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

Daniel Ryan¹, Laura Jenniches¹, Sarah Reichardt¹, Lars Barquist¹,² & Alexander J. Westermann¹,³

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

https://doi.org/10.1038/s41467-020-17348-5
Bacteroides are Gram-negative, obligate anaerobic, non-motile, non-spore forming rods, and among the most abundant bacterial genera in the human intestine. 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 conditions. The extensive metabolic potential encoded in the B. thetaiotaomicron genome, including 88 defined polysaccharide utilization loci (PULs), 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^Abf, as well as an arsenal of alternative, extra-cytoplasmic function (ECF) sigma factors. The consensus recognition sequence of σ^Abf was deduced by manual inspection of 23 promoter sequences in B. fragilis as ‘TAnnTTTG’ at the –35 region and a ‘TTTG’ motif at around the –33 region. Inspection of global RNA-seq data recently confirmed the –7 consensus in another Bacteroidetes member, Flavobacterium johnsoniae, whereas the –33 motif could not be identified. While in B. fragilis the sequence motif recognized by the oxidative stress-related alternative ECF sigma-factor EcFO was recently identified, no such motif is currently known for any of the B. thetaiotaomicron alternative ECFs. Besides sigma factors, SusR-like regulators and hybrid two-component signal transduction systems, 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.

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 efficiency. 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 initiation. As of now, only two sRNAs are known in Bacteroides spp. RteR is a trans-encoded sRNA of 90 nt that promotes discordant expression of the tra operon, required to assemble the mating apparatus for the transfer of the conjugative transposon, CTnDot. 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 glycans. 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 process. 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) 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, 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. 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 log to exponential phase, in mid-exponential and stationary phase—and analyzed the resulting data using the ANNOgesic bioinformatic tool. 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-extension. 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 System revealed enrichment of gene ontology (GO)-terms including DNA recombination, DNA integration, and integral membrane components (Supplementary Data 1).
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 protocol. While functions for a number of specific antisense transcripts have been described, 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 predicted aTSSs in 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.\textsuperscript{26,27}

**Global assessment of transcript features.** By combining Trans-termHP\textsuperscript{30} and RNAfold\textsuperscript{31}, ANNOgesic predicts intrinsic transcription terminators, which—together with the experimentally mapped TSSs—enables the deduction of transcription 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.\textsuperscript{20,32} In line with leaderless mRNAs being considered rare in Gram-negative species\textsuperscript{33}, 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 only other long 5′ UTRs to actually contain short open reading frames (sORFs; Supplementary Data 3) reminiscent of leader peptides mediating transcription attenuation.\textsuperscript{34} 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.\textsuperscript{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 study\textsuperscript{37} (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. \textsuperscript{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 negligence 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) were relatively highly expressed in our mRNA-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 MEME\textsuperscript{36} and GLAM2\textsuperscript{39} 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 σB\textsuperscript{ABE} promoter of *B. fragilis*. 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 platforms\textsuperscript{40} such as the SalCom\textsuperscript{41} and Acineteto\textsuperscript{29} 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 Plothy.\textsuperscript{42} Moreover, a linkout to JBrowse\textsuperscript{43} 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-hire.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* phylum\textsuperscript{44}, 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 θ-proteobacterium *Helicobacter pylori*\textsuperscript{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 database\textsuperscript{45}) 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 re-identified three putative thiamine pyrophosphate
(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. Conserva-
tional 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,
including that of *B. fragilis*\(^4\). In contrast, no CRISPR array nor any Cas (CRISPR-associated) proteins have been annotated in *B. thetaiotaomicron*. In agreement with this, the CRISPR Recognition Tool\(^5\), which is included in the ANNOgesic pipeline, failed to identify any CRISPR-RNAs in our dataset.

**Experimental validation of *B. thetaiotaomicron* 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 RNA\(^18\), was not expressed when *B. thetaiotaomicron* 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, 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 RNAs), 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 sRNAs.

