

A high-resolution transcriptome map identifies small RNA regulation of metabolism in the gut microbe *Bacteroides thetaiotaomicron*

Daniel Ryan¹, Laura Jenniches¹, Sarah Reichardt¹, Lars Barquist ^{1,2} & Alexander J. Westermann ^{1,3} 

Bacteria of the genus *Bacteroides* are common members of the human intestinal microbiota and important degraders of polysaccharides in the gut. Among them, the species *Bacteroides thetaiotaomicron* has emerged as the model organism for functional microbiota research. Here, we use differential RNA sequencing (dRNA-seq) to generate a single-nucleotide resolution transcriptome map of *B. thetaiotaomicron* grown under defined laboratory conditions. An online browser, called 'Theta-Base' (www.helmholtz-hiri.de/en/datasets/bacteroides), is launched to interrogate the obtained gene expression data and annotations of ~4500 transcription start sites, untranslated regions, operon structures, and 269 non-coding RNA elements. Among the latter is GibS, a conserved, 145 nt-long small RNA that is highly expressed in the presence of *N*-acetyl-D-glucosamine as sole carbon source. We use computational predictions and experimental data to determine the secondary structure of GibS and identify its target genes. Our results indicate that sensing of *N*-acetyl-D-glucosamine induces GibS expression, which in turn modifies the transcript levels of metabolic enzymes.

¹Helmholtz Institute for RNA-based Infection Research (HIRI), Helmholtz Centre for Infection Research (HZI), Würzburg, Germany. ²Faculty of Medicine, University of Würzburg, Würzburg, Germany. ³Institute of Molecular Infection Biology (IMIB), University of Würzburg, Würzburg, Germany.
✉email: alexander.westermann@uni-wuerzburg.de

Bacteroides are Gram-negative, obligate anaerobic, non-motile, non-spore forming rods, and among the most abundant bacterial genera in the human intestine¹. *Bacteroides thetaiotaomicron* has emerged as a model representative of the gut microbiota, due to its widespread distribution among human populations and the relative ease of studying these bacteria under laboratory conditions¹. The extensive metabolic potential encoded in the *B. thetaiotaomicron* genome, including 88 defined polysaccharide utilization loci (PULs)^{2,3}, forms the prerequisite for gut colonization. However, persistence within this notoriously dynamic niche requires coordinated control of gene expression in response to fluctuating nutrient levels.

Bacteroidetes lack the classical housekeeping sigma factor (σ^{70}) encoded by the proteobacterial *rpoD* gene, and consequently, the classical -10 and -35 boxes recognized by σ^{70} are absent from their promoters. Rather, members of this phylum contain an unusual, RpoD-like primary transcription factor, σ^{ABfr} , as well as an arsenal of alternative, extra-cytoplasmic function (ECF) sigma factors^{4,5}. The consensus recognition sequence of σ^{ABfr} was deduced by manual inspection of 23 promoter sequences in *B. fragilis* as 'TAnnTTTG' at the -7 region and a 'TTTG' motif at around the -33 region^{6,7}. Inspection of global RNA-seq data recently confirmed the -7 consensus in another Bacteroidetes member, *Flavobacterium johnsoniae*, whereas the -33 motif could not be identified⁸. While in *B. fragilis* the sequence motif recognized by the oxidative stress-related alternative ECF sigma-factor EcfO was recently identified⁹, no such motif is currently known for any of the *B. thetaiotaomicron* alternative ECFs. Besides sigma factors, SusR-like regulators¹⁰ and hybrid two-component signal transduction systems^{11,12}, wherein sensor kinase and response regulator are fused into a single polypeptide, contribute further to the adaptation of *Bacteroides* gene expression upon sensing of defined environmental cues, and DNA recognition motifs have been predicted for some of these regulators based on comparative genomics^{13,14}.

Knowledge of post-transcriptional control mechanisms is sparse in *B. thetaiotaomicron*. Lacking classical Shine-Dalgarno (SD) sequences, messenger RNAs (mRNAs) of Bacteroidetes were recently shown to be enriched for adenine residues at position -3, -6, and -11 to -15 relative to the translational start codon, with adenine overrepresentation at these positions positively correlating with translation efficiency⁸. In Proteobacteria, small regulatory RNAs (sRNAs) can regulate target mRNAs through imperfect base-pairing interactions that typically occlude the SD sequence and/or the start codon, thus interfering with translation initiation¹⁵. As of now, only two sRNAs are known in *Bacteroides* spp. RteR is a *trans*-encoded sRNA of 90 nt that promotes discoordinate expression of the *tra* operon, required to assemble the mating apparatus for the transfer of the conjugative transposon, CTnDot^{16,17}. DonS is a representative of a family of 78–128 nt-long *cis*-antisense RNAs divergently encoded to—and negatively impacting expression of—*susC* homologs of PUL systems involved in the binding and degradation of mucosa-derived glycans¹⁸. Lastly, a mRNA leader sequence was recently shown to tie *B. thetaiotaomicron* colonization to the presence of dietary sugars—and the authors speculated an sRNA could be involved in this process¹⁹. Together, these examples imply that RNA-mediated control mechanisms may be commonly employed by *B. thetaiotaomicron* to couple expression of metabolic genes to nutrient availability. However, the mode-of-action for all of these regulators is currently unknown, and RNA biology in *B. thetaiotaomicron* has not yet been investigated in a systematic manner. Particularly, given the lack of SD sequences in Bacteroidetes mRNAs, it is an open question if and how *trans*-encoded sRNAs could post-transcriptionally regulate target transcripts in this phylum.

Here, we performed differential RNA sequencing (dRNA-seq)^{20,21} of *B. thetaiotaomicron* grown in rich medium in three defined growth phases. This led to the annotation of a total of 4507 transcription start sites (TSSs) and untranslated regions (UTRs), as well as the identification of promoter motifs, RNA processing sites, and operon structures. To provide easy access to our transcriptome data, we have developed the intuitive Open-Access online database 'Theta-Base'. We report the identification of 269 noncoding RNA candidates from all major classes including *cis*- and *trans*-encoded, 5'- and 3'-derived, and intra-operonic sRNAs as well as riboswitches, RNA thermometers, and putative type-I toxin-antitoxin (TA) systems. We selected one of the newly identified intergenic sRNAs (GibS), determined its secondary structure, and—as the first example of a *trans*-encoded Bacteroidetes noncoding RNA—identified its target mRNAs in a genome-wide screen. Biochemical and genetic experiments revealed that GibS utilizes a highly conserved, single-stranded seed sequence in its 5' portion to mediate base-pairing with the translation initiation regions of two metabolic target mRNAs (*BT_0771*, *BT_3893*), resulting in repression of the respective transcripts. Altogether, the presented data imply that *B. thetaiotaomicron* employs riboregulatory mechanisms for adapting its gene expression to changing environmental conditions and should foster future studies to explore RNA biology in bacterial members of the human microbiota.

Results

TSSs of *B. thetaiotaomicron* VPI-5482 grown in rich medium. By selectively enriching triphosphates at the 5' end of primary transcripts, dRNA-seq identifies TSSs in a genome-wide manner, resulting in high-resolution transcriptome maps^{20,21}. To globally identify TSSs in *B. thetaiotaomicron*, we transferred the dRNA-seq protocol to total RNA samples extracted from the type strain *B. thetaiotaomicron* VPI-5482 in three defined growth phases in TYG (tryptone-yeast extract glucose) medium—at the transition from lag to exponential phase, in mid-exponential and stationary phase—and analyzed the resulting data using the ANNOgesic bioinformatic tool set²² (Fig. 1a). This approach reliably mapped TSSs in the *B. thetaiotaomicron* genome, as illustrated for the *roc* (*regulator of colonization*) mRNA for which dRNA-seq identified the exact TSS as previously mapped using primer-extension¹⁹.

In the *B. thetaiotaomicron* core genome and plasmid, we identified a total of 4507 TSSs that were classified into five categories based on their genomic location relative to annotated CDSs (Fig. 1b). Primary TSSs (pTSSs; 40% of all TSSs) were defined as the TSS with the highest coverage and secondary TSSs (sTSSs; 8%) as all remaining TSSs within a 300 nt window upstream of a given CDS. Internal TSSs (iTSSs; 21%) and antisense TSSs (aTSSs; 24%) arise from within a given CDS (or in case of an aTSS, within a 100 nt window flanking the CDS), in either sense or antisense orientation, respectively. The remaining 5% of predicted TSSs, not associated with any CDS, were classified as orphan (oTSSs) and might reflect an incomplete annotation of open reading frames or be indicative of intergenic sRNA genes (see below). Of the predicted TSSs, 1951 were detected in all growth phases with the total number of TSSs detected per condition ranging between 2265 and 3606 (Supplementary Data 1). The majority of TSSs were detected in stationary phase (~34% expressed exclusively under this condition; Supplementary Data 1). Functional analysis of genes with stationary phase-specific pTSSs using the PANTHER Classification System²³ revealed enrichment of gene ontology (GO)-terms including DNA recombination, DNA integration, and integral membrane components (Supplementary Data 1).

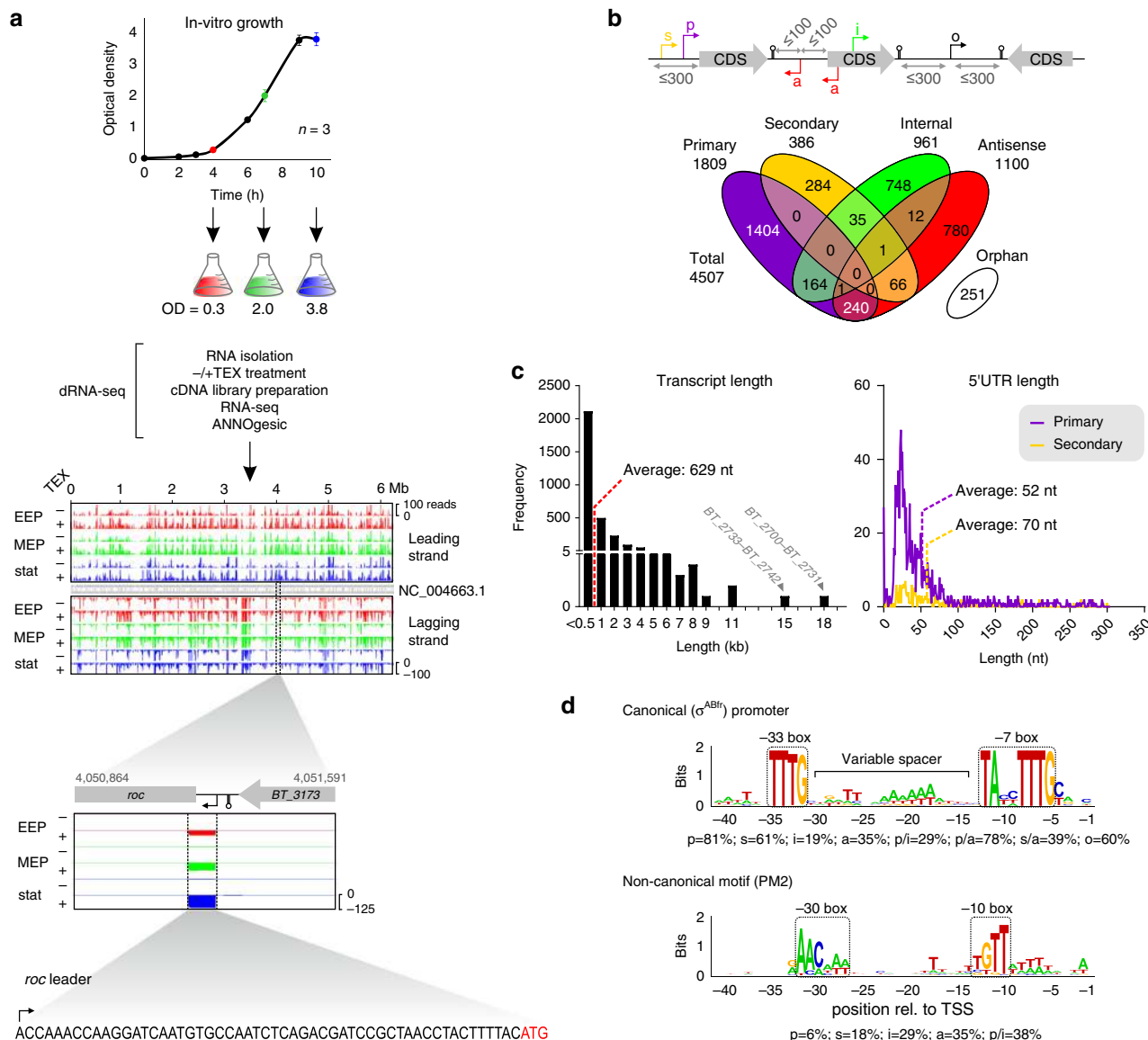


Fig. 1 High-resolution view at the *B. thetaiotaomicron* transcriptome. **a** Experimental workflow of the dRNA-seq approach and close-up view on the transcription start site (TSS; indicated by a bent arrow) of the *regulator of colonization* (*roc*) gene. Bacteria (AWS-001) from three different growth phases in TYG medium—early exponential (EEP), mid-exponential (MEP), and stationary phase (stat)—were analyzed. TEX, terminator exonuclease. Growth data refer to the mean \pm standard deviation from three biological replicates. For the read coverage plots, one representative replicate out of three is depicted. **b** Definition of categories for TSS annotation (upper) and overlap among TSS categories (lower). CDS, coding sequence. Note that pTSS and sTSS are mutually exclusive (thus no overlap in the Venn diagram). **c** Length distribution of transcripts (left) and 5' UTRs (right). Long operons are labeled by name. **d** Motif searches upstream of *B. thetaiotaomicron* TSSs reveal the canonical σ^{ABfr} promoter (upper) and a second, near-palindromic motif ('PM2'; lower). Percentage values below the respective motif denote the frequency of that sequence upstream of the respective TSS category as defined in **b**. Source data to this figure are provided in the Source Data file.

