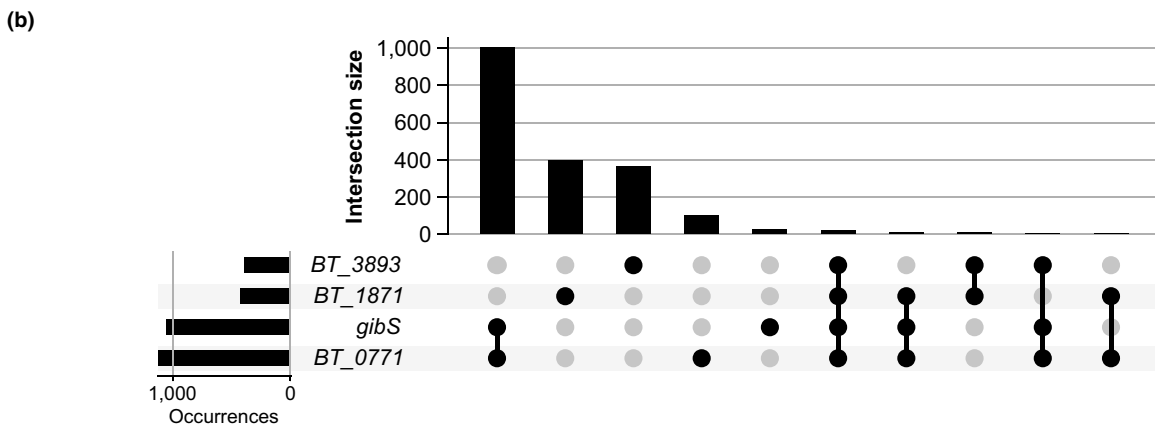
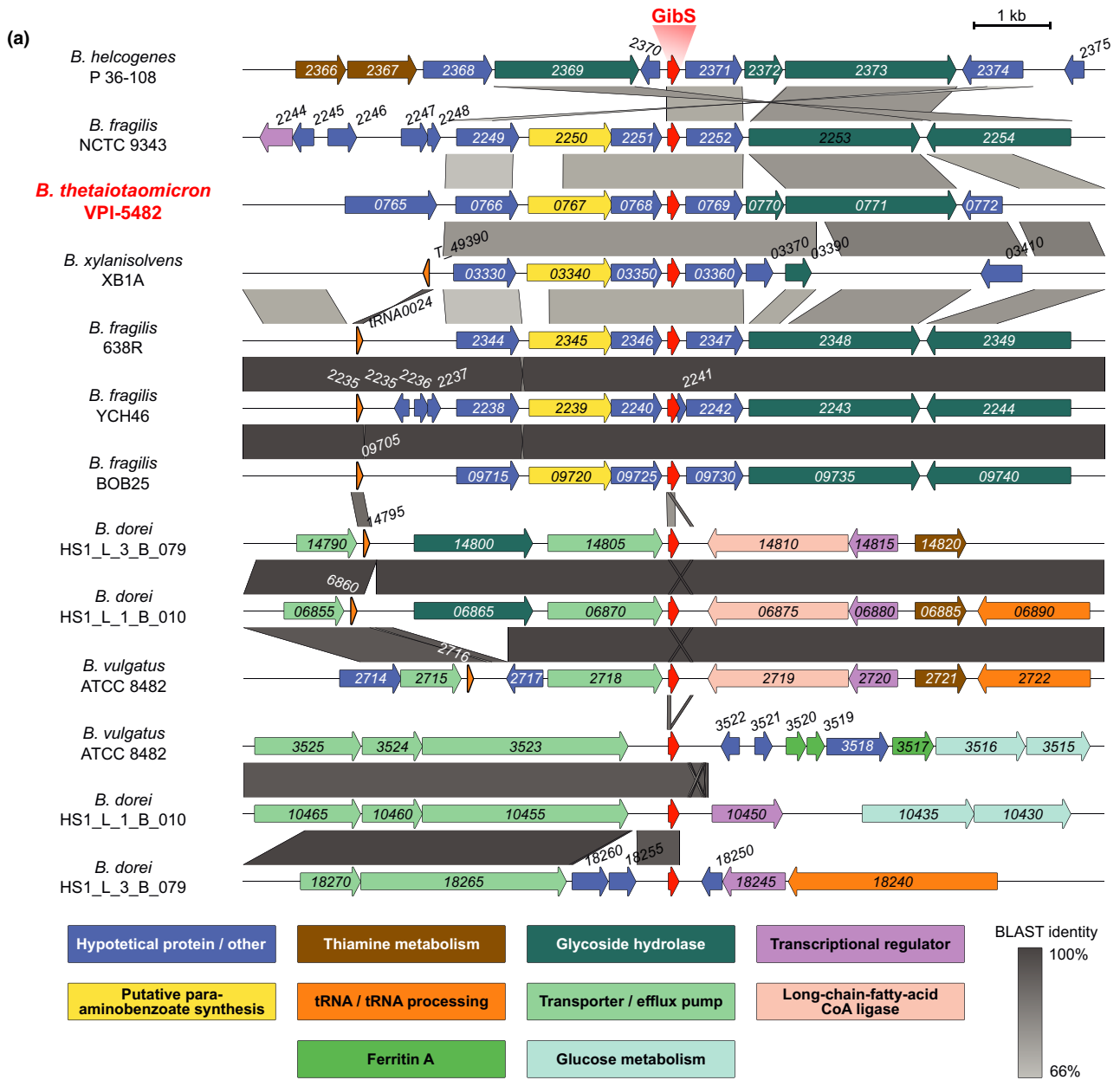
between this sRNA and its targets across *Bacteroides* spp. (Figure 5b). The known GibS target *BT\_0771* (Ryan et al., 2020a) exhibits an overall similar prevalence pattern to the sRNA itself. While this might suggest that regulation is also conserved beyond *B. thetaiotaomicron*, we note that genomic linkage likely contributes to their co-occurrence, as GibS is encoded adjacent to the *BT\_0769–0771* operon. We also report ~30 instances, where GibS was present in a genome devoid of any of its three known target genes, arguing that this sRNA could indeed have additional, currently unknown targets. Future studies might employ approaches orthogonal to transcriptomic profiling, such as sRNA affinity purification and sequencing (Correia Santos et al., 2021; Lalaouna et al., 2015), to more fully define GibS targets and gain further insight into the physiological role of this sRNA.

### 3 | DISCUSSION

sRNAs are key to the success of bacteria to occupy dynamic niches and survive in the face of diverse stresses (Holmqvist & Wagner, 2017; Wagner & Romby, 2015). A basic molecular toolkit to experimentally study the function of sRNAs and their interacting RBPs has been developed for the model proteobacterial species *E. coli* and *Salmonella enterica* (Hör et al., 2020; Sharma & Vogel, 2009). However, these methods have not yet been widely transferred to obligate anaerobic bacteria (Ryan et al., 2020b), hampering the functional and mechanistic study of sRNAs in many medically relevant gut microbiota species. In parallel to establishing comparable experimental tools, computational methods offer a fast-track solution to uncover sRNA functions in those species. Here, we employed a set of computational approaches to study global and specific aspects of sRNA biology in *Bacteroides*—a predominant bacterial genus in the human intestinal microbiota.

#### 3.1 | Identification of K homology, RNA recognition motif, and cold-shock domain proteins in *Bacteroides*

In Gram-negative bacteria, sRNA functionality often depends on RNA chaperones (Holmqvist & Vogel, 2018). Hfq-dependent sRNAs, for example, associate with this widely conserved RBP, which protects them from cellular ribonucleases, facilitates annealing to complementary regions within target transcripts, and may recruit endonucleases to cleave the bound target, effecting rapid decay (Kavita et al., 2018; Santiago-Frangos & Woodson, 2018; Vogel & Luisi, 2011). FinO domain proteins (Olejniczak & Storz, 2017), such as proteobacterial ProQ (Holmqvist et al., 2020), have only recently emerged as global RBPs but seem to also stabilize their ligands, thereby aiding sRNA-mediated control. CsrA is a prevalent regulatory protein that binds to GGA motifs within loop regions in mRNA 5' UTRs and influences translation initiation, an activity that in many bacteria is controlled by antagonistic sRNAs that sequester the protein in an inactive complex (Romeo & Babitzke, 2018). However, when we computationally searched for *Bacteroides* RBP candidates,



no homolog for any of the above global RBPs was identified. Rather, our screen pinpointed other *B. thetaiotaomicron* proteins with putative KH, RRM-1, or CSD RNA-binding domains.