Within its gut niche, B. thetaiotaomicron uses simple sugars as signals to control the expression of metabolic modules. Given their ability to mediate rapid responses to metabolic stimuli, 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 showed expression pattern, we probed significant 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 (∆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 PITD35 promoter (designated PITD) was re-inserted at an unrelated position into the ∆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 anhydroetacrylne (aT; 200 ng mL−1), the resulting strain expressed GibS to wild-type levels (Supplementary Fig. 6). Interfering with GibS expression in bacteria grown in minimal medium in the 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, ∆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 ∆gibS), comprise a diguconin encoding an α-galactosidase and a β-glucosidase, respectively, and are predicted to be part of the putative PUL22 involved in arabinoxylan degradation (Supplementary Fig. 7b). The genes seemingly repressed by GibS (de-repressed in ∆gibS) included BT_1655 (encoding a hypothetical protein within the CPS5 locus; Supplementary Fig. 7c) and BT_3893 (the second gene within a
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 program predicted GibS to anneal within the CDS of *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 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. Secondary structure prediction of *B. thetaiotaomicron* GibS was performed using the webserver for aligning structural RNAs and reflects the consensus structure of 14 different prediction programs, considering both, minimal free energy and covariation. The colors of the ribonucleobases represent the conserved and less-conserved ribonucleobases, respectively. The numbers denote the position relative to the 5' end of GibS comprised its highly conserved and unstructured 5' portion (Fig. 4). To experimentally test these predictions, we performed electrophoretic mobility shift assays (EMSAs) using

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 *gibS* mutant (Supplementary Fig. 7d). We rule out polar effects of sRNA deletion, as *BT_0769* expression reverted to wild-type levels when GibS was re-introduced in trans in the ΔgibS 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).
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 (Kd = 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.
Discussion

Only recently and thanks to technological breakthroughs that allow anaerobic culturing without the need for extensive lab equipment and the development of versatile protocols for genetic manipulation, 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 community.

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 eABr 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 promoters, but could not previously be verified globally within the Bacteroidetes phylum. 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 oxidative oxygen species. However, this motif differs substantially from the recognition sequence of the sigma factor EcO, which protects B. fragilis against oxidative stress 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 adaptation, 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, where they regulate target mRNAs through imperfect base-pairing interactions mediated by short seed sequences and, in many Gram-negative species, dedicated RNA chaperones. In the majority of the described cases, sRNA annealing to the SD sequence and/or start codon, interferes with translation initiation and induces target decay. Conversely, certain sRNAs may promote translation of their target by unfolding inhibitory structures that otherwise block ribosome binding. In addition, activating trans-encoded sRNAs in y-proteobacteria may induce target gene expression by binding to nascent mRNA leaders and interfering with Rho-mediated premature transcription termination. 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 protein. Conceptually related, an emerging class of sRNA sponges bind and titrate other sRNAs, creating complex post-transcriptional gene expression control networks. 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 sequenced 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 transcripts is not restricted to Proteobacteria. The full repertoire of B. thetaiotaomicron non-coding RNAs therefore appears consistent with the numbers reported in other bacteria. However, we want to point out that antisense transcription might at least partially be due to the appearance of random promoters, 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 straightforward construction of a clean deletion mutant. In the course of our studies, we validated the TSS of GibS predicted by sRNA-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 ΔgibS 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. 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-branched 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 organisms. 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 decay. Here we 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 mRNA. 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 control. 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 length. 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, −III, −III, −P, −R, and −Z are present in B. thetaiotaomicron according to Pfam) 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 ΔgibS 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. 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. 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 recovery. Similarly, luciferase assays—albeit depending on molecular oxygen—have already been successfully employed in Bacteroides and β-galactosidase assays with lacZ-target fusions represent an oxygen-independent alternative.

Modern sequencing-based technologies have proven useful to gain global perspectives on bacterial sRNAs 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 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, 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 (ΔgibS) was generated as previously described. Briefly, 1-kb sequences flanking the region to be deleted were amplified by PCR (using oligos AWO-053/−054, AWO-052/−055, 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-μL aliquot of this reaction was transformed into electro-competent E. coli S17-1 Δpir. Transformants were conjugated with a tdk deletion mutant of B. thetaiotaomicron (Δtkd) and conjugants counter-selected on 5-fluoro-2′-deoxyuridine (FUDR) plates. Single recombinants were selected on Brain Heart Infusion Supplemented (BHS) agar containing 200 μg mL−1 gentamicin and 25 μg mL−1 erythromycin. Double recombinants, resulting in either scarless deletion mutants or wild-type revertants, were selected by growth on BHS agar containing 25 μg mL−1 FUDR and an inability to grow on BHS agar containing 25 μg 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 gib5 compensation strain (gib5+) was constructed using a variant of the pNB2U vector system (Supplementary Data 4). The full gib5 gene was integrated into the vector by Gibson assembly (Addgene-105757) to ensure fusion of the gene from its native TSS. This resulted in a deletion of the proximal tetO2 (Tt) operator downstream of the promoter P1T07, while leaving the second operator (Tt0) intact (45). This construct, expressing the Gib5 sRNA from its +1 nucleotide under control of the P1T07 promoter, was conjugated into the wild strain and selected on BHI agar containing 25 μg·mL⁻¹ erythromycin. Successful insertion was confirmed by PCR (AWO-160-161) and Sanger sequencing.