Antisense transcription is widespread in bacteria, with between 5 and 75% of CDSs exhibiting antisense transcription depending on bacterial species, experimental condition, and RNA-seq protocol²⁴. While functions for a number of specific antisense transcripts have been described²⁵, their relevance as a general functional class remains unclear. Manual inspection of the predicted aTSSs in *B. thetaiotaomicron* revealed ~22% of them to serve as pTSS for CDSs further downstream (i.e., exceeding the set window of 300 nt; Fig. 1b), whereas the majority (73%) of identified aTSSs gave rise to relatively short, weakly expressed transcripts, suggesting some aTSSs may arise from spurious transcription initiation rather than being associated with a functional transcript²⁶. For future analyses of specific antisense

transcripts in *B. thetaiotaomicron* we advise readers to consider the here-reported aTSSs and consult the provided database (which will be described below). Manual inspection of predicted iTSSs revealed several examples of transcriptionally interconnected genes/operons with related functions, e.g., the TSS for *BT_3336*, involved in the biosynthesis of lipid A, is located within the upstream *oprM* gene for an efflux pump involved in multidrug resistance, highlighting the link between cell envelope modifications and antibiotic stress.

Sequence alignment of all identified TSSs revealed the purine bases adenine and guanine as the preferred initiating nucleotides (45% or 41% of cases, respectively). As purine triphosphates serve as major energy storage molecules in cells, purine base

overrepresentation in initiator nucleotides may help couple transcriptional activity to metabolic state, as observed previously in aerobic bacteria^{27–29}.

Global assessment of transcript features. By combining TransTermHP³⁰ and RNAfold³¹, ANNOgesic predicts intrinsic transcription terminators, which—together with the experimentally mapped TSSs—enables the deduction of transcript boundaries. *B. thetaiotaomicron* expressed transcripts in a range from 20 nt to ~18 kb (Fig. 1c, left). The longest transcripts correspond to ribosomal operons (*BT_2700–BT_2731* and *BT_2733–BT_2742*), cell surface and iron transport operons (*BT_1950–BT_1958*), and metabolic gene clusters (*BT_1099–BT_1108*).

5' UTRs were inferred from the identified TSSs and annotated CDSs, revealing average and median lengths of 52 and 32 nt, respectively (Fig. 1c, right). Most mRNAs harbored a 5' UTR of 23 nt; ~10 nt shorter than the assumed optimal length for translation initiation in Proteobacteria^{20,32}. In line with leaderless mRNAs being considered rare in Gram-negative species³³, our dRNA-seq screen identified only ~3.7% *Bacteroides* 5' UTRs shorter than 10 nt, e.g. of mRNAs for several transposases (*BT_0280*, *BT_1996*), two-component sensor kinases (*BT_3967*, *BT_2166*), and transporters (*BT_0161*, *BT_0158*). In contrast, 13.5% of mRNAs had unusually long (>100 nt) 5' UTRs that might serve as targeting platforms for sRNAs or RNA-binding proteins (RBPs), or contain *cis*-regulatory RNA elements. Indeed, as inferred from homology to known RNA elements from other bacteria, ~2% of the long 5' UTRs were predicted to harbor a putative riboswitch (Supplementary Data 2). Manual inspection revealed yet other long 5' UTRs to actually contain short open reading frames (sORFs; Supplementary Data 3) reminiscent of leader peptides mediating transcription attenuation³⁴. For instance, the 5' UTR associated with the mRNA for the hypothetical protein *BT_4401* (462 nt in length) contains several putative sORFs (sORF_378 to –381), whose existence needs to be validated in the future.

Small proteins have long gone unnoticed, due to difficulties in annotation and detection, but recently this class of molecules has gained attention as biological functions for several representatives could be demonstrated^{35,36}. In total, ANNOgesic predicted 409 sORF candidates in *B. thetaiotaomicron* over growth in TYG medium (Supplementary Data 3). However, only seven of those are supported by a recent ribosome profiling study³⁷ (Supplementary Fig. 1a). To some extent, this divergence can be explained by the Ribo-seq study considering only conserved sORFs, thus deliberately accepting false-negatives for the sake of enriching true-positives. sORF candidates called exclusively in our screen might thus be strain-specifically encoded, condition-specifically expressed (growth for maximally 10 h in TYG in the present study vs. 72 h growth in brain infusion medium in ref. ³⁷), or false-positives. Inspection of the Ribo-seq-exclusive sORFs revealed 18 of the 21 remaining candidates to possess few to no aligned RNA-seq reads in our experimental conditions, explaining their neglect in the present screen. The other three sORF candidates (NC_004663.1_3066620_3066739_–1, NC_004663.1_951512_951613_–1 and NC_004663.1_4594728_4594793_1 in ref. ³⁷) were relatively highly expressed in our RNA-seq data, but located within annotated CDSs for larger proteins and, consequently, were not called as sORFs by ANNOgesic.

***B. thetaiotaomicron* promoter architectures.** A search for promoters upstream (–50 to +1) of the cognate TSS using the MEME³⁸ and GLAM2³⁹ toolkits identified two conserved motifs (Fig. 1d). Motif 1, consisting of two sequence elements centered at the –7 and –33 nt positions separated by an AT-rich spacer of

variable length, strongly resembles the canonical σ^{ABfr} promoter of *B. fragilis*⁶. In *B. thetaiotaomicron*, this promoter was found upstream of 81% of pTSSs as well as 60% of sTSSs and oTSSs (Fig. 1d).

The second motif comprised sequence elements centered at the –10 and –30 positions relative to the TSS and was generally associated with more lowly expressed genes (Supplementary Fig. 1b). As is exemplified for *BT_4614* (Supplementary Fig. 1c), we noticed certain cases where transcription of genes driven by promoter motif 2 (PM2) initiated from a different TSS in stationary phase compared to earlier growth stages. More generally, functional annotation of all 75 PM2-containing coding genes in our study revealed an enrichment of the oxidative stress response (Supplementary Fig. 1d) and differential expression analysis an upregulation of oxidative stress-related genes in the stationary growth phase (Supplementary Fig. 1e).

Theta-Base: an interactive online browser to interrogate the *Bacteroides* transcriptome. *B. thetaiotaomicron* is emerging as a model anaerobic bacterium; however, the community currently lacks an online repository for transcriptomic features and gene expression profiles. Inspired by online community data visualization platforms⁴⁰ such as the SalCom⁴¹ and AcinetoCom²⁹ databases compiled by the Hinton group for the bacterial pathogens *Salmonella enterica* and *Acinetobacter baumannii*, we generated Theta-Base as an intuitive online tool to easily interrogate the here-presented transcriptomic data. Theta-Base enables search queries for any annotated *B. thetaiotaomicron* coding gene as well as the here-identified noncoding genes (see below) and visualizes their expression profiles over growth in TYG in a simple heatmap format (Fig. 2a) that can be exported to the interactive graphing environment Plotly⁴². Moreover, a linkout to JBrowse⁴³ allows the transcriptomic data to be viewed in the context of the *B. thetaiotaomicron* chromosome or plasmid (Fig. 2b), and additionally features the annotations of transcriptomic parts including TSSs and Rho-independent terminators. We hope this database will be routinely consulted by researchers of the *Bacteroides* community to retrieve information about transcript borders (relevant for cloning purposes or the design of PCR primers) or relative expression levels, and plan to further complement the current dataset with transcriptome profiles derived from *B. thetaiotaomicron* grown under various experimental conditions in the future. Theta-Base can freely be interrogated at www.helmholtz-iri.de/en/datasets/bacteroides.

The noncoding RNA landscape of *B. thetaiotaomicron*. Our dRNA-seq analysis uncovered a wealth of noncoding RNAs in *B. thetaiotaomicron* scattered throughout the genome (Fig. 3a). This included abundant housekeeping transcripts such as the RNA component of RNase P (M1 RNA) and transfer-messenger RNA (tmRNA). We also detected the *Bacteroides* 6S RNA and obtained evidence for the existence of product RNAs (pRNAs) (Supplementary Fig. 2a), which were previously predicted in the Bacteroidetes phylum⁴⁴, but remained to be validated. In fact, we detected two classes of pRNAs in *B. thetaiotaomicron* (denoted pRNA and pRNA* in Supplementary Fig. 2a), reminiscent of previous observations in the ϵ -proteobacterium *Helicobacter pylori*²⁰.

Noncoding RNA candidates were identified and classified on the basis of homology to described representatives of *cis*-regulatory elements into riboswitches and RNA thermometers (Rfam database⁴⁵) or based on their genomic location into *cis*-antisense, intergenic, 5'- or 3'-derived and intra-operonic sRNAs (Fig. 3b; Supplementary Fig. 2b–i; Supplementary Data 2). ANNOgesic rediscovered three putative thiamine pyrophosphate



Fig. 2 Theta-Base allows easy access to *B. thetaiotaomicron* transcriptomic features. **a** Heatmap representation of absolute (reads per kilobase of transcript per million mapped reads [rpkm]; left) or relative (rpkm divided by row average; right) transcript levels illustrates the anti-correlation in expression of GibS (a.k.a. BTnc035) and its repressed targets (*BT_0771*, *BT_3893*; as identified in Fig. 5) over growth in TYG. For visualization, query genes (coding and noncoding) may either be entered individually in the upper left box or selected from pre-defined gene sets. Biological replicates can be displayed individually or (as in the given example) the average expression over each three replicates may be shown. ‘EEP’ refers to early, ‘MEP’ to mid-exponential, and ‘stat’ to stationary phase. The heatmap can be exported as a graphic file in the upper right corner. At the lower left corner is a link to the genome browser (**b**). **b** JBrowse view on the read coverages over the GibS-encoding locus within the *B. thetaiotaomicron* genome. Tracks for display can be selected on the left (i.e., annotated coding and noncoding genes and transcript features [transcription start sites, terminators] as well as different experimental conditions and replicates). At the top, gene names or genome coordinates may be entered; zoom-in and -out functions exist. Inset: By clicking on any of the annotated features, primary sequences can be retrieved in FASTA format.

(TPP) riboswitches^{46,47} in the 5' UTRs of *BT_0653* (encoding the thiamine biosynthesis protein ThiS), *BT_2390* (for an outer membrane thiamine transporter), and *BT_2396* (for a protein involved in nicotinamide mononucleotide transport). Further, *B. thetaiotaomicron* expressed 78 *cis*-antisense and 124 intergenic sRNAs, respectively, during growth in rich medium; the latter number was reduced to 49 when requiring the presence of a Rho-independent terminator structure at the 3' end as an additional selection criterion (along with a cognate TSS and a minimum read coverage; see Methods for details). We refer to this subset of 49 candidates as “high-confidence” intergenic sRNAs. Conservation analysis revealed 22 of these sRNAs to be conserved within two or more genera within the Bacteroidetes phylum, while the remaining 27 are strain-specific (Supplementary Fig. 3a). Analysis

of the nucleobase composition of *cis*-antisense and intergenic sRNAs revealed GC contents of 38% or 35%, respectively, which is lower than that of CDSs, transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs), in line with observations previously made in *Escherichia coli*⁴⁸ (Fig. 3c). Selected *cis*-antisense RNA representatives were predicted to fold into extensive secondary structures, whereas several intergenic sRNAs exhibited extended unfolded regions (Supplementary Fig. 3b). Promoter analysis revealed the canonical σ^{AbfR} promoter to be associated with 52% or 76%, and the non-canonical PM2 with 24% or 13% of *cis*-antisense or intergenic sRNA candidates, respectively (Supplementary Data 2).

CRISPR (clustered regularly interspaced short palindromic repeats) systems protect their host against bacteriophage infections and are encoded by ~45% of sequenced bacterial genomes,