Each of these three domains derives from the small  $\beta$ -barrel “urfold”—a recurrent feature of global RBPs (Youkharibache et al., 2019). KH, RRM-1, and CSD consist of each  $\sim$ 70 amino acids, can

**FIGURE 5** Synteny and target co-occurrence analyses for GibS homologs across *Bacteroides* spp. (a), Comparison of the regions flanking *gibS* in *Bacteroides* genomes for which a homolog of this sRNA was predicted. For each strain, the 5,000 nucleotides upstream and downstream of the *gibS* gene (red arrow) are plotted. Genes within these regions are represented as arrows and colored according to their predicted function (see legend). Trapezoids connecting regions of adjacent genomes indicate the BLAST-derived sequence identity, on a gradient from 66% (light gray) to 100% (dark gray) identity. Gene locus tags (without the literal prefix, e.g., “BT\_” for *Bacteroides thetaiotaomicron*) are included within or above each gene arrow. (b), Co-occurrence analysis of GibS and its three established targets *BT\_0771*, *BT\_1871*, and *BT\_3893* (Ryan et al., 2020a) across *Bacteroides* genomes. The top bar graph indicates the number of genomes that contain a single gene or specific combination of the four genes, as denoted by the black dots below each bar. The bar graph to the left shows the number of genomes in which each gene was detected

bind to single-stranded regions within nucleic acids, and are prevalent in all three domains of life (Graumann & Marahiel, 1998; Maris et al., 2005; Valverde et al., 2008). Type II KH domains—which are predominant in prokaryotes—consist of three  $\beta$ -strands, two of which are in parallel orientation. Nucleic acid binding occurs in a hydrophobic cleft formed between two  $\alpha$ -helices and a variable loop (Nicastro et al., 2015). KH domains are, for example, contained within PNPase and ribosomal protein S3, where they mediate binding to RNA ligands, but also in NusA-like transcription elongation proteins, in which they initiate binding to chromosomal DNA. The Firmicute KH-containing protein KhpB (also known as EloR) has been recently described as a new globally acting RBP in Gram-positives (Lamm-Schmidt et al., 2021; Winther et al., 2019; Zheng et al., 2017). However, no KhpB homolog was predicted for *Bacteroides* spp. by our screen. Rather, the KH domain proteins found showed homology to known transcriptional regulators, ribosomal factors, or ribonucleases.

RRM domains consist of each four antiparallel  $\beta$ -strands and two  $\alpha$ -helices, and are best understood in the Eukarya, where they engage in various posttranscriptional processes (Maris et al., 2005). While our study was in review, the three *B. thetaiotaomicron* RRM-1-containing proteins (Figure 1c) were independently predicted as *Bacteroides* RBPs (Adams et al., 2021). What is more, the deletion of two of them—renamed to RbpA (BT\_0784) and RbpB (BT\_1887)—resulted in widespread gene expression changes. Additionally, purified RbpB bound single-stranded RNA baits with high affinity *in vitro*, suggesting these RRM proteins function as global RBPs in the Bacteroidetes phylum (Adams et al., 2021).

CSDs also adopt a  $\beta$ -barrel structure and the binding of CSPs remodels the folding of their RNA ligands (Phadtare & Severinov, 2010). The best-studied CSP, *E. coli* CspA, is produced upon a temperature decrease to 20°C and modulates both the transcription and translation of its mRNA ligands to grant survival below the optimum growth temperature (Bae et al., 2000; Giuliodori et al., 2010; Jiang et al., 1997). Specifically, CspA prevents the formation of inhibitory secondary structures in the 5' portion of mRNAs to overcome premature transcription termination and facilitate ribosome binding and translation initiation in the cold (Jiang et al., 1997). Despite their name, not all CSPs are induced at low temperatures. *Salmonella* CspC and -E are highly expressed at 37°C and were recently identified as global RBPs, with a largely overlapping ligand set of mRNAs and sRNAs that seem to be stabilized by CspC/E binding (Michaux et al., 2017).

Here, we predicted the conserved *Bacteroides* CSP BT\_1884 as a putative RBP in this genus. Compared with proteobacterial CspC and -E, *Bacteroides* BT\_1884 bore an N-terminal extension, which may indicate functional differences between the Bacteroidetes and Proteobacteria homologs. Nevertheless, BT\_1884 effectively cross-complemented the phenotypes resulting from *cspC/E* double knockout in *Salmonella*. Our initial characterization, however, cannot yet answer whether or not BT\_1884 may act as a global RBP in its native host. Deletion or overexpression had no influence on *B. thetaiotaomicron* growth in rich medium. At the same time, CspC/E levels similarly do not affect *Salmonella* growth, at least under optimal conditions (Michaux et al., 2017). In *Salmonella*, altered CspC/E levels are reflected in the steady-state levels of their direct RNA ligands (Michaux et al., 2017). In the absence of RIP-seq (RNA immune-precipitation and sequencing) or CLIP-seq (cross-linking immunoprecipitation-high-throughput sequencing) data, we randomly selected *Bacteroides* sRNAs for probing. This revealed 2 out of 11 sRNAs to have mildly elevated expression levels when BT\_1884 was overexpressed, although their cellular half-lives remained largely unchanged.

The binding of CSPs and the consequent suppression of local RNA folding may also impact RNA processing. For example, the plant CSP GRP2 melts secondary structures—and thereby likely interferes with the processing—of its RNA ligands (Kim et al., 2007; Sasaki et al., 2007). Likewise, specific CSD-containing proteins associate with rRNA and are implicated in ribosomal maturation. For instance, the proteobacterial cold-induced ribosome-binding factor A (RbfA) binds 16S rRNA and is required for efficient processing of this ribosomal constituent (Dammel & Noller, 1995; Gualerzi et al., 2003). Here, our presented data suggest that BT\_1884 affects 5S rRNA processing in *B. thetaiotaomicron*. Of note, BTnc013—one of the putative sRNAs that were upregulated when this CSP was overexpressed—is a 5' UTR-derived sRNA. Future studies might therefore explore whether BT\_1884 plays a general role in *Bacteroides* RNA remodeling and processing.

Absence of Hfq, ProQ, and CsrA homologs, yet the presence of RRM-1- and CSD-containing proteins in the Bacteroidetes phylum might be indicative of an RNA biology fundamentally different from that of Proteobacteria, whose members have served us as bacterial models of RNA biology for decades. This, in turn, could imply that many new RNA-related mechanisms and functions await discovery and should foster future studies of RBPs in *Bacteroides* and other microbiota members.

### 3.2 | What secondary structures imply about *Bacteroides* ncRNA biology

Global *Bacteroides* ncRNA characteristics, including the number of individual sRNA representatives per genome, their average length and degree of folding, and generally narrow conservation all are in line with the sRNA complement of well-studied aerobic model species, or indeed all other bacteria studied to date. This raises interesting questions about sRNA evolution (Jose et al., 2019): how exactly do these transcripts arise? How do they acquire function? And perhaps most importantly, how can we discriminate functional sRNAs from “transcriptional noise”? High-throughput techniques that assay sRNA function genome-wide through targeted pull-downs and/or crosslinking are promising approaches to begin answering these questions. Unfortunately, most protocols developed to date crucially depend on having a known central sRNA chaperone (Melamed et al., 2018), such as Hfq or ProQ, to query, limiting applications in *Bacteroides* for the time being.