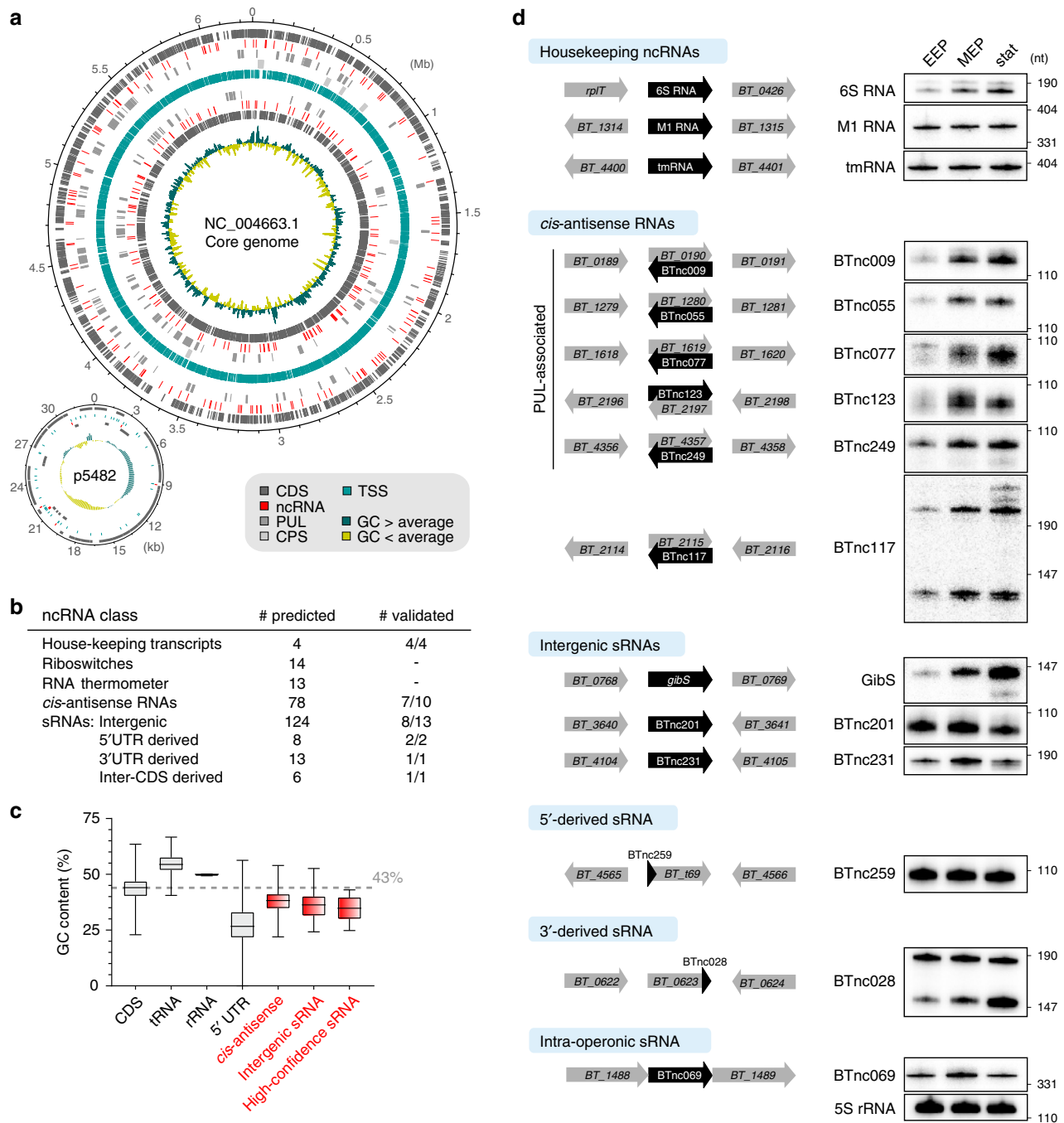


Fig. 3 The noncoding transcriptome of *B. thetaioaomicron*. **a** DNAPlotter maps¹¹¹ denote the genomic position of the identified noncoding genes within the chromosome or plasmid, respectively. Position of coding sequences (CDS), polysaccharide utilization loci (PUL), and capsular polysaccharide synthesis loci (CPS) were retrieved from NCBI and published literature⁵. Positions of transcription start sites (TSS) and noncoding RNAs (ncRNA) were identified in the present study. **b** Overview table on the numbers of predicted and experimentally validated (probed/detected on northern blot) ncRNA candidates of the different classes. **c** GC content of different transcript classes. The center lines refer to the medians, the lower and upper borders of the boxes refer to the 25% and 75%, respectively, and whiskers indicate the minimal/maximal GC contents of CDSs ($n = 4816$), tRNAs ($n = 71$), rRNAs ($n = 15$), 5' UTRs ($n = 1373$), *cis*-antisense RNAs ($n = 78$), and all putative or the high-confidence intergenic sRNAs ($n = 124$; 49). The dashed line represents the average GC content of the genome. **d** Northern blot validation of predicted ncRNA candidates. Left: sRNA locus orientation. Right: Total RNA was extracted from wild-type *B. thetaioaomicron* (AWS-001) grown in TYG to early exponential (EEP), mid-exponential (MEP) or stationary phase (stat) and used for Northern blotting (representative images from each two biological replicates are shown). Apparent sizes in nucleotides are given to the right of the blot. 5S rRNA served as loading control. Source data to this figure are provided in the Source Data file.

including that of *B. fragilis*⁴⁹. In contrast, no CRISPR array nor any Cas (CRISPR-associated) proteins have been annotated in *B. thetaioaomicron*. In agreement with this, the CRISPR Recognition Tool⁵⁰, which is included in the ANNOgesic pipeline, failed to identify any CRISPR-RNAs in our dataset.

Experimental validation of *B. thetaioaomicron* sRNAs. To validate the predicted sRNAs, we performed Northern blot assays for 31 randomly selected RNA candidates from diverse classes (Fig. 3d). DonS, the prototypical PUL-associated antisense RNA¹⁸, was not expressed when *B. thetaioaomicron* grew in TYG

medium and consequently, not identified in our screen. However, our transcriptomic approach readily detected ten further *cis*-antisense RNAs with a DonS-like genomic orientation, and we validated the expression of five of them (BTnc009, -055, -077, -249, and -123) by Northern blot (Fig. 3d). Of note, we observed anti-correlation in the expression of some of these antisense RNAs and their cognate *susC* homologs in our RNA-seq data (highlighted in red in Supplementary Fig. 4a), arguing that *B. thetaiotaomicron cis*-antisense RNAs may repress their corresponding PUL system in the presence of a prioritized carbon source, similarly to DonS in *B. fragilis*. Another seven pairs of *cis*-encoded antisense RNAs had a genomic architecture reminiscent of type-I TA systems, with each one of the divergently encoded genes harboring a candidate sORF. These putative TA systems showed a similar expression pattern with the presumed toxin mRNAs expressed at constant—albeit low—levels and the putative antitoxin RNA specifically induced in stationary phase (Supplementary Fig. 4b).

To assess the reliability of intergenic sRNA predictions, we independently tested the existence of selected candidates by Northern blot. This way, we validated nine of eleven tested candidates, from which eight (including BTnc035 [renamed to GibS; see below], BTnc201, and BTnc231; Fig. 3d; Supplementary Fig. 2d–f) possess both a TSS and an intrinsic terminator—features characteristic of canonical intergenic sRNAs. Thus—while not a hard selection criterion in the default ANNOgesic pipeline—the presence of a 3' terminator may enhance our confidence in sRNA predictions and we therefore prioritized these high-confidence intergenic sRNAs in the following.

We also uncovered several putative sRNAs originating from the 5' or 3' UTR of mRNAs of which we validated one representative candidate each (BTnc259, BTnc028; Supplementary Fig. 2g, h) by Northern blot (Fig. 3d). Finally, ANNOgesic classified sRNAs into a fifth group termed inter-CDS-derived (more appropriately, intra-operonic) sRNAs. These are similar to 5' UTR-derived sRNAs, however, they may originate at either a TSS or a processing site and can end at a Rho-independent terminator or a processing site. We identified six members belonging to this class, and experimentally validated BTnc069 (Fig. 3d; Supplementary Fig. 2i).

GibS—an intergenic sRNA conserved within *Bacteroides* spp.

To successfully colonize and persist in the human gut, *B. thetaiotaomicron* relies in large part on its ability to rapidly adapt to the ever-changing conditions associated with this dynamic environment. Given that sRNAs in other organisms are known to regulate gene expression in response to specific environmental cues^{15,51}, we wondered whether *B. thetaiotaomicron* would also harness its sRNA repertoire to adapt to environmental changes. Of the 49 high-confidence intergenic sRNAs identified, we focused on GibS (GlcNAc-induced *Bacteroides* sRNA; renamed from BTnc035 for reasons below)—a sRNA encoded in between a putative para-aminobenzoate synthase cluster (*BT_0763–68*) and a glycogen biosynthesis operon (*BT_0769–71*) (Fig. 4a). GibS was highly expressed during growth in TYG media, especially upon entry into the stationary growth phase (Fig. 3d). Moreover, its primary sequence and the non-canonical PM2 (–10 and –30 box; Fig. 1d) are well conserved within the *Bacteroides* genus (Fig. 4b, upper; Supplementary Fig. 3a); features indicative of a functional transcript.

Northern blotting indicated a ~145 nt-long RNA species to be the major GibS transcript form (Fig. 3d), in line with a potential terminator structure after which the read coverage in the RNA-seq experiment drops (Fig. 4b, lower). Primer-extension analysis confirmed the TSS of GibS as mapped by dRNA-seq

(Supplementary Fig. 5). In silico RNA alignment and folding of GibS homologs using WAR (webserver for aligning structural RNAs⁵²), which provides a consensus alignment and structure based on a range of methods that consider both minimal free energy and residue co-variation, indicated GibS to adopt a relatively unfolded conformation (Fig. 4c). Indeed, chemical and enzymatic in vitro probing largely confirmed this prediction, with the highly conserved 5' sequence being mainly single-stranded (positions 1–38), the middle region forming two consecutive meta-stable hairpins, followed by another linear region and the Rho-independent terminator (Fig. 4d). Apart from the extended single-stranded 5' region, this is reminiscent of the structure of well-characterized proteobacterial *trans*-acting sRNAs⁵³.

Within its gut niche, *B. thetaiotaomicron* uses simple sugars as signals to control the expression of metabolic modules^{12,54}. Given their ability to mediate rapid responses to metabolic stimuli⁵⁵, sRNAs might be involved in the adaptation of *Bacteroides* gene expression to nutrient availability. To test if GibS exhibits a monosaccharide-specific expression pattern, we profiled its steady-state levels in *B. thetaiotaomicron* grown to stationary phase in minimal medium with defined simple sugars as the sole carbon source (Fig. 4e). While GibS expression showed considerable variation across the panel of carbon sources, growth in the presence of N-acetyl-D-glucosamine (GlcNAc) resulted in the strongest induction (~1.6 fold higher GibS levels in stationary phase in GlcNAc than in TYG). Thus, GibS is a conserved, largely unstructured sRNA in *B. thetaiotaomicron* that is induced in the presence of select monosaccharides.

Identification of GibS targets. Toward characterizing the role of GibS in *B. thetaiotaomicron* physiology, we constructed a sRNA deletion mutant (Δ *gibS*) by removing the *gibS* sequence from the chromosome. In addition, a *trans*-complementation strain (*gibS*+) was generated for which the *gibS* gene under control of a modified P1T_{DP} promoter⁵⁶ (designated P1T_D) was re-inserted at an unrelated position into the Δ *gibS* chromosome. The promoter modification involved removal of the proximal *tetO2* operator to ensure transcription of the sRNA from its native TSS, albeit with the cost of leaky expression (see Methods). However, upon addition of anhydrotetracycline (aTC; 200 ng mL⁻¹), the resulting strain expressed GibS to wild-type levels (Supplementary Fig. 6). Interfering with GibS expression in bacteria grown in defined minimal medium in presence of GlcNAc as the sole carbon source resulted in subtle growth variations (Supplementary Fig. 7a). In contrast, strains grew nearly identically in the presence of glucose (Supplementary Fig. 7a).

Next, we employed a genome-wide comparative transcriptomic approach to search for GibS-dependent expression changes. To this end, wild-type *B. thetaiotaomicron*, Δ *gibS*, and *gibS*+ strains were grown to stationary phase in TYG—i.e., a growth phase when endogenous GibS is highly expressed (Fig. 3d)—and their total RNA was extracted, depleted for rRNA, and sequenced. In the absence of GibS, four genes were repressed and five genes induced as compared to their wild-type expression levels (Fig. 5a). Notably, with the exceptions of *BT_0294* and *BT_0823*, expression of all these genes reverted to wild-type levels in the *trans*-complementation strain (Fig. 5a), indicating a sRNA-specific effect. *BT_1871–72*, i.e. the two genes seemingly activated by GibS (and thus, repressed in strain Δ *gibS*), comprise a dicistron encoding an α -galactosidase and a β -glucosidase, respectively, and are predicted to be part of the putative PUL22 involved in arabinan degradation (Supplementary Fig. 7b)⁵⁷. The genes seemingly repressed by GibS (de-repressed in Δ *gibS*) included *BT_1655* (encoding a hypothetical protein within the CPS5 locus; Supplementary Fig. 7c) and *BT_3893* (the second gene within a

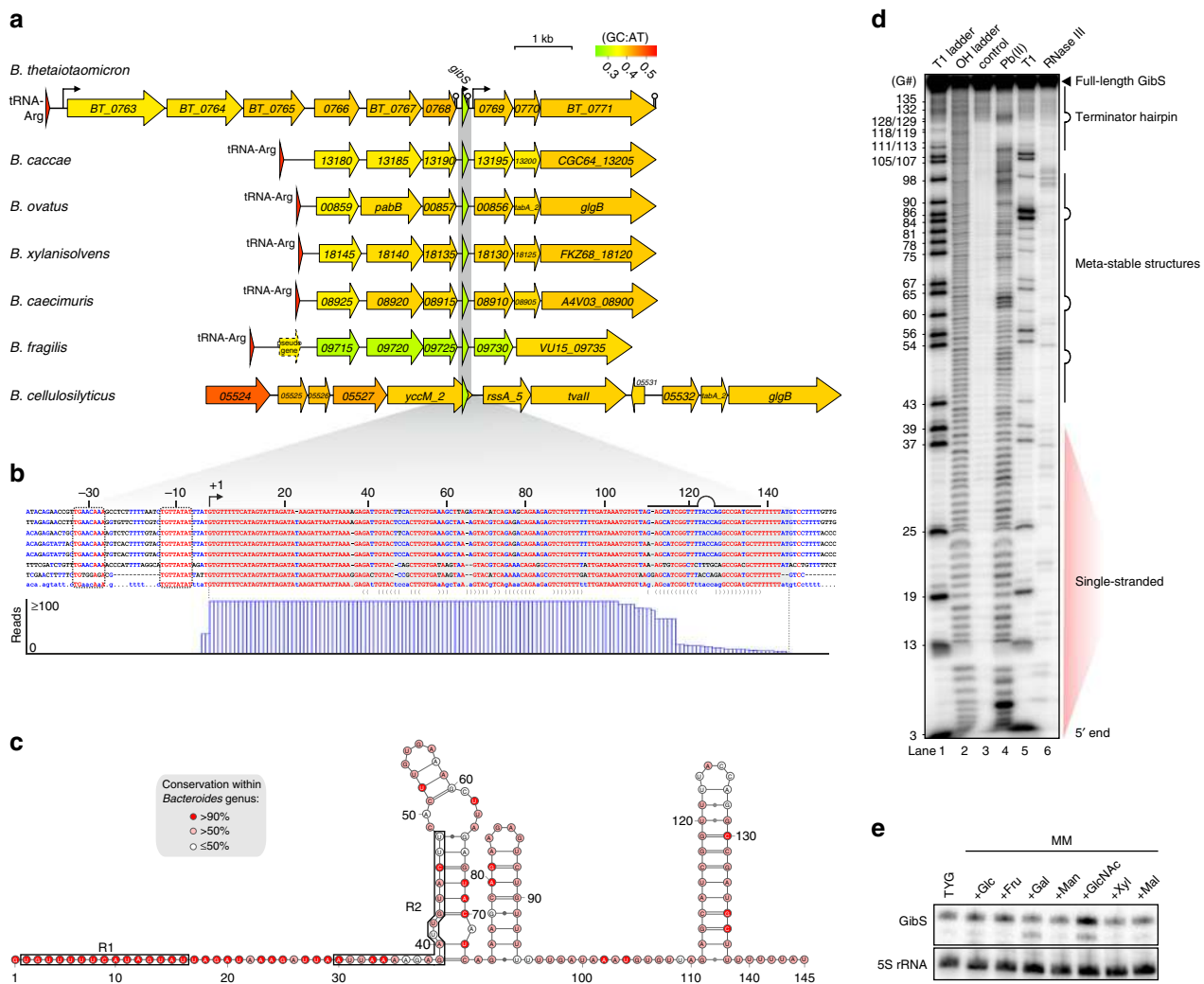
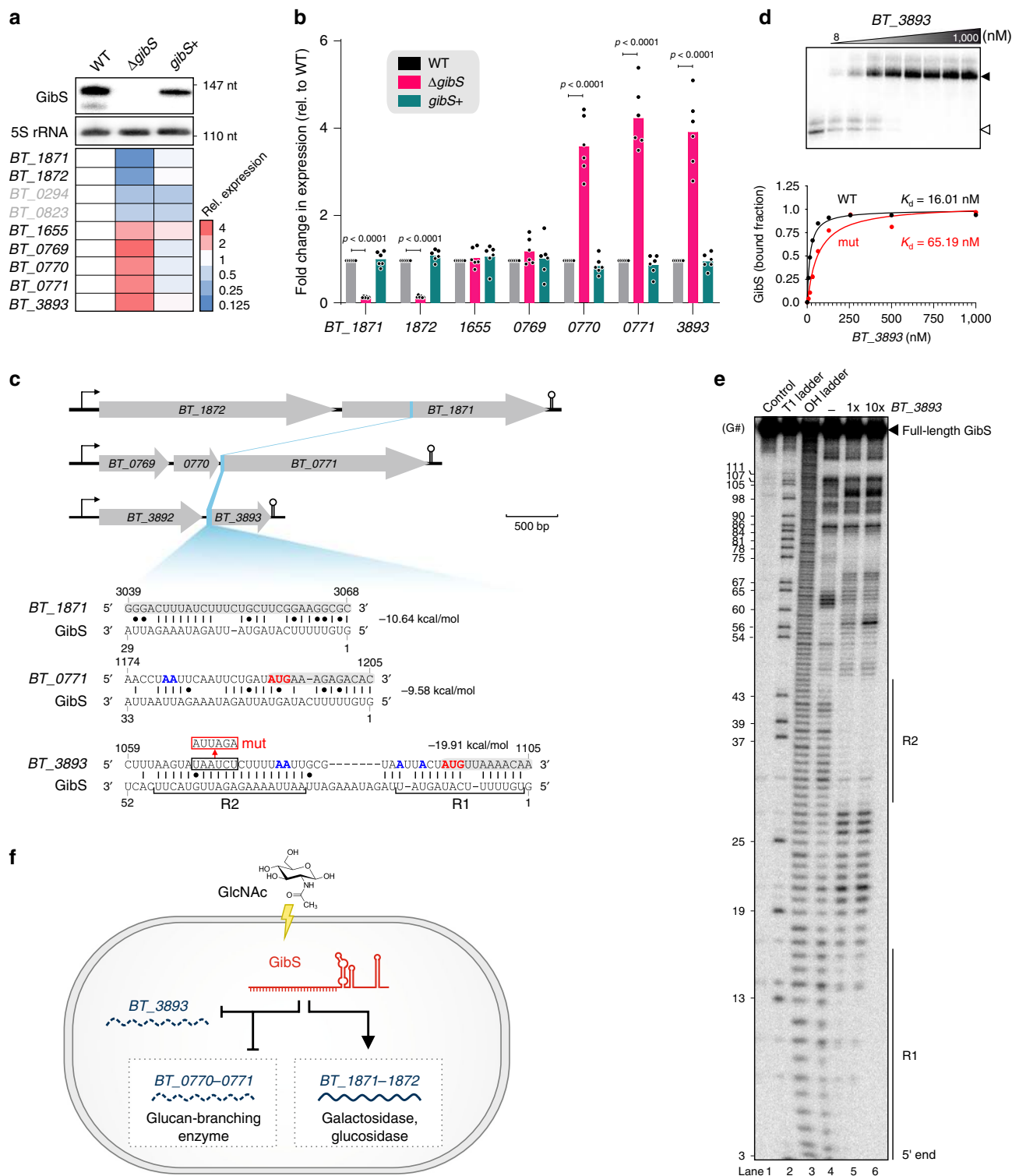


Fig. 4 Conservation, secondary structure, and expression profile of GibS sRNA. a Shown is the overview of the *gibS* locus in seven indicated *Bacteroides* species. **b** Upper: Sequence alignment of the *gibS* gene and flanking regions in the same *Bacteroides* species as in **a**. The sequence at the bottom refers to the consensus and the brackets below indicate the determined intra-molecular base-pairings (see **c** and **d**). Red and blue colors indicate perfectly conserved and less-conserved ribonucleobases, respectively. The numbers denote the position relative to the 5' end of GibS (+1 position; validated in Supplementary Fig. 5). Lower: Read coverage plot over the *gibS* locus (stationary phase, '-TEX' sample). At the position of the presumed intrinsic terminator (indicated above the alignment) read coverage drops, likely due to strong hairpins being less efficiently converted into cDNA. **c** Secondary structure prediction of *B. thetaiotaomicron* GibS was performed using the webserver for aligning structural RNAs⁵² and reflects the consensus structure of 14 different prediction programs, considering both, minimal free energy and covariation. The colors of the ribobases reflect their conservation within *Bacteroides* and numbering is as in **b**. R1 and R2 seed regions as identified in Fig. 5e are labeled. **d** In vitro structure probing of 5' end-labeled GibS confirms the in silico prediction (representative image from two independent replicate experiments is depicted). T1 and OH ladders refer to partial digestion under denaturing conditions with nuclease T1 (lane 1; cleaves unpaired G residues, indicated to the left), or alkali (lane 2; cleaves at all positions), respectively, and 'control' (lane 3) refers to untreated GibS. Lanes 4–6 reveal cleavages induced by lead (II) acetate (cleaves single-stranded nucleotides), RNase T1, or RNase III (cleaves double-stranded regions), respectively, under native conditions. **e** Northern blot-based expression profiling of GibS in wild-type *B. thetaiotaomicron* (AWS-001) grown to stationary phase in either TYG medium or in minimal medium (MM) in presence of glucose (Glc), fructose (Fru), galactose (Gal), mannose (Man), *N*-acetylglucosamine (GlcNAc), xylose (Xyl) or maltose (Mal), respectively, as the sole carbon source (representative image from two biological replicates). 5S rRNA was the loading control. Source data are provided in the Source Data file.

dicistron and encoding a hypothetical protein; Supplementary Fig. 7e). *BT_3892*—the first gene in this operon and encoding a branched-chain amino acid aminotransferase—was unaffected by *gibS* status. Finally, the glycogen synthesis *BT_0769–71* operon, which is encoded adjacent to *gibS* itself (Fig. 4a), was de-repressed in the Δ *gibS* mutant (Supplementary Fig. 7d). We rule out polar effects of sRNA deletion, as *BT_0769–71* expression reverted to wild-type levels when GibS was re-introduced in *trans* in the Δ *gibS* background (strain *gibS* + in Fig. 5a). Altered expression levels of all target candidates except *BT_1655* and *BT_0769* were independently validated by quantitative real-time PCR (qRT-PCR) (Fig. 5b).

To test whether the observed GibS-dependent expression changes result from direct sRNA:mRNA base-pairing events, we searched for regions within the target candidates that showed partial complementarity to the GibS sequence. Among the differentially expressed genes identified above, the IntaRNA program⁵⁸ predicted GibS to anneal within the CDS of *BT_1871* and the region spanning the respective start codons of *BT_0771* and *BT_3893* (Fig. 5c). The correspondingly predicted seed region of GibS comprised its highly conserved and unstructured 5' portion (Fig. 4c). To experimentally test these predictions, we performed electrophoretic mobility shift assays (EMSA) using



radiolabeled GibS and increasing concentrations of ~150 nt-long mRNA segments encompassing the assumed targeting sites. Binding of GibS to the 5' region of both *BT_3893* and *BT_0771* occurred with high affinity ($K_d = 16.01$ and 18.06 nM) resulting in an upshift of the sRNA with increasing concentrations of each mRNA target segment (Fig. 5d; Supplementary Fig. 8a). This suggests *BT_3893* and *BT_0771* mRNAs to be direct targets of GibS. Indeed, in-line probing of radiolabeled GibS with or without in vitro-transcribed *BT_3893* mRNA (Fig. 5e) validated the computationally predicted base-pairing events involving two

adjacent regions in the 5' portion of GibS (termed R1 and R2 in Fig. 5c), and mutating a hexameric sequence within the *BT_3893* region targeted by R2 reduced its affinity to GibS (Fig. 5d; Supplementary Fig. 8b). In contrast, the in vitro interaction of GibS with *BT_1871*, for which binding was proposed to occur within the CDS, was of very low affinity (Supplementary Fig. 8c). This argues that in the bacterial cell, base-pairing between GibS and *BT_1871* may depend on auxiliary factors, such as a RNA chaperone, or that activation of this gene by GibS (Fig. 5a, b) is indirect.

Fig. 5 GibS regulates the expression of target genes likely via direct base-pairing. **a** Comparative transcriptomics of isogenic wild-type *B. thetaiotaomicron* (WT; AWS-003), a *gibS* deletion mutant (Δ *gibS*; AWS-028), and complementation strain (*gibS*+; AWS-035). Upper: Northern blot showing stationary phase expression of GibS in the three strains. Heatmap below: RNA-seq data for putative targets under the same condition (i.e., fold change >2 or <0.5 at an FDR <0.05 between WT and Δ *gibS*; mean of two biological replicates). Gray labeled genes were not considered further, as their expression pattern was not predictive of a GibS-specific effect. **b** qRT-PCR-based validation of the identified target candidates. 16S rRNA was used as a reference transcript. Bars denote the mean from three biological replicates each in technical duplicates (represented as single dots). Significant differences compared to wild-type expression levels were assessed using two-way ANOVA (Sidak's multiple comparisons test; *p* values for significant [*p* < 0.05] comparisons are given). **c** Locus representation of target genes (upper) and their predicted interactions with GibS (lower). Red/blue nucleotides: start codon/Kozak-like sequences; gray shading: coding region; R1, R2: seed regions interacting with *BT_3893* (see **e**); 'mut': mutated nucleotides of *BT_3893* (for **d**). Minimal free energy values are to the right. Coordinates are relative to the TSS. **d** EMSAs support the predicted sRNA targeting site within *BT_3893*. T7-transcribed and 5' end-labeled GibS was incubated with increasing concentrations of a ~150 nt-long 5' segment of either wild-type (WT; black) or mutated variant of *BT_3893* ('mut'; red; see **c**). K_d values represent the means of two independent replicate experiments. A representative gel from the WT is shown at the top and an image for binding to the mutated variant is given in Supplementary Fig. 8b (white and black arrowheads refer to free and bound GibS, respectively). **e** In-line probing of ~0.2 pmol ³²P-labeled GibS in the absence (lane 4) or presence of either 20 nM (lane 5) or 200 nM *BT_3893* mRNA leader (lane 6). Untreated RNA (lane 1; 'control'), partially RNase T1- (lane 2; 'T1') or alkali-digested (lane 3; 'OH') GibS served as ladders. R1 and R2 denote GibS seed regions protected from cleavage in the presence of the target. Two independent replicate experiments were performed of which a representative image is depicted. **f** Working model for the role of GibS in *B. thetaiotaomicron*. Source data are provided as a Source Data file.

Discussion

Only recently and thanks to technological breakthroughs that allow anaerobic culturing without the need for extensive lab equipment⁵⁹ and the development of versatile protocols for genetic manipulation^{56,60–62}, has it become possible for a wider group of researchers to functionally characterize anaerobic gut commensals. In this context, *Bacteroides* species—predominant members of the human microbiota—are gaining increasing attention by the scientific community⁶³.

In the present work, we have compiled a high-resolution transcriptome map of the *B. thetaiotaomicron* type strain VPI-5482, which is freely accessible to the research community as an intuitive online database. Our screen identified ~4500 TSSs within the core genome and plasmid, and revealed high plasticity in the transcriptome structure of *B. thetaiotaomicron*. Inspection of the sequences upstream of the identified TSSs revealed the canonical σ^{ABfr} promoter. Not only did our genome-wide motif search confirm the -7 box, but it also identified the -33 box that was initially inferred from manual inspection of a few dozen promoters⁶, but could not previously be verified globally within the Bacteroidetes phylum⁸. We assume that the variable spacing between the two boxes (Fig. 1d) might have hampered previous motif searches. In addition, our global promoter analysis revealed a second sequence motif enriched in front of oxygen tension-related genes, suggesting the motif could be recognized by a stationary phase-inducible alternative sigma factor to pre-adapt microbes against exposure to reactive oxygen species⁶⁴. However, this motif differs substantially from the recognition sequence of the sigma factor EcfO, which protects *B. fragilis* against oxidative stress⁹ and deserves further investigation.

Comprehensive transcriptomic analyses are reliant on an accurate genome annotation. We intend to regularly update Theta-Base with the most recently reported genomic features. In addition, in the future, plasticity of the *B. thetaiotaomicron* transcriptome may be refined even further by similar RNA-seq studies under additional experimental settings, including growth on defined carbon sources or in face of specific stress conditions. For example, given that *Bacteroides* promoter inversions are commonly observed under stressful conditions such as antibiotics exposure or gut adaptation⁶⁵, those data would be useful to explore the effects of invertible *Bacteroides* promoters on global gene expression.

In addition to bacterial transcriptional control networks, post-transcriptional regulation mediated by regulatory RNA molecules has been revealed as a second layer of adapting global gene expression to changing environmental and intrinsic cues.