The presented data reveal additional interesting facets of *Bacteroides* RNA biology. For instance, the consensus secondary structure of *Bacteroides* 6S RNA is a bulged hairpin, with pRNA transcription initiating from the central bulge region, as previously described for well-characterized proteobacterial 6S RNA (Barrick et al., 2005; Wassarman, 2018). However, the central bulge of *Bacteroides* 6S RNA is substantially larger than that of any other described 6S molecule. It is thus tempting to speculate that the absence of the canonical sigma factor ( $\sigma^{70}$ ; encoded by the gene *rpoD*) and the presence of a different “housekeeping” sigma factor in this phylum ( $\sigma^{\text{ABfr}}$ ; Bayley et al., 2000; Vingadassalom et al., 2005) could be responsible for the divergent 6S structure. In other words, our findings might imply that the transcription bubble in actively transcribed genomic DNA, which is mimicked by 6S RNA to sequester the RNA polymerase holoenzyme, may be larger in the Bacteroidetes than in other phyla.

The bacterial SRP, consisting of 4.5S RNA and the Ffh protein, binds hydrophobic residues in nascent polypeptides. The ribonucleoprotein complex then docks to its membrane receptor FtsY, thereby cotranslationally inserting proteins into the membrane or tethering them for transport out of the cytosol (Peluso et al., 2000). Reflecting this housekeeping function, the SRP system is essential across the bacterial kingdom including in *B. thetaiotaomicron* (Goodman et al., 2009; Liu et al., 2021). Our analysis was unable to resolve a central tetraloop structure in the *Bacteroides* 4.5S RNA by either experimental structure probing or computational conservation-based analysis. In *E. coli*, this loop contains the sequence GGAA, which is important for stimulating GTPase activity of the SRP–FtsY complex (Siu et al., 2007) and whose mutation abolishes SRP function (Jagath et al., 2001). Interestingly, additional base pair disruptions in the tetraloop-adjacent stem partially restore *E. coli* 4.5S RNA function (Jagath et al., 2001). Consequently, the extended loop of our predicted *Bacteroides* 4.5S

structure might represent a molecular compromise to retain SRP–receptor interaction in the absence of a GGAA tetraloop, though it remains unclear if the *Bacteroides* 4.5S adopts a different central structure that may fulfill a similar function.

Our structural compendium of *Bacteroides* ncRNA families was deposited in the Rfam database and will be useful for the community as a resource for future mechanistic studies. For example, certain RBPs such as FinO-like proteins (Olejniczak & Storz, 2017), exhibit high affinities toward RNA stem-loop structures. Although no FinO domain-containing homolog is present in *Bacteroides*, similar structural preferences can be envisaged for other RBPs. In fact, both cold-shock and RRM-1 domains tend to bind single-stranded stretches within target RNA (Maris et al., 2005; Phadtare & Severinov, 2010). The presented folding predictions could thus be interrogated in the future to search for a common structural code for RBP ligands.

### 3.3 | Conclusion and perspective

Studying RNA processes in bacteria has had a tremendous impact on modern science, ranging from basic research to biotechnology and medicine (Pickar-Oliver & Gersbach, 2019; Serganov & Patel, 2007; Vogel, 2020). It is important, however, to emphasize that our current knowledge of bacterial RNA biology is strongly biased toward that of Gammaproteobacteria and Bacilli, whose species have been harnessed as model Gram-negative or Gram-positive organisms, respectively, for decades. In contrast, due to technical hurdles associated with experimentation with obligate anaerobic species, as of now we only have a vague idea of the RNA complement of most gut microbiota species. Mapping and functionally analyzing this “RNA microbiome” bears great potential to uncover novel functions and mechanisms employed by bacterial RNA molecules and the proteins they interact with. While experimental toolkits are being developed to study RNA-related processes in different representatives of the microbiota (Fuchs et al., 2021; Ponath et al., 2021; Ryan et al., 2020b), informatic approaches, such as comparative genomics, offer an entry point to explore RNAs expressed by human microbiotas (Fremin & Bhatt, 2021; Weinberg et al., 2010). In the present study, we have leveraged *in silico* tools and combined them with experimental approaches to improve our understanding of the RNA biology of the gut microbiota member *B. thetaiotaomicron*. This provided insights into potential RBPs, ncRNA structure, and function in *Bacteroides* spp. Collectively, our findings imply that much is still to be learned from the RNA biology of obligate anaerobes. Even GibS—arguably the best-characterized Bacteroidetes sRNA to date (Ryan et al., 2020a)—confronts us with many open questions regarding its mechanistic function and physiological role. However, the present study may serve as starting ground for future targeted investigations, to functionally dissect and—eventually—exploit the RNA of these predominant gut bacteria.

## 4 | EXPERIMENTAL PROCEDURES

### 4.1 | Bacterial strains, plasmids, and growth conditions

*B. thetaiotaomicron* VPI-5482 is referred to as wild type throughout this study and was grown as previously described (Ryan et al., 2020a). All strains, plasmids, and oligonucleotides used in this study are indicated in Supplementary Table S1. Growth curves were generated on a Biotek Epoch2 microplate reader, inoculating fresh tryptone yeast extract glucose (TYG) medium with 1:100 dilutions of overnight cultures of the respective strain.

### 4.2 | Phylogenetic tree visualization

The bacterial phylogenetic tree shown in Figure 1a was obtained from GTDB release 95 (Parks et al., 2020) and pruned with the APE R package (Paradis & Schliep, 2019). For simplicity, we kept the historical assignment of *B. dorei* and *B. vulgatus* to the *Bacteroides*, rather than the *Phocaicola* genus of GTDB r95.

### 4.3 | In silico search for RNA-binding domain-containing proteins

Hidden Markov models (HMMs) from Pfam A (release 32) were downloaded from the Pfam ftp server, and used with hmmer (version 3.2.1; Eddy, 2011) to annotate the *B. thetaiotaomicron* VPI-5482 (accession AE015928.1) proteome using the Pfam trusted bit score cutoff as a threshold for reporting results. The same was done for other bacterial species (see Figure 1b). These annotations were then manually searched for the presence of known RNA-binding domains. Domain hits for Hfq (PF17209), ProQ (PF04352), CsrA (PF02599), KH (PF00013, PF07650, and PF17905), RRM (PF00076, PF04059, PF08777, PF10598, PF13893, PF16367, and PF11835), or CSD (PF00313) were plotted in Figure 1b. *B. thetaiotaomicron* proteins containing hits were further investigated including the determination of their domain structure using the Pfam website sequence search (see Figure 1c).

### 4.4 | Generation of *BT\_1884* mutant, complementation, and overexpression strains

*BT\_1884* was deleted from the *B. thetaiotaomicron* VPI-5482 genome as previously described (Bencivenga-Barry et al., 2020), generating strain AWS-186. Briefly, plasmid AWP-039 (Supplementary Table S1) was generated by cloning 1 kb regions upstream and downstream of *BT\_1884* (amplified with oligonucleotides AWO-543/AWO-544 and AWO-545/AWO-546, respectively) into the multiple cloning site of pSIE1 (a gift from Andrew Goodman,

Addgene plasmid #136355; <http://n2t.net/addgene:136355>; RRID:Addgene\_136355). The plasmid was then conjugated into *B. thetaiotaomicron*. Transconjugant colonies (confirmed by PCR with oligonucleotides AWO-539/AWO-547) were grown overnight in TYG medium and streaked on brain heart infusion agar plates containing anhydrotetracycline. Mutants that had correctly excised the plasmid backbone were identified by PCR (oligonucleotides AWO-539/AWO-563) and the modified genomic sequence was confirmed by Sanger sequencing.

*BT\_1884* complementation (AWS-193) and overexpression (AWS-192) strains were generated by conjugating plasmids AWP-036 and AWP-035, respectively, into the  $\Delta BT_1884$  strain (AWS-186). Both plasmids, generated through Gibson cloning, contain the NBU2 integration system (Lim et al., 2017) that promotes their integration into the *B. thetaiotaomicron* genome. AWP-035 was generated by replacing the promoter and *GFP* cassette present in pWW3452 (Whitaker et al., 2017) with the native promoter and ribosome binding site of the *BT\_1887*–*BT\_1884* operon (100 bp upstream of the transcription start site), followed by the coding sequence of *BT\_1884* (using oligonucleotides AWO-571/AWO-572, AWO-573/AWO-570). Plasmid AWP-036 was similarly generated by replacing *GFP* in pWW3452 with the coding sequence of *BT\_1884*, but keeping the original phage promoter and strong RBS (oligonucleotides AWO-567/AWO-568, AWO-569/AWO-570).

Heterologous expression of *BT\_1884* in *Salmonella enterica* serovar Typhimurium was achieved by transforming plasmid AWP-044 into the  $\Delta cspCE$  *Salmonella* background (JVS-5084; Michaux et al., 2017). Cloning of AWP-044 was done as in (Michaux et al., 2017). Briefly, the coding sequence of *BT\_1884* was amplified with oligonucleotides AWO-654 and AWO-655 and the amplicon ligated in the *Xba*I site of pBAD33.

### 4.5 | RNA extraction, northern blotting, and quantitative real-time PCR (qRT-PCR)

*B. thetaiotaomicron* was grown overnight in 5 ml TYG medium and total RNA was extracted from stationary phase cultures (OD  $\approx$  4.8) as previously described (Ryan et al., 2020a). For northern blotting, DNase I-digested (NEB) total RNA (5  $\mu$ g/lane) was denatured for 5 min at 95°C followed by 5 min incubation on ice. Samples were run on a TBE 6% (vol/vol) polyacrylamide (PAA) 7 M urea gel at 300 V, 150 mA for  $\sim$ 2 hr and blotted onto nylon membranes (Sigma #15356) at 50 V, 4°C for 1 hr. After UV cross-linking the RNA to the membrane, blots were probed with <sup>32</sup>P-labeled gene-specific oligonucleotides (Supplementary Table S1) in Hybri-Quick buffer (Carl Roth AG) at 42°C and exposed as required on phosphor screens. Signals were visualized using a phosphorimager (FLA-3000 Series, Fuji). qRT-PCR was performed using Takyon SYBR MasterMix (Eurogentech #UF-NSMT-B0701) with the addition of EuroScript II reverse transcriptase (Eurogentech #UF-RTAD-D0701) on 10 ng DNase-digested total RNA per reaction.



#### 4.6 | RNA stability assay

To monitor cellular half-lives of *B. thetaiotaomicron* ncRNAs in different *BT\_1884* backgrounds, we applied rifampicin-mediated inhibition of de novo transcription. Briefly, *B. thetaiotaomicron* strains AWS-001 (WT), AWS-186 ( $\Delta BT_1884$ ), and AWS-192 (*BT\_1884*<sup>+</sup>) were inoculated from single colonies and grown for ~16 hr in liquid TYG to an OD of ~4.8 (the condition where we observed altered steady-state levels; see Figure 1d) and then treated with rifampicin (final concentration: 500  $\mu$ g/ml) to halt transcription. RNA samples were drawn at the indicated time points after rifampicin addition and RNA decay was assessed by northern blot analysis.

#### 4.7 | Polymyxin B sensitivity assay

Polymyxin B survival assays were performed as in Michaux et al. (2017) with minor modifications. Overnight cultures of the indicated *Salmonella* strains were subcultured 1:100 in Lysogeny Broth (LB) and grown for 2 hr at 37°C, 220 rpm. L-arabinose (0.2% final concentration) was included in the overnight and subculturing media to induce the expression of CSPs in the complementation strains. Polymyxin B (Sigma-Aldrich #81271) was then added at a final concentration of 2  $\mu$ g/ml and cultures were incubated for a further hour at 37°C, 220 rpm. Before and after the polymyxin challenge, serial dilutions of the cultures were prepared and each 5  $\mu$ l spotted on LB agar plates to determine bacterial survival.

#### 4.8 | Swimming assay

The motility of *Salmonella* strains was assessed via swimming assays as previously described (Michaux et al., 2017). Twenty milliliters of soft LB-agar (0.3% agar; prewarmed to 60°C) were poured per Petri dish on the lab bench and dried with an open lid for 15 min. L-arabinose (0.2% final concentration) was supplemented to the overnight media and agar plates to induce the expression of CSPs in the complementation strains. Then, 10  $\mu$ l of each bacterial overnight culture were spotted on the center of a plate, followed by incubation for 5.5 hr at 37°C and measurement of swimming distances.

#### 4.9 | Interspecies sRNA structuredness comparison

We define the structuredness of an sRNA as the deviation (Z score) of its predicted MFE from the distribution of the MFEs of 1,000 sequences of the same length as the sRNA and having the same dinucleotide frequency as the respective genome. For Figure 2b, we computed MFE values with Vienna RNAfold (Lorenz et al., 2011) and generated the 1,000 background sequences with esl-shuffle (-L

length -N 1000 -d options, with length being the sRNA length) from the hmmer3 suite (Eddy, 2011).

#### 4.10 | Manual curation/extension of *B. thetaiotaomicron* ncRNA annotations

We reannotated previous sRNA candidate BTnc259 (Ryan et al., 2020a) as the *B. thetaiotaomicron* 4.5S RNA homolog (*ffs* gene). Reannotation was based on a keyword search for the term “Bacteroides” in the Rfam database with subsequent hits filtered for “Bacterial small-signal recognition particle RNA.” The resulting sequence table file was downloaded and filtered for species of the Bacteroidetes/Chlorobi group (Figure 1a), retrieving strain *Chlorobium phaeobacteroides* DSM 266 (NC\_008639.1). A subsequent BLAST search of the annotated *ffs* gene in that strain identified a sequence overlapping with the BTnc259 annotation. A further search for BTnc259 in the Rfam database along with our in silico and in vitro structure analyses supported this candidate to indeed be the 4.5S RNA of *B. thetaiotaomicron* VPI-5482.

RteR was originally described as an sRNA transcribed from a locus downstream of the *exc* gene on conjugative transposon CTnDOT (Jeters et al., 2009; Waters & Salyers, 2012). However, RteR was not included in our original annotation of sRNAs in *B. thetaiotaomicron* strain VPI-5482 (Ryan et al., 2020a). A BLAST search for the *exc* gene suggested *BT\_0105* (annotated as a type IA DNA topoisomerase) within the conjugative transposon CTn1 as a putative *exc* homolog. By manually inspecting our previous RNA-seq data (Ryan et al., 2020a) mapping to this locus, we identified a lowly expressed, ~75 nt-long transcript derived from a region downstream of *BT\_0105*, associated with a canonical  $\sigma^{ABFr}$  promoter (Bayley et al., 2000), and not harboring any potential open reading frame (Stothard, 2000). It was suggested previously that the secondary structure of RteR is important for its function as a repressor of conjugative transfer (Waters & Salyers, 2012). Based on the shared genomic organization and similar RNA secondary structure (Waters & Salyers, 2012), we propose this CTn1-encoded RNA as a functional homolog of RteR.

#### 4.11 | Analysis of ncRNA conservation

A custom genome database was constructed using bacterial genomes belonging to class Bacteroidia (Taxonomy ID: 200643) and selected from the ENA (<https://www.ebi.ac.uk/genomes/bacteria.html>, accessed 1/12/2017). An iterative search of each ncRNA candidate using nhmmer (Wheeler & Eddy, 2013) was performed as previously described (Lindgreen et al., 2014). With each iteration, nhmmer was set with the flags, -popen 0.4999 -E 0.001 -incE 0.001. Additionally, hits with E values  $\leq 0.001$  were required to have almost full-length alignments with no more than 10% missing sequence at either end. The resultant alignment served as input for hmmbuild and was utilized in the next iteration. Alignments were

subsequently manually inspected using the RALEE editor (Griffiths-Jones, 2005).

#### 4.12 | In vitro transcription and radiolabeling of RNA

RNA candidates selected for structure probing were in vitro-transcribed using primer pairs AWO-243/-244 (BTnc201), AWO-442/-443 (6S RNA), AWO-458/-459 (4.5S RNA), and AWO-503/-504 (RteR) carrying a T7 promoter in the sense oligonucleotide to amplify genomic DNA templates. In vitro transcription reactions of these templates were performed using the MEGAscript T7 kit (Ambion) and the DNA templates were subsequently digested with DNase I (NEB). The in vitro-transcribed RNA bands of interest were excised from a 6% (vol/vol) PAA 7 M urea gel and extracted in elution buffer (0.1 M NaOAc, 0.1% SDS, 10 mM EDTA) at 8°C and 1,400 rpm, overnight. The RNA was precipitated using a 30:1 mix of ethanol and NaOAc, washed with chilled 75% ethanol, and resuspended in 20–30 µl water at 65°C. Dephosphorylation of 50 pmol of the above RNA was carried out using 25 U of calf intestinal alkaline phosphatase (NEB) in a 50 µl volume and incubated at 37°C for 1 hr. The dephosphorylated RNA was extracted using phenol, chloroform, isoamylalcohol (25:24:1) and precipitated as described above. Twenty picomoles of the RNA were 5' end-labeled using <sup>32</sup>P-γATP (20 µCi) and polynucleotide kinase (NEB) at 37°C for 1 hr in a reaction volume of 20 µl. The labeled RNA was then column-purified (G-50, GE Healthcare) and gel-extracted as above.