Particularly, *trans*-encoded sRNAs have been described in species across the bacterial phylogenetic tree^{15,53}, where they regulate target mRNAs through imperfect base-pairing interactions mediated by short seed sequences and, in many Gram-negative species, dedicated RNA chaperones⁶⁶. In the majority of the described cases, sRNA annealing to the SD sequence and/or start codon, interferes with translation initiation and induces target decay¹⁵. Conversely, certain sRNAs may promote translation of their target by unfolding inhibitory structures that otherwise block ribosome binding^{67,68}. In addition, activating *trans*-encoded sRNAs in γ -proteobacteria may induce target gene expression by binding to nascent mRNA leaders and interfering with Rho-mediated premature transcription termination^{69,70}. Besides targeting mRNAs, some sRNAs may bind to, and titrate, regulatory RBPs, thereby indirectly controlling the levels of the target mRNA set of the bound protein⁷¹. Conceptually related, an emerging class of sRNA sponges bind and titrate other sRNAs, creating complex post-transcriptional gene expression control networks⁷². In contrast, if and how *trans*-encoded sRNAs regulate target gene expression in *Bacteroides* has previously barely been addressed.

Overall, our study provides evidence for the existence of >200 noncoding RNAs from diverse classes in *B. thetaiotaomicron*. As an example, our screen experimentally validated the 6S RNA homologue of *Bacteroides* and provided evidence for the existence of pRNAs that rescue sequestered RNA polymerases, further highlighting 6S RNA-mediated regulation of transcriptional activity to be an ultra-conserved mechanism. Taking into account that dRNA-seq was performed in just three defined growth stages in a single (rich) medium, the reported numbers for *cis*-antisense (78) and intergenic sRNA candidates (124, of which we consider 49 as high-confidence) are probably conservative. In addition, the screen identified 21 UTR-derived sRNA candidates, suggesting that evolution of dual-function transcripts⁷³ is not restricted to Proteobacteria. The full repertoire of *B. thetaiotaomicron* noncoding RNAs therefore appears consistent with the numbers reported in other bacteria^{20,29,74–77}. However, we want to point out that antisense transcription might at least partially be due to the appearance of random promoters²⁶, which is particularly likely in a genome with a low GC content. Consequently, not all identified *cis*-antisense RNA candidates might actually be functional.

Here, we began to mechanistically characterize the GibS sRNA. We selected GibS for several reasons: (i) its strong sequence conservation (especially in its 5' portion) within the *Bacteroides* genus argues for functionality; (ii) the presence of PM2 upstream of its TSS renders it a representative for a larger set of transcripts;

(iii) the generally high steady-state levels of this sRNA—particularly in stationary growth phase and in the presence of the simple sugar GlcNAc—enables robust detection and analysis; and (iv) *gibS* does not overlap with any other gene, which allowed for the straight-forward construction of a clean deletion mutant. In the course of our studies, we validated the TSS of *GibS* predicted by dRNA-seq and determined the secondary structure of the sRNA. Unlike many well-characterized *trans*-encoded sRNAs in Proteobacteria, *GibS* contains long single-stranded regions, particularly in its 5' part. Expression profiling of wild-type *B. thetaiotaomicron* and a Δ *gibS* mutant identified potential target mRNAs of this sRNA and provided a possible explanation for the lack of structure at the 5' end: *GibS* harbors an extended seed region, comprising the first 30–50 nt of its 5' portion, that mediates base-pairing with the region around the target mRNA's start codon (*BT_0771*, *BT_3893*) or, potentially, a region deep within the CDS of *BT_1871*.

What is the physiological role of this sRNA? We find *GibS* expression to peak when *B. thetaiotaomicron* is cultured in the presence of GlcNAc as the sole carbon source. This monosaccharide is a constituent of host-derived glycosaminoglycans such as chondroitin sulfate (CS), dermatan sulfate, and heparin/heparan sulfate (HS), with CS and HS being the priority nutrients for *B. thetaiotaomicron*^{78,79}. The identified *GibS* targets are related to metabolic processes: The *BT_1871*–*BT_1872* operon, which is activated by *GibS*, encodes a galactosidase and a periplasmic glucosidase, and *BT_0770*–*BT_0771*, that is repressed by *GibS*, encodes a glucan-branching enzyme. Together, this suggests a regulatory network, wherein sensing of a specific glycosaminoglycan-derived monosaccharide results in the induction of *GibS*, which in turn leads to post-transcriptional metabolic rearrangements in the bacterial cell (Fig. 5f). If true, this places *GibS* among the growing number of sRNAs involved in carbon catabolite repression in diverse bacterial organisms⁸⁰. Dissecting the exact role of *GibS*, however, requires further work.

GibS targeting the translation initiation region of *BT_0771* and *BT_3893* mRNAs is reminiscent of the classical mechanism of sRNA-mediated target control, preventing ribosome loading and interfering with translation initiation, often accompanied by enhanced target mRNA decay¹⁵. We here provided the first example showing that, despite the lack of a SD sequence, *Bacteroidetes* mRNAs may still be repressed by sRNAs that anneal to their translation initiation region. Consequently, deleting *gibS* from the *B. thetaiotaomicron* genome resulted in an upregulation of both these mRNAs, and *trans*-complementation reverted their expression to wild-type levels. In contrast, sRNA targeting within the CDS is relatively rare, but could mask an endonucleolytic cleavage site, thereby stabilizing the target mRNA⁶⁸. In line with such a mechanism, *BT_1871* levels dropped in the absence of *GibS* as compared to both the wild-type and complementation strain. Alternatively, activation of *BT_1871* mRNA in presence of *GibS* may occur indirectly.

Commonly, protein co-factors are involved in sRNA-mediated target control⁸¹. In *Bacteroides*, in absence of an obvious homologue of both Hfq and ProQ, sRNAs might either regulate targets in a protein-independent manner, or depend on an alternative, elusive RNA chaperone. The average GC content of the identified intergenic sRNAs was relatively low (~35%). Moreover, we identified several *B. thetaiotaomicron* sRNAs that appear rather unstructured (Supplementary Fig. 3b). Well-studied Hfq-dependent sRNAs in Proteobacteria typically contain seeds of ~6–10 nt in length⁸². The above features—AT richness and single-strandedness—might indicate that certain *B. thetaiotaomicron* sRNAs need more extended seeds for efficient target annealing, as is exemplified by *GibS*. Whether such extensive base-pairing events would be more or less likely to require assistance by RNA chaperones needs further

investigation. While in the present case, *GibS* interactions with the 5' region of *BT_0771* and *BT_3893* occurred at very high affinities even without a protein co-factor in vitro, long single-stranded seed regions may be inherently vulnerable to ribonucleases (homologs of RNase BN, –G, –HII, –III, –P, –R, and –Z are present in *B. thetaiotaomicron* according to Pfam⁸³) and stability of the respective sRNAs in the bacterial cytosol could thus depend on RBP association.

Compared to species that have long served us as model organisms for prokaryotic RNA biology, there is currently no toolbox available to mechanistically decipher the functions of sRNAs in *Bacteroides* species. Here, we employed an aTC-triggered sRNA induction system, but leaky expression and limited dynamic range (Supplementary Fig. 6) prompted us to grow the respective strain (*gibS*+) for 2 h in presence of the inducer to reach high sRNA levels. Despite successfully complementing expression changes in the Δ *gibS* mutant, more tightly inducible sRNA pulse-expression systems might be harnessed in the future to facilitate discrimination between direct and indirect effects of sRNA overexpression^{84,85}. For target validation, two-plasmid systems, consisting of an inducible sRNA expression vector and a constitutively expressed fluorescent reporter fusion with the respective target mRNA, are routinely used in Proteobacteria^{86,87}. Despite the oxygen-dependent maturation of commonly used fluorescent proteins, similar reporters should also work in anaerobic bacteria because for sample preparation, cell suspensions are typically shifted to normoxic conditions that allow for fluorescence recovery⁸⁸. Similarly, luciferase assays—albeit depending on molecular oxygen—have already been successfully employed in *Bacteroides*^{47,89} and β -galactosidase assays with *lacZ*-target fusions⁹⁰ represent an oxygen-independent alternative.

Modern sequencing-based technologies have proven useful to gain global perspectives on bacterial sRNAs⁹¹ and their generic nature should allow sRNA systems biology approaches to be applied to currently understudied bacterial species. For example, pooled knockdown or knockout mutant libraries represent invaluable tools for global screens. Transposon-sequencing (Tn-seq) data exist for *B. thetaiotaomicron*^{92–94} and it will be intriguing to re-analyze them for phenotypes associated with a disruption of the here-identified noncoding RNA candidates. However, like any random mutagenesis approach, Tn-seq is inherently biased toward longer genes, typically resulting in an underrepresentation of sRNA mutants. In this context, targeted approaches such as CRISPR interference, whose applicability was already demonstrated for *B. thetaiotaomicron*⁸⁹, appear as promising alternatives to simultaneously knock down hundreds of sRNAs and screen the resulting mutant pool under a variety of conditions. Obviously, implementing these technologies requires time and effort. The here-presented data, however, chart a rich world of *Bacteroides* RNA biology for us to explore.

Methods

Bacterial strains and genetics. Strains, plasmids, and oligonucleotides used in this study are listed in Supplementary Data 4. *B. thetaiotaomicron* type strain VPI-5482 is referred to as wild-type throughout the study. The *gibS* deletion mutant (Δ *gibS*) was generated as previously described⁹⁵. Briefly, 1-kb sequences flanking the region to be deleted were amplified by PCR (using oligos AWO-053/–054, AWO-055/–56, AWO-314/–315, and AWO-316/–317) and assembled into the suicide vector pExchange-*tdk* by Gibson assembly (NEB) as per the manufacturer's protocol. A 2- μ L aliquot of this reaction was transformed into electro-competent *E. coli* S17-1 λ pir. Transformants were conjugated with a *tdk* deletion mutant of *B. thetaiotaomicron* (Δ *tdk*) and conjugants counter-selected on 5-fluoro-2'-deoxyuridine (FudR) plates. Single recombinants were selected on Brain Heart Infusion Supplemented (BHIS) agar containing 200 μ g mL⁻¹ gentamicin and 25 μ g mL⁻¹ erythromycin. Double recombinants, resulting in either scarless deletion mutants or wild-type revertants, were selected by growth on BHIS agar containing 200 μ g mL⁻¹ FudR and an inability to grow on BHIS agar containing 25 μ g mL⁻¹ erythromycin. Successful deletions were subsequently confirmed by PCR (AWO-150/–151, AWO-111/–112, AWO-318/–319, AWO-340/–341, AWO-342/–343) and Sanger sequencing.

A *gibS* complementation strain (*gibS*⁺) was constructed using a variant of the pNBU2 vector system⁹⁵ (Supplementary Data 4). The full *gibS* gene was integrated into the vector by Gibson assembly (AWO-156/–157) to ensure transcription from its native TSS. This resulted in a deletion of the proximal *tetO2* (T_D) operator downstream of the promoter P1T_D, while leaving the second operator (T_D) intact⁵⁶. This construct, expressing the *GibS* sRNA from its +1 nucleotide under control of the P1T_D promoter, was conjugated into the Δ *tdk* strain and selected on BHIS agar containing 25 μ g mL⁻¹ erythromycin. Successful insertion was confirmed by PCR (AWO-160/–161) and Sanger sequencing.

B. thetaiotaomicron culture conditions. *Bacteroides* strains were cultured in an anaerobic chamber (Coy Laboratory Products) in presence of an anoxic gas mix (85% N₂, 10% CO₂, 5% H₂) at 37 °C. Routine cultivation involved the use of complex media; TYG (20 g L⁻¹ tryptone, 10 g L⁻¹ yeast extract, 0.5% glucose, 5 mg L⁻¹ hemin, 1 g L⁻¹ cysteine, 0.0008% CaCl₂, 19.2 mg L⁻¹ MgSO₄·7H₂O, 40 mg L⁻¹ KH₂PO₄, 40 mg L⁻¹ K₂HPO₄, 80 mg L⁻¹ NaCl, 0.2% NaHCO₃) and BHIS (52 g L⁻¹ BHI agar powder, 1 g L⁻¹ cysteine, 5 mg L⁻¹ hemin, 0.2% NaHCO₃). Carbohydrate growth assays were performed in minimal medium (1 g L⁻¹ L-cysteine, 5 mg L⁻¹ hemin, 20 mg L⁻¹ L-methionine, 4.17 mg L⁻¹ FeSO₄, 0.2% NaHCO₃, 0.9 g L⁻¹ KH₂PO₄, 0.02 g L⁻¹ MgCl₂·6H₂O, 0.026 g L⁻¹ CaCl₂·2H₂O, 0.001 g L⁻¹ CoCl₂·6H₂O, 0.01 g L⁻¹ MnCl₂·4H₂O, 0.5 g L⁻¹ NH₄Cl, 0.25 g L⁻¹ Na₂SO₄) supplemented with 0.5% of the indicated carbon sources⁹⁶.

RNA extraction, TEX treatment, cDNA library preparation, and sequencing.

For dRNA-seq (Figs. 1–3), wild-type *B. thetaiotaomicron* VPI-5482 (AWS-001) was grown overnight in 5 mL pre-reduced TYG medium followed by sub-culturing (1:100) into 50 mL fresh pre-reduced TYG. Total RNA was isolated by hot phenol extraction from culture aliquots of each three biological replicates at early exponential (4 h), mid-exponential (7 h), and stationary growth phase (10 h). Briefly, 4 OD equivalents of culture were harvested and 1.6 mL stop mix⁹⁷ was added (95% vol vol⁻¹ ethanol, 5% vol vol⁻¹ water saturated phenol, pH >7.0). The bacterial cells were lysed by incubation with 600 μ L lysozyme (0.5 mg mL⁻¹) and 60 μ L 10% SDS for 2 min at 64 °C, followed by the addition of 66 μ L of 3 M NaOAc. Phenol extraction (750 μ L; Roti-Aqua phenol) was performed at 64 °C for 6 min with the subsequent addition of 750 μ L chloroform. RNA was precipitated from the aqueous phase with twice the volume of 30:1 (ethanol:3 M NaOAc, pH 6.5) mix and incubated at –80 °C overnight. After centrifugation, pellets were washed with 75% (vol vol⁻¹) ethanol and re-suspended in 50 μ L H₂O. Contaminating genomic DNA was removed by treating 40 μ g total RNA with 5 U of DNase I (Fermentas) and 0.5 μ L Superase-In RNase Inhibitor (Ambion) in a 50 μ L reaction. RNA quality was checked using a 2100 Bioanalyzer and the RNA 6000 Nano kit (Agilent Technologies). RNA integrity (RIN) values for all samples were between 9.2 and 9.6.