#### 4.13 | Structure probing

Structure probings were performed in 10 µl reaction volumes using radiolabeled RNA (0.2 pmol) that was first denatured at 95°C for 1 min and then chilled on ice for 5 min. Following the addition of 1 µg yeast RNA and 10x structure buffer (Ambion), the reactions were incubated for 10 min at 37°C. To these reactions, either 2 µl of fresh lead (II) acetate (25 mM, Fluka), 2 µl of RNase T1 (0.01 U/µl, Ambion), or 2 µl RNase III (NEB) and 1 mM dithiothreitol were added and incubated at 37°C for 45 s, 3 min or 10 min, respectively. The reactions were stopped by addition of 12 µl loading buffer II (Ambion) and placed on ice. Alkaline hydrolysis ladders were prepared by incubating 0.4 pmol labelled RNA with 9 µl of 1x alkaline hydrolysis buffer (Ambion). RNase T1 ladders were similarly prepared in 8 µl of 1x sequencing buffer (Ambion) and incubated at 95°C for 1 min prior to the addition of 1 µl RNase T1 (0.1 U/µl). Both reactions were incubated for 5 min at 37°C and stopped as above. Prior to loading on an 8% (vol/vol) PAA 7 M urea sequencing gel, all samples were denatured at 95°C for 3 min.

#### 4.14 | Curation of ncRNA families

The curation of our RNA families largely followed the protocol described in Barquist et al. (2016). Initial sequences were collected

using nhmmer as described above in “Analysis of ncRNA conservation” then fed to the WAR webserver. The resulting alignments were manually evaluated and edited for consistency using the RALEE emacs RNA editor mode (Griffiths-Jones, 2005) by at least four independent editors, and further evaluated using the R-scape webserver (Rivas et al., 2017). Details of editing are provided in Supplementary Table S2, “Consensus structure curation”. These alignments and structures were further revised in light of structure probing data as described in the main text, where relevant, and deposited on the Rfam database (Supplementary Table S3).

#### 4.15 | sRNA synteny analysis

Comparison plots of 5,000 nucleotides flanking the respective sRNA's transcriptional start site were generated through Easyfig (Sullivan et al., 2011). The plots were then manually curated to include gene functional annotation from NCBI (Tatusova et al., 2016). Where available, a function was assigned to the uncharacterized proteins with eggNOG-mapper version 2.0.1b (Huerta-Cepas et al., 2017) using the following flags: `-m diamond -d none -tax_scope auto -go_evidence non-electronic -target_orthologs all -seed_ortholog_evalue 0.001 -seed_ortholog_score 60 -query-cover 20 -subject-cover 0`.

#### 4.16 | sRNA-target co-occurrence analysis

The nucleotide sequences of GibS, BT\_0771, BT\_1871, and BT\_3893 were searched with blastn (7 nt word size) (Altschul et al., 1997) against the refseq\_genomes database filtered for *Bacteroides* (taxid 816). Hits with an *E* value of >0.01 or query cover of <75% were discarded. The results of the co-occurrence analysis (Figure 5b) were visualized with the UpSetPlot Python package (<https://github.com/jnothman/UpSetPlot>).

#### ACKNOWLEDGMENTS


We thank Charlotte Michaux for advice on *Salmonella* motility assays, Jörg Vogel for sharing *Salmonella* strains, and Nancy Ontiveros and Anton Petrov for assistance with depositing alignments in the Rfam database. Research in the Westermann lab is supported by the German Research Foundation (DFG): Individual Research Grant We6689/1-1. Open Access funding enabled and organized by Projekt DEAL.

#### DATA AVAILABILITY STATEMENT

The results of the integrated structure and sequence analysis of *B. thetaiotaomicron* ncRNAs were deposited in the Rfam database (RF04177 to RF04184) and are publicly available as of release 14.6. The existing family RF01693 was updated with a description indicating it may be the Bacteroidetes 6S, though the underlying alignment has not been updated as of publication. The original Stockholm format alignments underlying the conclusions in this manuscript have been deposited on Zenodo with <https://doi.org/10.5281/zenodo.5156251> (<https://doi.org/10.5281/zenodo.5156250>).

## ORCID

Gianluca Prezza  <https://orcid.org/0000-0002-0032-4369>

Alexander J. Westermann  <https://orcid.org/0000-0003-3236-0169>

[org/0000-0003-3236-0169](https://orcid.org/0000-0003-3236-0169)

## REFERENCES

- Adams, A.N.D., Azam, M.S., Costliow, Z.A., Ma, X., Degnan, P.H. & Vanderpool, C.K. (2021) A novel family of RNA-binding proteins regulate polysaccharide metabolism in *Bacteroides thetaiotaomicron*. *Journal of Bacteriology*. [Epub ahead of print]. <https://doi.org/10.1128/JB.00217-21>
- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25, 3389–3402.
- Argaman, L., Hershberg, R., Vogel, J., Bejerano, G., Wagner, E.G., Margalit, H. & Altuvia, S. (2001) Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. *Current Biology*, 11, 941–950.
- Bae, W., Xia, B., Inouye, M. & Severinov, K. (2000) *Escherichia coli* CspA-family RNA chaperones are transcription antiterminators. *Proceedings of the National Academy of Sciences USA*, 97, 7784–7789.
- Barquist, L., Burge, S.W. & Gardner, P.P. (2016). Studying RNA homology and conservation with infernal: from single sequences to RNA families. *Current Protocols in Bioinformatics*, 54, 12.13.11–12.13.25.
- Barrick, J.E., Sudarsan, N., Weinberg, Z., Ruzzo, W.L. & Breaker, R.R. (2005) 6S RNA is a widespread regulator of eubacterial RNA polymerase that resembles an open promoter. *RNA*, 11, 774–784.
- Bauriedl, S., Gerovac, M., Heidrich, N., Bischler, T., Barquist, L., Vogel, J. & Schoen, C. (2020) The minimal meningococcal ProQ protein has an intrinsic capacity for structure-based global RNA recognition. *Nature Communications*, 11, 2823.
- Bayley, D.P., Rocha, E.R. & Smith, C.J. (2000) Analysis of cepA and other *Bacteroides fragilis* genes reveals a unique promoter structure. *FEMS Microbiology Letters*, 193, 149–154.
- Bencivenga-Barry, N.A., Lim, B., Herrera, C.M., Trent, M.S. & Goodman, A.L. (2020) Genetic manipulation of wild human gut *Bacteroides*. *Journal of Bacteriology*, 202, e00544–19.
- Bernhart, S.H., Hofacker, I.L., Will, S., Gruber, A.R. & Stadler, P.F. (2008) RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, 9, 474.
- Buffie, C.G. & Pamer, E.G. (2013) Microbiota-mediated colonization resistance against intestinal pathogens. *Nature Reviews Immunology*, 13, 790–801.
- Cao, Y., Forstner, K.U., Vogel, J. & Smith, C.J. (2016) cis-encoded small RNAs, a conserved mechanism for repression of polysaccharide utilization in *Bacteroides*. *Journal of Bacteriology*, 198, 2410–2418.
- Chao, Y. & Vogel, J. (2010) The role of Hfq in bacterial pathogens. *Current Opinion in Microbiology*, 13, 24–33.
- Chen, H., Shiroguchi, K., Ge, H. & Xie, X.S. (2015) Genome-wide study of mRNA degradation and transcript elongation in *Escherichia coli*. *Molecular Systems Biology*, 11, 808.
- Consortium, R.N. (2021) RNACentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Research*, 49, D212–D220.
- Correia Santos, S., Bischler, T., Westermann, A.J. & Vogel, J. (2021) MAPS integrates regulation of actin-targeting effector SteC into the virulence control network of *Salmonella* small RNA PinT. *Cell Reports*, 34, 108722.
- Dammel, C.S. & Noller, H.F. (1995) Suppression of a cold-sensitive mutation in 16S rRNA by overexpression of a novel ribosome-binding factor, RbfA. *Genes & Development*, 9, 626–637.
- Dugar, G., Herbig, A., Forstner, K.U., Heidrich, N., Reinhardt, R., Nieselt, K. & Sharma, C.M. (2013) High-resolution transcriptome maps reveal strain-specific regulatory features of multiple *Campylobacter jejuni* isolates. *PLoS Genetics*, 9, e1003495.
- Eddy, S.R. (2011) Accelerated profile HMM searches. *PLoS Computational Biology*, 7, e1002195.
- El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C. et al. (2019) The Pfam protein families database in 2019. *Nucleic Acids Research*, 47, D427–D432.
- Fremin, B.J. & Bhatt, A.S. (2021) Comparative genomics identifies thousands of candidate structured RNAs in human microbiomes. *Genome Biology*, 22, 100.
- Fu, Y., Deiorio-Haggar, K., Anthony, J. & Meyer, M.M. (2013) Most RNAs regulating ribosomal protein biosynthesis in *Escherichia coli* are narrowly distributed to Gammaproteobacteria. *Nucleic Acids Research*, 41, 3491–3503.
- Fuchs, M., Lamm-Schmidt, V., Sulzer, J., Ponath, F., Jenniches, L., Kirk, J.A. et al. (2021) An RNA-centric global view of *Clostridioides difficile* reveals broad activity of Hfq in a clinically important gram-positive bacterium. *Proceedings of the National Academy of Sciences of the United States of America*, 118, e2103579118.
- Gardner, P.P., Daub, J., Tate, J.G., Nawrocki, E.P., Kolbe, D.L., Lindgreen, S. et al. (2009) Rfam: updates to the RNA families database. *Nucleic Acids Research*, 37, D136–D140.
- Giuliodori, A.M., Di Pietro, F., Marzi, S., Masquida, B., Wagner, R., Romby, P., Gualerzi, C.O. & Pon, C.L. (2010) The cspA mRNA is a thermosensor that modulates translation of the cold-shock protein CspA. *Molecular Cell*, 37, 21–33.
- Glowacki, R.W.P. & Martens, E.C. (2020) If you eat it, or secrete it, they will grow: the expanding list of nutrients utilized by human gut bacteria. *Journal of Bacteriology*, 203, e00481–20.
- Gonzalez, G.M., Hardwick, S.W., Maslen, S.L., Skehel, J.M., Holmqvist, E., Vogel, J. et al. (2017) Structure of the *Escherichia coli* ProQ RNA-binding protein. *RNA*, 23, 696–711.
- Goodman, A.L., McNulty, N.P., Zhao, Y., Leip, D., Mitra, R.D., Lozupone, C.A., Knight, R. & Gordon, J.I. (2009) Identifying genetic determinants needed to establish a human gut symbiont in its habitat. *Cell Host & Microbe*, 6, 279–289.
- Graumann, P.L. & Marahiel, M.A. (1998) A superfamily of proteins that contain the cold-shock domain. *Trends in Biochemical Sciences*, 23, 286–290.
- Griffiths-Jones, S. (2005) RALEE–RNA alignment editor in Emacs. *Bioinformatics*, 21, 257–259.
- Gualerzi, C.O., Giuliodori, A.M. & Pon, C.L. (2003) Transcriptional and post-transcriptional control of cold-shock genes. *Journal of Molecular Biology*, 331, 527–539.
- Guillier, M., Gottesman, S. & Storz, G. (2006) Modulating the outer membrane with small RNAs. *Genes & Development*, 20, 2338–2348.
- Harrison, P.W., Ahamed, A., Aslam, R., Alako, B.T.F., Burgin, J., Buso, N., Courtot, M., Fan, J. et al. (2021) The European nucleotide archive in 2020. *Nucleic Acids Research*, 49, D82–D85.
- Holmqvist, E., Berggren, S. & Rizvanovic, A. (2020) RNA-binding activity and regulatory functions of the emerging sRNA-binding protein ProQ. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1863, 194596.
- Holmqvist, E., Li, L., Bischler, T., Barquist, L. & Vogel, J. (2018) Global maps of ProQ binding in vivo reveal target recognition via RNA structure and stability control at mRNA 3' ends. *Molecular Cell*, 70(971–982), e976.
- Holmqvist, E. & Vogel, J. (2018) RNA-binding proteins in bacteria. *Nature Reviews Microbiology*, 16, 601–615.
- Holmqvist, E. & Wagner, E.G.H. (2017) Impact of bacterial sRNAs in stress responses. *Biochemical Society Transactions*, 45, 1203–1212.
- Hooper, L.V., Falk, P.G. & Gordon, J.I. (2000) Analyzing the molecular foundations of commensalism in the mouse intestine. *Current Opinion in Microbiology*, 3, 79–85.