Prior to cDNA synthesis, total RNA was fragmented using ultrasound (4 pulses of 30 s at 4 °C) and treated with T4 polynucleotide kinase (NEB). Subsequently, half of each total RNA sample was treated with Terminator exonuclease (TEX) to enrich for primary transcripts, whereas the other half was left untreated. RNA samples were then poly(A)-tailed using poly(A) polymerase and 5'-PPP was removed with 5' polyphosphatase (Epicentre). RNA adaptors were ligated and first-strand cDNA synthesis was carried out using oligo(dT) primers and M-MLV reverse transcriptase. The cDNA was PCR amplified to about 10–20 ng μ L⁻¹, purified using Agencourt AMPure XP kit (Beckman Coulter Genomics), and fractionated in a size range of 200–500 bp. Libraries were sequenced on an Illumina NextSeq platform (150 cycles) at the Core Unit SysMed of the University of Würzburg.

For conventional RNA-seq (Fig. 5a), *B. thetaiotaomicron* Δ *tdk* (AWS-003), Δ *gibS*, (AWS-028), and *gibS*⁺ (AWS-035) strains (two biological replicates) were grown in TYG to stationary phase as described above, followed by the addition of aTC at a final concentration of 200 ng mL⁻¹ (for maximal *GibS* expression with minimal growth attenuation; Supplementary Fig. 6a, b), and resumed growth therein for 2 h. Total RNA was extracted and RNA quality checked as above with RIN values between 7.7 and 9.5. Prior to library preparation, rRNA was depleted using the Pan-Prokaryote riboPOOLS kit (siTOOLS Biotech). In brief, 1 μ g of total RNA was incubated for 10 min at 68 °C and 30 min at 37 °C with 100 pmol of rRNA-specific biotinylated DNA probes in 2.5 mM Tris-HCl pH 7.5, 0.25 mM EDTA, and 500 mM NaCl. DNA-rRNA hybrids were depleted from total RNA by two consecutive 15 min incubations with 0.45 mg streptavidin-coated magnetic Dynabeads MyOne C1 (ThermoFisher Scientific) in 2.5 mM Tris-HCl pH 7.5, 0.25 mM EDTA, and 1 M NaCl at 37 °C. The rRNA-depleted RNA samples were purified using the Zymo RNA Clean & Concentrator kit combined with DNase treatment on a solid support (Zymo Research).

cDNA libraries were prepared using the NEBNext Multiplex Small RNA Library Prep kit for Illumina (NEB) in accordance with the manufacturers' instructions, except for the following modifications: RNA samples were fragmented with Mg²⁺ for 2.75 min at 94 °C using the NEBNext Magnesium RNA Fragmentation Module (NEB) followed by RNA purification with the Zymo RNA Clean & Concentrator kit. Fragmented RNA was dephosphorylated at the 3' end, phosphorylated at the 5' end, and decapped using 10 U T4-PNK +/- 40 nmol ATP and 5 U RppH, respectively (NEB). After each enzymatic treatment, RNA was purified with the Zymo RNA Clean & Concentrator kit. RNA fragments were ligated for cDNA synthesis to the 3' SR and 5' SR adapters pre-diluted 1:3 with nuclease-free water. PCR

amplification to add Illumina adaptors and indices to the cDNA was performed for 14 cycles with 1:3 pre-diluted primers. Barcoded cDNA libraries were purified using magnetic MagSi-NGSPREP Plus beads (amsbio) at a 1.8:1 ratio of beads to sample volume. Libraries were quantified with the Qubit 3.0 Fluometer (ThermoFisher) and the library quality and size distribution was checked using a 2100 Bioanalyzer with the high sensitivity DNA kit (Agilent). Sequencing of ten pooled libraries spiked with 5% PhiX control library was performed in single-end mode on the NextSeq 500 platform (Illumina) with the Mid Output Kit v2.5 (75 cycles).

Read processing and mapping. Sequencing reads were quality filtered with the local run manager software from Illumina version 2.2.0. Generated reads were then trimmed for the NEBNext adapter sequence using Cutadapt version 2.5 with default parameters. In addition, Cutadapt was given the –nextseq-trim=20 switch to handle two-color sequencing chemistry and reads that were trimmed to length 0 were discarded.

For both sequencing protocols (dRNA-seq and conventional RNA-seq), reads were mapped to the *B. thetaiotaomicron* VPI-5482 reference genome (NC_004663.1) and plasmid (NC_004703.1) using the READemption pipeline⁹⁸. Details of alignment statistics can be found in Supplementary Data 5. Aligned reads in wiggle format were visualized using both the Integrated Genome Browser⁹⁹ and JBrowse⁴³. Differential gene expression analysis was performed using DESeq2¹⁰⁰ with log fold-change shrinkage using the DESeq2 betaprior method.

Identification of TSSs. We employed the ANNOgesic pipeline (version 0.7.33) that integrates a suite of tools to annotate bacterial genomes from both dRNA-seq and conventional RNA-seq data²². TSSs were identified using the ANNOgesic implementation of TSSpredator⁷⁷ (usage: annogesic tss_ps) which compares the relative enrichment of reads between TEX-treated samples and their untreated counterparts. This resulted in characteristic enrichment peaks that were indicative of a 5' triphosphate that protects against TEX digestion. Default settings were used with the addition of the gene validation option that relates identified TSSs to annotated genes (Supplementary Data 1). TSSs were categorized based on their enrichment and location relative to the start of the cognate coding gene. Primary TSSs were classified as having the highest coverage within 300 bp upstream of an ORF, while all other TSSs within this region were defined as secondary TSSs. Internal TSSs were identified as originating on the sense strand within a coding gene and antisense TSSs were located on the antisense strand overlapping with, or within a 100 bp flanking region, of a given sense gene. All remaining TSSs were classified as orphan TSSs. TSSs called by TSSpredator were manually curated based on read coverage plots to ensure accuracy of the assignments and the thus validated TSSs are incorporated in Theta-Base (www.helmholtz-hiri.de/en/datasets/bacteroides).

Search for promoter motifs. Conserved DNA sequences upstream of TSSs were identified by ANNOgesic (usage: annogesic promoter) integrating both, MEME for gapless motifs³⁸ and GLAM2 for gapped motifs³⁹. Default parameters were used with the addition of the flag –n set to 50 to increase sensitivity. Representative sequences containing consensus motifs were aligned using ClustalW (version 2.1)¹⁰¹ and logos generated with WebLogo (version 2.8)¹⁰².

Prediction of intrinsic terminators. ANNOgesic employs two heuristic algorithms for the prediction of Rho-independent terminators (usage: annogesic terminator). TransTermHP scans genome sequences for the presence of Rho-independent terminators³⁰. To further substantiate these predictions and to detect the presence of Rho-independent terminators also in between convergent gene pairs, RNA-seq data were referenced to detect a significant decrease in coverage associated with the predicted terminators. All parameters were run at default values. The combined results of these predictions are available in Supplementary Data 6 and incorporated into Theta-Base (www.helmholtz-hiri.de/en/datasets/bacteroides).

sRNA identification. To detect putative sRNAs in *B. thetaiotaomicron* from our dataset, ANNOgesic (usage: srna) tested transcripts for several criteria. First, detected transcripts were compared with RNAs contained within the sRNA database (BSRD; <http://www.bac-srna.org/BSRD/index.jsp>) and the non-redundant protein database (nr database; <ftp://ftp.ncbi.nih.gov/blast/db/FASTA/>). Homologs identified in the former were classified as putative sRNAs, while those contained in the latter were excluded from further analysis.

The remaining transcripts were further assessed and classified as intergenic sRNA candidates in case of the presence of a defined TSS, a stable secondary structure (folding energy change as calculated by RNAfold and normalized to sRNA length < –0.05), and a length ranging from 30 to 500 nt. *Cis*-antisense RNAs were identified in a similar manner except that in addition to the above criteria, they originated in antisense orientation to annotated genes. sRNAs sharing a TSS with a mRNA were classified as 5' UTR-derived if they were associated with a sharp drop in coverage and/or a processing site in front of the cognate CDS. 3' UTR-derived sRNAs required a TSS or processing site within the 3' region of the cognate coding gene and either a processing site or shared terminator with the parental mRNA. Intra-operonic sRNAs required a TSS or processing site at the 5' end as well as a coverage drop or processing site at their 3' end.

To maximize detection of sRNAs while minimizing the number of false positives, we used a list of known or proposed *cis*-antisense RNAs from a previous study¹⁸ as a benchmark and lowered the minimum average coverage to 5. With this modification, ANNOgesic recovered five out of seven benchmark RNAs (Supplementary Data 2). The high success rate of experimental validation for the identified sRNA candidates from all categories (Fig. 3d) further supports the predictions.

sRNA conservation analysis. We constructed a custom genome database consisting of completed bacterial genomes from the ENA (<https://www.ebi.ac.uk/genomes/bacteria.html>), accessed 1/12/2017) belonging to class Bacteroidia according to the NCBI Taxonomy (Taxonomy ID: 200643). We then performed an iterative search for each candidate sRNA sequence using nhmmer 3.1b1¹⁰³, similarly to as previously described¹⁰⁴. In each round of iteration nhmmer was run with the flags `--popen 0.4999 -E 0.001 --incE 0.001`, hit sequences with an *E* value of 0.001 or less were additionally required to have near-full length alignments not missing more than 10% of sequence length at either end, and the resulting alignment was used as input for hmmbuild and fed into the next round of iteration. The alignments were then manually examined using the RALEE alignment editor¹⁰⁵. The final homologous sequences identified for GibS were additionally subjected to realignment and structure prediction using the webserver for aligning structural RNAs (WAR) to improve prediction of secondary structure⁵².

Northern blotting and qRT-PCR analysis. The sequences of DNA oligonucleotides used for Northern blot and qRT-PCR are given in Supplementary Data 4. Northern blotting was performed as described previously¹⁰⁶. Briefly, total RNA (2.5–10 µg) was run on a 6% (vol vol⁻¹) polyacrylamide (PAA)-7 M urea gel and electro-blotted onto Hybond XL membranes (Amersham) at 50 V, 4 °C for 1 h. Blots were probed with ³²P-labeled gene-specific oligonucleotides in Hybri-Quick buffer (Carl Roth AG) at 42 °C and exposed as required. Visualization was achieved using a phosphorimager (FLA-3000 Series, Fuji) and images were quantified using ImageJ¹⁰⁷.

For qRT-PCR assays, reverse transcription and PCR amplification were performed in the same reaction mix containing 1 µL of DNase I-treated RNA (adjusted in water to 10 ng µL⁻¹), 5 µL master mix (No ROX SYBR MasterMix blue dTTP kit, Takyon), 0.1 µL of forward and reverse primer (10 µM each), and 0.08 µL reverse transcriptase (One-Step Kit converter, Takyon) per well in a 96-well format. Two technical replicates per each biological replicate were pipetted and plates analyzed on a CFX96 instrument (Biorad).

Primer-extension analysis. Primer-extensions were performed in 20 µL reactions containing 9 µL (10 µg) of DNase I-treated total RNA and 1 µL of 5' end-labeled DNA oligonucleotide (AWO-348). After an initial denaturation step at 95 °C for 1 min followed by a 5-min incubation on ice, 10 µL of the elongation mix (4 µL 5x first-strand buffer, 0.5 mM dNTP mix, 0.5 µL RNase Inhibitor, and 5 mM DTT) were added to the sample and incubated for 5 min at 42 °C for annealing to occur. Thereafter, 1 µL of the reverse transcriptase (1:1 dilution; SuperScript III, Thermo Fischer Scientific) was added, followed by an incubation for 1 h at 50 °C. The reaction was terminated at 70 °C for 15 min, followed by the addition of 1 µL RNase H. For preparation of the sequencing ladder, oligonucleotides (AWO-347/–348) were used to PCR-amplify the region –50 nt to +87 nt relative to the TSS of GibS from genomic DNA. A sequencing reaction was set up using the labeled oligonucleotide and the DNA cycle sequencing kit (Jena Bioscience) according to the manufacturers' instructions. Both, sample and sequencing ladder (10 µL each) were electrophoresed on a sequencing gel (10% [vol vol⁻¹] PAA-7 M urea). The gel was dried, exposed, and visualized on a phosphorimager (FLA-3000 Series, Fuji).

In vitro transcription and radiolabeling of RNA. Primer pairs carrying a T7 promoter and amplifying templates from genomic DNA are listed in Supplementary Data (AWO-241/–311, AWO-329/–313, AWO-331/–332, AWO-333/–334). Similarly, a hexanucleotide substitution (UAAUCU → AUUAGA) in the R2-targeted segment of *BT_3893* ('mut') was generated by overlap extension PCR using primer pairs AWO-329/–369 and 359/–313. In vitro transcription was carried out using the MEGAscript T7 kit (Ambion) followed by DNase I digestion (1 U, 37 °C for 15 min). The in vitro-transcribed RNA product was extracted from a 6% (vol vol⁻¹) PAA-7 M urea gel by comparison to the LowRange RNA ladder (ThermoFisher Scientific) and subsequently eluted in RNA elution buffer (0.1 M NaOAc, 0.1% SDS, 10 mM EDTA) overnight in a thermoblock at 8 °C with shaking at 1400 rpm. The RNA was precipitated using ethanol:NaOAc (30:1) mix, washed with 75% ethanol, and re-suspended in 20 µL water (65 °C for 5 min).

For radioactive labeling, 50 pmol of the in vitro transcript were dephosphorylated using 25 U of calf intestine alkaline phosphatase (NEB) in a 50 µL reaction volume and incubated for 1 h at 37 °C. Following extraction with phenol:chloroform:isoamylalcohol (P:C:I, 25:24:1), the RNA was precipitated as described above. The dephosphorylated RNA (20 pmol) was 5' end-labeled (20 µCi of ³²P-γATP) using 1 U of polynucleotide kinase (NEB) for 1 h at 37 °C in a 20 µL reaction volume. The labeled RNA was purified on a G-50 column (GE Healthcare) and extracted from a PAA gel as above.

Structure and in-line probing. Structure probing was carried out in 10 µL reactions¹⁰⁸. Briefly, 5' end-labeled RNA (0.2 pmol) was denatured at 95 °C for 1 min and chilled on ice for 5 min. 1 µg of yeast RNA was added followed by the addition of 10x structure buffer (Ambion) and the reaction incubated at 37 °C for 10 min, prior to the addition of 2 µL of freshly prepared lead (II) acetate (25 mM; Fluka), 2 µL of RNase T1 (0.01 U µL⁻¹; Ambion) or 2 µL of RNase III (New England Biolabs) and 1 mM DTT. The reactions were incubated for 45 s (lead [II] acetate), 3 min (RNase T1) or 10 min (RNase III), respectively, at 37 °C, stopped by adding 12 µL gel loading buffer II (Ambion), and stored on ice. The alkaline hydrolysis ladder was prepared by incubating 0.4 pmol labeled RNA with 9 µL 1x alkaline hydrolysis buffer (Ambion) and incubated at 95 °C for 5 min. The RNase T1 ladder was prepared by incubating 0.4 pmol labeled RNA in 8 µL of 1x sequencing buffer (Ambion) at 95 °C for 1 min followed by the addition of 1 µL RNase T1 (0.1 U µL⁻¹) and incubation at 37 °C for 5 min. Both reactions were stopped by the addition of 12 µL loading buffer II and stored on ice. Immediately prior to loading, the samples were denatured at 95 °C for 3 min and resolved on an 8% (vol vol⁻¹) PAA-7 M urea sequencing gel. The gel was visualized after appropriate exposure as described above.

In-line probing assays were performed by incubating 0.2 pmol labeled RNA for 40 h at room temperature in 1x in-line probing buffer (100 mM KCl, 20 mM MgCl₂, 50 mM Tris-HCl, pH 8.3). RNase T1 and alkaline hydrolysis ladders were prepared as mentioned above. Reactions were stopped by the addition of 10 µL colorless loading dye (1.5 mM EDTA, pH 8, 10 M urea) on ice. Samples were run on a 10% (vol vol⁻¹) PAA-7 M urea sequencing gel and visualized as described above.

Electrophoretic mobility shift assays. EMSAs were performed in 10 µL reactions as described¹⁰⁹, containing 1x RNA structure buffer (SB; Ambion), 5' end-labeled GibS RNA (4 nM final concentration), 1 µg yeast RNA (~4 µM final concentration), and putative target mRNA segments (~150 nt in length) at the following final concentrations: 0, 8, 16, 32, 64, 128, 256, 512, and 1024 nM. The reactions were incubated at 37 °C for 1 h following which 3 µL of 5x native loading dye (0.2% bromophenol blue, 0.5x TBE, 50% glycerol) were added. Electrophoresis of the samples was carried out on a native 6% (vol vol⁻¹) PAA gel in 0.5x TBE buffer at 4 °C and 300 V, for 3 h. The gels were dried and visualized as above.

Theta-Base. The Theta-Base website provides an interactive tool to interrogate the transcriptome structure and gene expression profile of *B. thetaiotaomicron* type strain VPI-5482 as determined in the course of this work. The user interface is implemented in Python using Dash by Plotly⁴². The Dash app uses the Flask web framework for the back end and is deployed with Gunicorn, a Python WSGI HTTP Server for UNIX. The experimental data are stored in an efficient SQLite database¹¹⁰. The app offers the possibility to create heatmaps of user- or pre-defined lists of coding or noncoding RNAs over the experimental conditions tested in this work (early exponential, mid-exponential, and stationary phase in TYG). The data can be exported to, and modified in, the interactive graphing library Plotly and the heatmaps can be saved. In addition, the website runs an instance of JBrowse⁴³ to explore operon structures, the position of coding and noncoding loci, TSSs, and terminators in the *B. thetaiotaomicron* transcriptome.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All sequencing data are available at NCBI Gene Expression Omnibus (<http://www.ncbi.nlm.nih.gov/geo>) under the accession number GSE144492. The source data underlying Figs. 1a–d, 3c, d, 4d, e, 5a, b, d, e and Supplementary Figs. 1b, d, e, 5, 6a, b, 7a and 8a–c are provided as a Source Data file. Sequences in FASTA format were downloaded from the sRNA database (BSRD; <http://www.bac-srna.org/BSRD/index.jsp>) and non-redundant protein database (nr database; <ftp://ftp.ncbi.nih.gov/blast/db/FASTA/>). Complete genome sequences were downloaded from ENA (<https://www.ebi.ac.uk/genomes/bacteria.html>), accessed 1/12/2017).

Code availability

Core software central to the conclusions drawn in this study are publicly available and their usage parameters described in the appropriate sections above.

Received: 24 February 2020; Accepted: 23 June 2020;

Published online: 16 July 2020

References

- Wexler, H. M. *Bacteroides*: the good, the bad, and the nitty-gritty. *Clin. Microbiol. Rev.* **20**, 593–621 (2007).
- Martens, E. C. et al. Recognition and degradation of plant cell wall polysaccharides by two human gut symbionts. *PLoS Biol.* **9**, e1001221 (2011).

3. Lee, S. M. et al. Bacterial colonization factors control specificity and stability of the gut microbiota. *Nature* **501**, 426–429 (2013).
4. Martens, E. C., Roth, R., Heuser, J. E. & Gordon, J. I. Coordinate regulation of glycan degradation and polysaccharide capsule biosynthesis by a prominent human gut symbiont. *J. Biol. Chem.* **284**, 18445–18457 (2009).
5. Martens, E. C., Chiang, H. C. & Gordon, J. I. Mucosal glycan foraging enhances fitness and transmission of a saccharolytic human gut bacterial symbiont. *Cell Host Microbe* **4**, 447–457 (2008).
6. Bayley, D. P., Rocha, E. R. & Smith, C. J. Analysis of cepA and other *Bacteroides fragilis* genes reveals a unique promoter structure. *FEMS Microbiol. Lett.* **193**, 149–154 (2000).
7. Vingadassalom, D. et al. An unusual primary sigma factor in the Bacteroidetes phylum. *Mol. Microbiol.* **56**, 888–902 (2005).
8. Baez, W. D. et al. Global analysis of protein synthesis in *Flavobacterium johnsoniae* reveals the use of Kozak-like sequences in diverse bacteria. *Nucleic Acids Res.* **47**, 10477–10488 (2019).
9. Ndamukong, I. C., Gee, J. & Smith, C. J. The extracytoplasmic function sigma factor EcfO protects *Bacteroides fragilis* against oxidative stress. *J. Bacteriol.* **195**, 145–155 (2013).
10. D'Elia, J. N. & Salyers, A. A. Effect of regulatory protein levels on utilization of starch by *Bacteroides thetaiotaomicron*. *J. Bacteriol.* **178**, 7180–7186 (1996).
11. Sonnenburg, E. D. et al. A hybrid two-component system protein of a prominent human gut symbiont couples glycan sensing in vivo to carbohydrate metabolism. *Proc. Natl Acad. Sci. USA* **103**, 8834–8839 (2006).
12. Sonnenburg, E. D. et al. Specificity of polysaccharide use in intestinal bacteroides species determines diet-induced microbiota alterations. *Cell* **141**, 1241–1252 (2010).
13. Ravcheev, D. A., Godzik, A., Osterman, A. L. & Rodionov, D. A. Polysaccharides utilization in human gut bacterium *Bacteroides thetaiotaomicron*: comparative genomics reconstruction of metabolic and regulatory networks. *Bmc Genomics* <https://doi.org/10.1186/1471-2164-14-873> (2013).
14. Chang, C. et al. A novel transcriptional regulator of L-arabinose utilization in human gut bacteria. *Nucleic Acids Res.* **43**, 10546–10559 (2015).
15. Wagner, E. G. & Romby, P. Small RNAs in bacteria and archaea: who they are, what they do, and how they do it. *Adv. Genet.* **90**, 133–208 (2015).
16. Jeters, R. T., Wang, G. R., Moon, K., Shoemaker, N. B. & Salyers, A. A. Tetracycline-associated transcriptional regulation of transfer genes of the *Bacteroides* conjugative transposon CTnDOT. *J. Bacteriol.* **191**, 6374–6382 (2009).
17. Waters, J. L. & Salyers, A. A. The small RNA RteR inhibits transfer of the *Bacteroides* conjugative transposon CTnDOT. *J. Bacteriol.* **194**, 5228–5236 (2012).
18. Cao, Y., Forstner, K. U., Vogel, J. & Smith, C. J. cis-Encoded small RNAs, a conserved mechanism for repression of polysaccharide utilization in *Bacteroides*. *J. Bacteriol.* **198**, 2410–2418 (2016).
19. Townsend, G. E. II et al. Dietary sugar silences a colonization factor in a mammalian gut symbiont. *Proc. Natl Acad. Sci. USA* **116**, 233–238 (2019).
20. Sharma, C. M. et al. The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature* **464**, 250–255 (2010).
21. Sharma, C. M. & Vogel, J. Differential RNA-seq: the approach behind and the biological insight gained. *Curr. Opin. Microbiol.* **19**, 97–105 (2014).
22. Yu, S. H., Vogel, J. & Forstner, K. U. ANNOgesic: a Swiss army knife for the RNA-seq based annotation of bacterial/archaeal genomes. *GigaScience* <https://doi.org/10.1093/gigascience/giy096> (2018).
23. Mi, H. Y., Muruganujan, A., Casagrande, J. T. & Thomas, P. D. Large-scale gene function analysis with the PANTHER classification system. *Nat. Protoc.* **8**, 1551–1566 (2013).
24. Georg, J. & Hess, W. R. Widespread antisense transcription in prokaryotes. *Microbiol. Spectrum* <https://doi.org/10.1128/microbiolspec.RWR-0029-2018> (2018).
25. Wade, J. T. & Grainger, D. C. Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. *Nat. Rev. Microbiol.* **12**, 647–653 (2014).
26. Llorens-Rico, V. et al. Bacterial antisense RNAs are mainly the product of transcriptional noise. *Sci. Adv.* <https://doi.org/10.1126/sciadv.1501363> (2016).
27. Ramachandran, V. K., Shearer, N. & Thompson, A. The primary transcriptome of *Salmonella enterica* Serovar Typhimurium and its dependence on ppGpp during late stationary phase. *PLoS ONE* <https://doi.org/10.1371/journal.pone.0092690> (2014).
28. Berger, P. et al. The primary transcriptome of the *Escherichia coli* O104:H4 pAA plasmid and novel insights into its virulence gene expression and regulation. *Sci. Rep.-Uk* **6**, 35307 (2016).
29. Kroger, C. et al. The primary transcriptome, small RNAs and regulation of antimicrobial resistance in *Acinetobacter baumannii* ATCC 17978. *Nucleic Acids Res.* **46**, 9684–9698 (2018).
30. Kingsford, C. L., Ayanbule, K. & Salzberg, S. L. Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biol.* <https://doi.org/10.1186/Gb-2007-8-2-R22> (2007).
31. Lorenz, R. et al. ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).
32. Kim, D. et al. Comparative analysis of regulatory elements between *Escherichia coli* and *Klebsiella pneumoniae* by genome-wide transcription start site profiling. *PLoS Genet.* **8**, e1002867 (2012).
33. Brock, J. E., Pourshahian, S., Gilberti, J., Limbach, P. A. & Janssen, G. R. Ribosomes bind leaderless mRNA in *Escherichia coli* through recognition of their 5'-terminal AUG. *Rna-a Publ. Rna Soc.* **14**, 2159–2169 (2008).
34. Yanofsky, C., Konan, K. V. & Sarsero, J. P. Some novel transcription attenuation mechanisms used by bacteria. *Biochimie* **78**, 1017–1024 (1996).
35. Duval, M. & Cossart, P. Small bacterial and phagic proteins: an updated view on a rapidly moving field. *Curr. Opin. Microbiol.* **39**, 81–88 (2017).
36. Storz, G., Wolf, Y. I. & Ramamurthi, K. S. Small proteins can no longer be ignored. *Annu. Rev. Biochem.* **83**, 753–777 (2014).
37. Sberro, H. et al. Large-scale analyses of human microbiomes reveal thousands of small, novel genes. *Cell* **178**, 1245–1259 (2019).
38. Bailey, T. L. et al. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–W208 (2009).
39. Frith, M. C., Saunders, N. F., Kobe, B. & Bailey, T. L. Discovering sequence motifs with arbitrary insertions and deletions. *PLoS Comput Biol.* **4**, e1000071 (2008).
40. Colgan, A. M., Cameron, A. D. & Kroger, C. If it transcribes, we can sequence it: mining the complexities of host-pathogen-environment interactions using RNA-seq. *Curr. Opin. Microbiol.* **36**, 37–46 (2017).
41. Kroger, C. et al. An infection-relevant transcriptomic compendium for *Salmonella enterica* Serovar Typhimurium. *Cell Host Microbe* **14**, 683–695 (2013).
42. Inc., P. T. *Collaborative Data Science*, <https://plot.ly> (2015).
43. Buels, R. et al. JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.* **17**, 66 (2016).
44. Wehner, S., Damm, K., Hartmann, R. K. & Marz, M. Dissemination of 6S RNA among Bacteria. *RNA Biol.* **11**, 1468–1479 (2014).
45. Kalvari, I. et al. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res.* **46**, D335–D342 (2018).
46. Costliow, Z. A. & Degnan, P. H. Thiamine acquisition strategies impact metabolism and competition in the gut microbe *Bacteroides thetaiotaomicron*. *Msystems* **2**, e00116–e00117 (2017).
47. Costliow, Z. A., Degnan, P. H. & Vanderpool, C. K. Thiamine pyrophosphate riboswitches in *Bacteroides* species regulate transcription or translation of thiamine transport and biosynthesis genes. Preprint at <https://www.biorxiv.org/content/10.1101/867226v1> (2019).
48. Hershberg, R., Altuvia, S. & Margalit, H. A survey of small RNA-encoding genes in *Escherichia coli*. *Nucleic Acids Res.* **31**, 1813–1820 (2003).
49. Tajkarimi, M. & Wexler, H. M. CRISPR-Cas systems in *Bacteroides fragilis*, an important pathobiont in the human gut microbiome. *Front. Microbiol.* **8**, 2234 (2017).
50. Bland, C. et al. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinforma.* **8**, 209 (2007).
51. Gottesman, S. & Storz, G. Bacterial small RNA regulators: versatile roles and rapidly evolving variations. *Cold Spring Harb. Perspect. Biol.* **3**, a003798 (2011).
52. Torarinsson, E. & Lindgreen, S. WAR: Webserver for aligning structural RNAs. *Nucleic Acids Res.* **36**, W79–W84 (2008).
53. Storz, G., Vogel, J. & Wassarman, K. M. Regulation by small RNAs in bacteria: expanding frontiers. *Mol. Cell* **43**, 880–891 (2011).
54. Schwalm, N. D. III, Townsend, G. E. II & Groisman, E. A. Multiple signals govern utilization of a polysaccharide in the gut bacterium *Bacteroides thetaiotaomicron*. *MBio* <https://doi.org/10.1128/mBio.01342-16> (2016).
55. Bobrovskyy, M. & Vanderpool, C. K. Regulation of bacterial metabolism by small RNAs using diverse mechanisms. *Annu. Rev. Genet.* **47**, 209–232 (2013).
56. Lim, B., Zimmermann, M., Barry, N. A. & Goodman, A. L. Engineered regulatory systems modulate gene expression of human commensals in the gut. *Cell* **169**, 547–558 (2017).
57. Terrapon, N. et al. PULDB: the expanded database of Polysaccharide Utilization Loci. *Nucleic Acids Res.* **46**, D677–D683 (2018).
58. Mann, M., Wright, P. R. & Backofen, R. IntaRNA 2.0: enhanced and customizable prediction of RNA-RNA interactions. *Nucleic Acids Res.* **45**, W435–W439 (2017).
59. Lagier, J. C. et al. Culturing the human microbiota and culturomics. *Nat. Rev. Microbiol.* **16**, 540–550 (2018).
60. Whitaker, W. R., Shepherd, E. S. & Sonnenburg, J. L. Tunable expression tools enable single-cell strain distinction in the gut microbiome. *Cell* **169**, 538–546.e512 (2017).
61. Garcia-Bayona, L. & Comstock, L. E. Streamlined genetic manipulation of diverse *Bacteroides* and *Parabacteroides* isolates from the human gut microbiota. *MBio* <https://doi.org/10.1128/mBio.01762-19> (2019).