- Hör, J., Matera, G., Vogel, J., Gottesman, S. & Storz, G. (2020) Transacting small RNAs and their effects on gene expression in *Escherichia coli* and *Salmonella enterica*. *EcoSal Plus*, 9. <https://doi.org/10.1128/ecosalplus.ESP-0030-2019>
- Horler, R.S. & Vanderpool, C.K. (2009) Homologs of the small RNA SgrS are broadly distributed in enteric bacteria but have diverged in size and sequence. *Nucleic Acids Research*, 37, 5465–5476.
- Howden, B.P., Beaume, M., Harrison, P.F., Hernandez, D., Schrenzel, J., Seemann, T., Francois, P. & Stinear, T.P. (2013) Analysis of the small RNA transcriptional response in multidrug-resistant *Staphylococcus aureus* after antimicrobial exposure. *Antimicrobial Agents and Chemotherapy*, 57, 3864–3874.
- Huerta-Cepas, J., Forslund, K., Coelho, L.P., Szklarczyk, D., Jensen, L.J., von Mering, C. & Bork, P. (2017) Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Molecular Biology and Evolution*, 34, 2115–2122.
- Irnov, I., Sharma, C.M., Vogel, J. & Winkler, W.C. (2010) Identification of regulatory RNAs in *Bacillus subtilis*. *Nucleic Acids Research*, 38, 6637–6651.
- Jagath, J.R., Matassova, N.B., de Leeuw, E., Warnecke, J.M., Lentzen, G., Rodnina, M.V., Luirink, J. & Wintermeyer, W. (2001) Important role of the tetraloop region of 4.5S RNA in SRP binding to its receptor FtsY. *RNA*, 7, 293–301.
- Jain, C. (2008) The *E. coli* RhlE RNA helicase regulates the function of related RNA helicases during ribosome assembly. *RNA*, 14, 381–389.
- Jeters, R.T., Wang, G.R., Moon, K., Shoemaker, N.B. & Salyers, A.A. (2009) Tetracycline-associated transcriptional regulation of transfer genes of the *Bacteroides* conjugative transposon CTnDOT. *Journal of Bacteriology*, 191, 6374–6382.
- Jiang, W., Hou, Y. & Inouye, M. (1997) CspA, the major cold-shock protein of *Escherichia coli*, is an RNA chaperone. *Journal of Biological Chemistry*, 272, 196–202.
- Jose, B.R., Gardner, P.P. & Barquist, L. (2019) Transcriptional noise and exaptation as sources for bacterial sRNAs. *Biochemical Society Transactions*, 47, 527–539.
- Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E.P., Rivas, E., Eddy, S.R., Bateman, A., Finn, R.D. et al. (2018) Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Research*, 46, D335–D342.
- Kavita, K., de Mets, F. & Gottesman, S. (2018) New aspects of RNA-based regulation by Hfq and its partner sRNAs. *Biochemical Society Transactions*, 42, 53–61.
- Kim, J.S., Park, S.J., Kwak, K.J., Kim, Y.O., Kim, J.Y., Song, J., Jang, B., Jung, C.H. & Kang, H. (2007) Cold shock domain proteins and glycine-rich RNA-binding proteins from *Arabidopsis thaliana* can promote the cold adaptation process in *Escherichia coli*. *Nucleic Acids Res*, 35, 506–516.
- Kim, K., Palmer, A.D., Vanderpool, C.K. & Slauch, J.M. (2019) The small RNA PinT contributes to PhoP-mediated regulation of the *Salmonella* pathogenicity island 1 type III secretion system in *Salmonella enterica* Serovar Typhimurium. *Journal of Bacteriology*, 201, e00312–19.
- Koo, J.T., Alleyne, T.M., Schiano, C.A., Jafari, N. & Lathem, W.W. (2011) Global discovery of small RNAs in *Yersinia pseudotuberculosis* identifies *Yersinia*-specific small, noncoding RNAs required for virulence. *Proceedings of the National Academy of Sciences USA*, 108, E709–E717.
- Kroger, C., Colgan, A., Srikumar, S., Handler, K., Sivasankaran, S.K., Hammarlof, D.L., Canals, R., Grissom, J.E. et al. (2013) An infection-relevant transcriptomic compendium for *Salmonella enterica* Serovar Typhimurium. *Cell Host & Microbe*, 14, 683–695.
- Lalaouna, D., Carrier, M.C., Semsey, S., Brouard, J.S., Wang, J., Wade, J.T. & Masse, E. (2015) A 3' external transcribed spacer in a tRNA transcript acts as a sponge for small RNAs to prevent transcriptional noise. *Molecular Cell*, 58, 393–405.
- Lamm-Schmidt, V., Fuchs, M., Sulzer, J., Gerovac, M., Hör, J., Dersch, P., Vogel, J. & Faber, F. (2021) Grad-seq identifies KhpB as a global RNA-binding protein in *Clostridioides difficile* that regulates toxin production. *microLife*, 2, uqab004. <https://doi.org/10.1093/femsm/l/uaq004>
- Larsen, N. & Zwieb, C. (1991) SRP-RNA sequence alignment and secondary structure. *Nucleic Acids Research*, 19, 209–215.
- Lim, B., Zimmermann, M., Barry, N.A. & Goodman, A.L. (2017) Engineered regulatory systems modulate gene expression of human commensals in the gut. *Cell*, 169(547–558), e515.
- Lindgreen, S., Umu, S.U., Lai, A.S., Eldai, H., Liu, W., McGimpsey, S., Wheeler, N.E., Biggs, P.J., Thomson, N.R., Barquist, L., Poole, A.M. & Gardner, P.P. (2014) Robust identification of noncoding RNA from transcriptomes requires phylogenetically-informed sampling. *PLoS Computational Biology*, 10, e1003907.
- Liu, H., Shiver, A.L., Price, M.N., Carlson, H.K., Trotter, V.V., Chen, Y., Escalante, V., Ray, J. et al. (2021) Functional genetics of human gut commensal *Bacteroides thetaiotaomicron* reveals metabolic requirements for growth across environments. *Cell Reports*, 34, 108789.
- Lorenz, R., Bernhart, S.H. Höner zu Siederdisen, C., Tafer, H., Flamm, C., Stadler, P.F. & Hofacker, I.L. (2011) ViennaRNA package 2.0. *Algorithms for Molecular Biology*. *AMB*, 6, 26.
- Madison, J.T., Everett, G.A. & Kung, H. (1966) Nucleotide sequence of a yeast tyrosine transfer RNA. *Science*, 153, 531–534.
- Maris, C., Dominguez, C. & Allain, F.H. (2005) The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *FEBS Journal*, 272, 2118–2131.
- McCown, P.J., Corbino, K.A., Stav, S., Sherlock, M.E. & Breaker, R.R. (2017) Riboswitch diversity and distribution. *RNA*, 23, 995–1011.
- Melamed, S., Faigenbaum-Romm, R., Peer, A., Reiss, N., Shechter, O., Bar, A., Altuvia, Y., Argaman, L. & Margalit, H. (2018) Mapping the small RNA interactome in bacteria using RIL-seq. *Nature Protocols*, 13, 1–33.
- Michaux, C., Holmqvist, E., Vasicek, E., Sharan, M., Barquist, L., Westermann, A.J., Gunn, J.S. & Vogel, J. (2017) RNA target profiles direct the discovery of virulence functions for the cold-shock proteins CspC and CspE. *Proceedings of the National Academy of Sciences USA*, 114, 6824–6829.
- Nicastro, G., Taylor, I.A. & Ramos, A. (2015) KH-RNA interactions: back in the groove. *Current Opinion in Structural Biology*, 30, 63–70.
- Notredame, C., Higgins, D.G. & Heringa, J. (2000) T-coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, 302, 205–217.
- Olejniczak, M. & Storz, G. (2017) ProQ/FinO-domain proteins: another ubiquitous family of RNA matchmakers? *Molecular Microbiology*, 104, 905–915.
- Pandey, S., Gravel, C.M., Stockert, O.M., Wang, C.D., Hegner, C.L., LeBlanc, H. & Berry, K.E. (2020) Genetic identification of the functional surface for RNA binding by *Escherichia coli* ProQ. *Nucleic Acids Research*, 48, 4507–4520.
- Papenfert, K., Bouvier, M., Mika, F., Sharma, C.M. & Vogel, J. (2010) Evidence for an autonomous 5' target recognition domain in an Hfq-associated small RNA. *Proc Natl Acad Sci U S A*, 107, 20435–20440.
- Paradis, E. & Schliep, K. (2019) Ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35, 526–528.
- Parks, D.H., Chuvochina, M., Chaumeil, P.A., Rinke, C., Mussig, A.J. & Hugenholtz, P. (2020) A complete domain-to-species taxonomy for bacteria and archaea. *Nature Biotechnology*, 38, 1079–1086.
- Peer, A. & Margalit, H. (2011) Accessibility and evolutionary conservation mark bacterial small-rna target-binding regions. *Journal of Bacteriology*, 193, 1690–1701.
- Peer, A. & Margalit, H. (2014) Evolutionary patterns of *Escherichia coli* small RNAs and their regulatory interactions. *RNA*, 20, 994–1003.
- Peluso, P., Herschlag, D., Nock, S., Freymann, D.M., Johnson, A.E. & Walter, P. (2000) Role of 4.5S RNA in assembly of the bacterial signal recognition particle with its receptor. *Science*, 288, 1640–1643.