62. Bencivenga-Barry, N. A., Lim, B., Herrera, C. M., Trent, M. S. & Goodman, A. L. Genetic manipulation of wild human gut *Bacteroides*. *J. Bacteriol.* <https://doi.org/10.1128/JB.00544-19> (2019).
63. Wexler, A. G. & Goodman, A. L. An insider's perspective: *Bacteroides* as a window into the microbiome. *Nat. Microbiol.* **2**, 17026 (2017).
64. Mishra, S. & Imlay, J. A. An anaerobic bacterium, *Bacteroides thetaiotaomicron*, uses a consortium of enzymes to scavenge hydrogen peroxide. *Mol. Microbiol.* **90**, 1356–1371 (2013).
65. Jiang, X. et al. Invertible promoters mediate bacterial phase variation, antibiotic resistance, and host adaptation in the gut. *Science* **363**, 181–187 (2019).
66. Holmqvist, E. & Vogel, J. RNA-binding proteins in bacteria. *Nat. Rev. Microbiol.* **16**, 601–615 (2018).
67. Frohlich, K. S. & Vogel, J. Activation of gene expression by small RNA. *Curr. Opin. Microbiol.* **12**, 674–682 (2009).
68. Papenfort, K. & Vanderpool, C. K. Target activation by regulatory RNAs in bacteria. *FEMS Microbiol. Rev.* **39**, 362–378 (2015).
69. Sedlyarova, N. et al. sRNA-mediated control of transcription termination in *E. coli*. *Cell* **167**, 111–121 (2016).
70. Silva, I. J. et al. SraL sRNA interaction regulates the terminator by preventing premature transcription termination of rho mRNA. *Proc. Natl Acad. Sci. USA* **116**, 3042–3051 (2019).
71. Romeo, T. & Babitzke, P. Global regulation by CsrA and its RNA antagonists. *Microbiol. Spectrum* <https://doi.org/10.1128/microbiolspec.RWR-0009-2017> (2018).
72. Bossi, L. & Figueroa-Bossi, N. Competing endogenous RNAs: a target-centric view of small RNA regulation in bacteria. *Nat. Rev. Microbiol.* **14**, 775–784 (2016).
73. Miyakoshi, M., Chao, Y. & Vogel, J. Regulatory small RNAs from the 3' regions of bacterial mRNAs. *Curr. Opin. Microbiol.* **24**, 132–139 (2015).
74. Albrecht, M. et al. The transcriptional landscape of *Chlamydia pneumoniae*. *Genome Biol.* **12**, R98 (2011).
75. Toledo-Arana, A. et al. The *Listeria* transcriptional landscape from saprophytism to virulence. *Nature* **459**, 950–956 (2009).
76. Vogel, J. et al. RNomics in *Escherichia coli* detects new sRNA species and indicates parallel transcriptional output in bacteria. *Nucleic Acids Res.* **31**, 6435–6443 (2003).
77. Dugar, G. et al. High-resolution transcriptome maps reveal strain-specific regulatory features of multiple *Campylobacter jejuni* isolates. *PLoS Genet.* **9**, e1003495 (2013).
78. Rogers, T. E. et al. Dynamic responses of *Bacteroides thetaiotaomicron* during growth on glycan mixtures. *Mol. Microbiol.* **88**, 876–890 (2013).
79. Pudlo, N. A. et al. Symbiotic human gut bacteria with variable metabolic priorities for host mucosal glycans. *mBio* **6**, e01282–01215 (2015).
80. Durica-Mitic, S., Gopel, Y. & Gorke, B. Carbohydrate utilization in bacteria: making the most out of sugars with the help of small regulatory RNAs. *Microbiol. Spectrum* <https://doi.org/10.1128/Microbiolspec.Rwr-0013-2017> (2018).
81. Woodson, S. A., Panja, S. & Santiago-Frangos, A. Proteins that chaperone RNA regulation. *Microbiol. Spectrum* <https://doi.org/10.1128/microbiolspec.RWR-0026-2018> (2018).
82. Gorski, S. A., Vogel, J. & Doudna, J. A. RNA-based recognition and targeting: sowing the seeds of specificity. *Nat. Rev. Mol. Cell Biol.* **18**, 215–228 (2017).
83. El-Gebali, S. et al. The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
84. Masse, E., Vanderpool, C. K. & Gottesman, S. Effect of RyhB small RNA on global iron use in *Escherichia coli*. *J. Bacteriol.* **187**, 6962–6971 (2005).
85. Papenfort, K. et al. SigmaE-dependent small RNAs of *Salmonella* respond to membrane stress by accelerating global omp mRNA decay. *Mol. Microbiol.* **62**, 1674–1688 (2006).
86. Urban, J. H. & Vogel, J. Translational control and target recognition by *Escherichia coli* small RNAs in vivo. *Nucleic Acids Res.* **35**, 1018–1037 (2007).
87. Corcoran, C. P. et al. Superfolder GFP reporters validate diverse new mRNA targets of the classic porin regulator, MicF RNA. *Mol. Microbiol.* **84**, 428–445 (2012).
88. Pinilla-Redondo, R., Riber, L. & Sorensen, S. J. Fluorescence recovery allows the implementation of a fluorescence reporter gene platform applicable for the detection and quantification of horizontal gene transfer in anoxic environments. *Appl. Environ. Microb.* **84**, e02507–e02517 (2018).
89. Mimee, M., Tucker, A. C., Voigt, C. A. & Lu, T. K. Programming a human commensal bacterium, *Bacteroides thetaiotaomicron*, to sense and respond to stimuli in the murine gut microbiota. *Cell Syst.* **1**, 62–71 (2015).
90. Huntzinger, E. et al. Staphylococcus aureus RNAPIII and the endoribonuclease III coordinately regulate spa gene expression. *EMBO J.* **24**, 824–835 (2005).
91. Barquist, L. & Vogel, J. Accelerating discovery and functional analysis of small RNAs with new technologies. *Annu. Rev. Genet.* **49**, 367–394 (2015).
92. Goodman, A. L. et al. Identifying genetic determinants needed to establish a human gut symbiont in its habitat. *Cell Host Microbe* **6**, 279–289 (2009).
93. Wu, M. et al. Genetic determinants of in vivo fitness and diet responsiveness in multiple human gut *Bacteroides*. *Science* **350**, aac5992 (2015).
94. Liu, H. P. M. et al. Large-scale chemical-genetics of the human gut bacterium *Bacteroides thetaiotaomicron*. Preprint at <https://www.biorxiv.org/content/10.1101/573055v1> (2019).
95. Koropatkin, N. M., Martens, E. C., Gordon, J. I. & Smith, T. J. Starch catabolism by a prominent human gut symbiont is directed by the recognition of amylose helices. *Structure* **16**, 1105–1115 (2008).
96. Bacic, M. K. & Smith, C. J. Laboratory maintenance and cultivation of *Bacteroides* species. *Curr. Protoc. Microbiol.* **Unit 13C**, 11 (2008).
97. Eriksson, S., Lucchini, S., Thompson, A., Rhen, M. & Hinton, J. C. Unravelling the biology of macrophage infection by gene expression profiling of intracellular *Salmonella enterica*. *Mol. Microbiol.* **47**, 103–118 (2003).
98. Forstner, K. U., Vogel, J. & Sharma, C. M. READemption—a tool for the computational analysis of deep-sequencing-based transcriptome data. *Bioinformatics* **30**, 3421–3423 (2014).
99. Nicol, J. W., Helt, G. A., Blanchard, S. G., Raja, A. & Loraine, A. E. The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. *Bioinformatics* **25**, 2730–2731 (2009).
100. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
101. Thompson, J. D., Higgins, D. G., Gibson, T. J. & CLUSTAL, W. improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680 (1994).
102. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).
103. Wheeler, T. J. & Eddy, S. R. nhmmer: DNA homology search with profile HMMs. *Bioinformatics* **29**, 2487–2489 (2013).
104. Lindgreen, S. et al. Robust identification of noncoding RNA from transcriptomes requires phylogenetically-informed sampling. *PLoS Comput Biol.* **10**, e1003907 (2014).
105. Griffiths-Jones, S. RALEE-RNA ALIGNMENT Editor in Emacs. *Bioinformatics* **21**, 257–259 (2005).
106. Sittka, A., Pfeiffer, V., Tedin, K. & Vogel, J. The RNA chaperone Hfq is essential for the virulence of *Salmonella typhimurium*. *Mol. Microbiol.* **63**, 193–217 (2007).
107. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).
108. Sharma, C. M., Darfeuille, F., Plantinga, T. H. & Vogel, J. A small RNA regulates multiple ABC transporter mRNAs by targeting C/A-rich elements inside and upstream of ribosome-binding sites. *Genes Dev.* **21**, 2804–2817 (2007).
109. Pernitzsch, S. R., Tirier, S. M., Beier, D. & Sharma, C. M. A variable homopolymeric G-repeat defines small RNA-mediated posttranscriptional regulation of a chemotaxis receptor in *Helicobacter pylori*. *Proc. Natl Acad. Sci. USA* **111**, E501–E510 (2014).
110. Hipp, R. et al. SQLite v3.22.0 (SQLite Development Team, 2015).
111. Carver, T., Thomson, N., Bleasby, A., Berriman, M. & Parkhill, J. DNAPlotter: circular and linear interactive genome visualization. *Bioinformatics* **25**, 119–120 (2009).

Acknowledgements

We are grateful to members of the Westermann, Barquist, Faber, Vogel, and Brochado groups for fruitful discussions of our research and to Jörg Vogel, Caroline Taouk, Erik Holmqvist, and Carsten Kröger for constructive comments on the manuscript. We thank the group of Andrew Goodman for beta-testing Theta-Base and Sebastian Winter for the *B. thetaiotaomicron* *Δtdk* strain (AWS-003) and the strain harboring the pEchange_Δtdk plasmid (AWS-011). Thanks to the Core Unit Systems Medicine in Würzburg, particularly to Tobias Hecker, Sascha Dietrich, and Elena Katzwitsch, for support with cDNA library preparation and RNA-seq, to Falk Ponath for help with ANNOgesic, and to Michael Kütt for technical assistance with the launch of Theta-Base. This work was supported by a DFG grant to A.J.W. (WE 6689/1-1) and by the IZKF at the University of Würzburg (project Z-6).

Author contributions

D.R. and A.J.W. conceptualized the study. D.R. and S.R. conducted experiments. D.R., L.B., and A.J.W. analyzed data. D.R., L.J., and L.B. performed computational studies. D.R., L.J., L.B., and A.J.W. established 'Theta-Base'. D.R. and A.J.W. wrote the original manuscript draft that all co-authors reviewed and edited. L.B. and A.J.W. supervised the project. A.J.W. acquired funding.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41467-020-17348-5>.

Correspondence and requests for materials should be addressed to A.J.W.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permission information is available at <http://www.nature.com/reprints>

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020