- Pfeiffer, V., Sittka, A., Tomer, R., Tedin, K., Brinkmann, V. & Vogel, J. (2007) A small non-coding RNA of the invasion gene island (SPI-1) represses outer membrane protein synthesis from the Salmonella core genome. *Molecular Microbiology*, *66*, 1174–1191.
- Phadtare, S. & Severinov, K. (2005) Nucleic acid melting by *Escherichia coli* CspE. *Nucleic Acids Research*, *33*, 5583–5590.
- Phadtare, S. & Severinov, K. (2010) RNA remodeling and gene regulation by cold shock proteins. *RNA Biology*, *7*, 788–795.
- Pickar-Oliver, A. & Gersbach, C.A. (2019) The next generation of CRISPR-Cas technologies and applications. *Nature Reviews Molecular Cell Biology*, *20*, 490–507.
- Ponath, F., Tawk, C., Zhu, Y., Barquist, L., Faber, F. & Vogel, J. (2021) RNA landscape of the emerging cancer-associated microbe *Fusobacterium nucleatum*. *Nature Microbiology*, *6*, 1007–1020.
- Richter, A.S. & Backofen, R. (2012) Accessibility and conservation: general features of bacterial small RNA-mRNA interactions? *RNA Biology*, *9*, 954–965.
- Rivas, E., Clements, J. & Eddy, S.R. (2017) A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nature Methods*, *14*, 45–48.
- Rivas, E., Klein, R.J., Jones, T.A. & Eddy, S.R. (2001) Computational identification of noncoding RNAs in *E. coli* by comparative genomics. *Current Biology*, *11*, 1369–1373.
- Romeo, T. & Babitzke, P. (2018) Global regulation by CsrA and its RNA antagonists. *Microbiology Spectrum*, *6*. <https://doi.org/10.1128/microbiolspec.RWR-0009-2017>
- Ryan, D., Jenniches, L., Reichardt, S., Barquist, L. & Westermann, A.J. (2020a) A high-resolution transcriptome map identifies small RNA regulation of metabolism in the gut microbe *Bacteroides thetaiotaomicron*. *Nature Communications*, *11*, 3557.
- Ryan, D., Prezza, G. & Westermann, A.J. (2020b) An RNA-centric view on gut bacteroidetes. *Biological Chemistry*, *402*, 55–72.
- Santiago-Frangos, A. & Woodson, S.A. (2018) Hfq chaperone brings speed dating to bacterial sRNA. *Wiley Interdisciplinary Reviews: RNA*, *9*, e1475.
- Sasaki, K., Kim, M.H. & Imai, R. (2007) Arabidopsis COLD SHOCK DOMAIN PROTEIN2 is a RNA chaperone that is regulated by cold and developmental signals. *Biochemical and Biophysical Research Communications*, *364*, 633–638.
- Serganov, A. & Patel, D.J. (2007) Ribozymes, riboswitches and beyond: regulation of gene expression without proteins. *Nature Reviews Genetics*, *8*, 776–790.
- Sharma, C.M., Hoffmann, S., Darfeuille, F., Reignier, J., Findeiss, S., Sittka, A., Chabas, S., Reiche, K. et al. (2010) The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature*, *464*, 250–255.
- Sharma, C.M. & Vogel, J. (2009) Experimental approaches for the discovery and characterization of regulatory small RNA. *Current Opinion in Microbiology*, *12*, 536–546.
- Sharma, C.M. & Vogel, J. (2014) Differential RNA-seq: the approach behind and the biological insight gained. *Current Opinion in Microbiology*, *19*, 97–105.
- Shinohara, A., Matsui, M., Hiraoka, K., Nomura, W., Hirano, R., Nakahigashi, K., Tomita, M., Mori, H. et al. (2011) Deep sequencing reveals as-yet-undiscovered small RNAs in *Escherichia coli*. *BMC Genomics*, *12*, 428.
- Siu, F.Y., Spangord, R.J. & Doudna, J.A. (2007) SRP RNA provides the physiologically essential GTPase activation function in cotranslational protein targeting. *RNA*, *13*, 240–250.
- Slager, J., Aprianto, R. & Veening, J.W. (2018) Deep genome annotation of the opportunistic human pathogen *Streptococcus pneumoniae* D39. *Nucleic Acids Research*, *46*, 9971–9989.
- Stein, E.M., Kwiatkowska, J., Basczok, M.M., Gravel, C.M., Berry, K.E. & Olejniczak, M. (2020) Determinants of RNA recognition by the FinO domain of the *Escherichia coli* ProQ protein. *Nucleic Acids Research*, *48*, 7502–7519.
- Stothard, P. (2000) The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *Biotechniques*, *28*(1102), 1104.
- Sullivan, M.J., Petty, N.K. & Beatson, S.A. (2011) Easyfig: a genome comparison visualizer. *Bioinformatics*, *27*, 1009–1010.
- Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E.P., Zaslavsky, L., Lomsadze, A., Pruitt, K.D. et al. (2016) NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Research*, *44*, 6614–6624.
- Torarinsson, E. & Lindgreen, S. (2008) WAR: Webserver for aligning structural RNAs. *Nucleic Acids Research*, *36*, W79–W84.
- Valverde, R., Edwards, L. & Regan, L. (2008) Structure and function of KH domains. *FEBS Journal*, *275*, 2712–2726.
- Vingadassalom, D., Kolb, A., Mayer, C., Rybkine, T., Collatz, E. & Podglajen, I. (2005) An unusual primary sigma factor in the *Bacteroidetes* phylum. *Molecular Microbiology*, *56*, 888–902.
- Vogel, J. (2020) An RNA biology perspective on species-specific programmable RNA antibiotics. *Molecular Microbiology*, *113*, 550–559.
- Vogel, J. & Luisi, B.F. (2011) Hfq and its constellation of RNA. *Nature Review Microbiology*, *9*, 578–589.
- Wagner, E.G.H. & Romby, P. (2015) Small RNAs in bacteria and archaea: who they are, what they do, and how they do it. *Advances in Genetics*, *90*, 133–208.
- Wassarman, K.M. (2018) 6S RNA, a global regulator of transcription. *Microbiology Spectrum*, *6*. <https://doi.org/10.1128/microbiolspec.RWR-0019-2018>
- Wassarman, K.M., Repoila, F., Rosenow, C., Storz, G. & Gottesman, S. (2001) Identification of novel small RNAs using comparative genomics and microarrays. *Genes Development*, *15*, 1637–1651.
- Wassarman, K.M. & Saecker, R.M. (2006) Synthesis-mediated release of a small RNA inhibitor of RNA polymerase. *Science*, *314*, 1601–1603.
- Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M. & Barton, G.J. (2009) Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, *25*, 1189–1191.
- Waters, J.L. & Salyers, A.A. (2012) The small RNA RteR inhibits transfer of the *Bacteroides* conjugative transposon CTnDOT. *Journal of Bacteriology*, *194*, 5228–5236.
- Wehner, S., Damm, K., Hartmann, R.K. & Marz, M. (2014) Dissemination of 6S RNA among bacteria. *RNA Biology*, *11*, 1467–1478.
- Weinberg, Z. & Breaker, R.R. (2011) R2R—software to speed the depiction of aesthetic consensus RNA secondary structures. *BMC Bioinformatics*, *12*, 3.
- Weinberg, Z., Wang, J.X., Bogue, J., Yang, J., Corbino, K., Moy, R.H. & Breaker, R.R. (2010) Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. *Genome Biology*, *11*, R31.
- Wexler, A.G. & Goodman, A.L. (2017) An insider's perspective: *Bacteroides* as a window into the microbiome. *Nature Microbiology*, *2*, 17026.
- Wexler, H.M. (2007) *Bacteroides*: the good, the bad, and the nitty-gritty. *Clinical Microbiology Reviews*, *20*, 593–621.
- Wheeler, T.J. & Eddy, S.R. (2013) nhmmer: DNA homology search with profile HMMs. *Bioinformatics*, *29*, 2487–2489.
- Whitaker, W.R., Shepherd, E.S. & Sonnenburg, J.L. (2017) Tunable expression tools enable single-cell strain distinction in the gut microbiome. *Cell*, *169*(538–546), e512.
- Williams, K.P. (2002) Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. *Nucleic Acids Research*, *30*, 866–875.
- Winther, A.R., Kjos, M., Stamsas, G.A., Havarstein, L.S. & Straume, D. (2019) Prevention of EloR/KhpA heterodimerization by introduction of site-specific amino acid substitutions renders the essential elongosome protein PBP2b redundant in *Streptococcus pneumoniae*. *Scientific Reports*, *9*, 3681.



- Woese, C.R., Gutell, R., Gupta, R. & Noller, H.F. (1983) Detailed analysis of the higher-order structure of 16S-like ribosomal ribonucleic acids. *Microbiology Reviews*, 47, 621–669.
- Wurtzel, O., Sesto, N., Mellin, J.R., Karunker, I., Edelheit, S., Becavin, C., Archambaud, C., Cossart, P. et al. (2012) Comparative transcriptomics of pathogenic and non-pathogenic *Listeria* species. *Molecular System Biology*, 8, 583.
- Youkharibache, P., Veretnik, S., Li, Q., Stanek, K.A., Mura, C. & Bourne, P.E. (2019) The small beta-barrel domain: a survey-based structural analysis. *Structure*, 27, 6–26.
- Zheng, J.J., Perez, A.J., Tsui, H.T., Massidda, O. & Winkler, M.E. (2017) Absence of the KhpA and KhpB (JAG/EloR) RNA-binding proteins suppresses the requirement for PBP2b by overproduction of FtsA in *Streptococcus pneumoniae* D39. *Molecular Microbiology*, 106, 793–814.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

**How to cite this article:** Prezza, G., Ryan, D., Mädler, G., Reichardt, S., Barquist, L. & Westermann, A.J. (2022) Comparative genomics provides structural and functional insights into *Bacteroides* RNA biology. *Molecular Microbiology*, 117, 67–85. <https://doi.org/10.1111/mmi.14793>