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

For dRNA-seq (Figs. 1–3), wild-type B. thetaiotaomicron VPI-5482 (AWP-001) was grown overnight in 5 mL pre-reduced TGY medium followed by sub-culturing (1:100) into 50 mL fresh pre-reduced TGY. Total RNA was isolated by hot phenol extraction and a mix of each of 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 mix 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 0.6% lysozyme (0.5 mg mL−1) and 60 μL 10% SDS for 15 min at 64 °C, followed by the addition of 66 μL of 3 M NaAc. Phenol extraction (750 mL; Roti-Aqua Phenol) was performed at 64 °C for 6 min with the subsequent addition of 750 mL Chloroform. RNA was precipitated from the aqueous phase with twice the volume of 30:1 ethanol3 (NaAcO, pH 6.5) mix and incubated at −80 °C overnight. After centrifugation, pellets were washed with 75% (vol−vol−1) ethanol and resuspended in 50 μL H2O. Contaminating genomic DNA was removed by treating 40 μg total RNA with 5 U of DNase I (Fermentas) and 9 mg RNAse-free DNaseI (Roche). DNA was digested with the addition of the gene validation option that relates identification of sRNAs 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-hzi.de/en/datasets/bacteroides).

**Search for promoter motifs.**

Conserved DNA sequences upstream of TSSs were identified by ANNOgesic (usage: anngenic promoter integrating both, MEME for gapless motifs68 and GLAM2 for gapped motifs59). Default parameters were used with the addition of the flag –net to 50 to increase sensitivity. Representative sequence containing conserved motifs were aligned using ClustalW (version 2.1) 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: anngenic terminator). TransTERMHP scans genome sequences for the presence of Rho-independent terminators99. 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-hzi.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 300 nt. 5′-antisense sRNAs 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 a processing site or shared terminator with the parental mRNA. Intronic sRNA candidates were ruled out for this step because the presence of a TSS at the 3′ 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 study\(^5\) as a benchmark. The lowest U of minimum average coverage to win 5 min of modifications, ANNOgesc 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 hmmner 3.1b\(^1\), similarly to as previously described\(^5\). In each round of iteration, a search was run with the flags --open 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 editor\(^105\).

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 structure\(^5\).

### 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 previously\(^106\). 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 70% (vol/vol\(^1\)) sulfuric acid washed RNA probes Hybridizing quick buffer (Carl Roth AG) at 42°C and exposed as described. Visualization was achieved using a phosphorimager (FLA-3000 Series, Fuji) and images were quantified using ImageJ\(^107\).

For qRT-PCR assays, reverse transcription and PCR amplification were performed in the same reaction mix containing 1 µL of DNase I-treated total RNA (adjusted in water to 10 µg/µL), 5 µL master mix (No ROX SYBR MasterMix blue 1
t M) and 0.1 µL forward 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 were dried, exposed, and visualized on a phosphorimager (FLA-1001221) or PAA-7 M urea sequencing gel and visualized as described above.

### Electrophoretic mobility shift assays

EMSAs were performed in 10 µL reactions as described\(^109\), containing 1 µRNA 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 sequences (~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 5 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 Plotly\(^42\). 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 database. 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 JBrowse\(^43\) 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, 4d, 5a, b, d, e and Supplementary Figs. 1b, d, e, 5, 6a, b, 7a and 8a are provided as a Source Data file. Sequences in FASTA format were downloaded from the sRNA database (BRSD; http://www.bac-sra.org/BRSD/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

1. Wester, H. M. *Bacteroides*: the good, the bad, and the nitty-gritty. *Clin. Microbiol. Rev.* 20, 593–621 (2007).
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 Escherichia coli and Klebsiella pneumoniae by genome-wide transcription start site profiling. *PLoS Genet.* **8**, e1002867 (2012).

5. Brock, J. E., Poursabzhan, S., Giliberti, J., Limbach, P. A. & Janssen, G. R. Ribosomes bind leaderless mRNAs in Escherichia coli through recognition of synonymous AUGs. *Cell Host Microbe* **5**, 447–450 (2009).

6. Yanovsky, K., Konan, K. V. & Sarsero, J. P. Some novel transcription attenuation mechanisms used by bacteria. *Biochimie* **78**, 1017–1024 (1996).

7. Duval, M. & Cossart, P. Small bacterial and phagic proteins: an updated view on a rapidly moving field. *Curr. Opin. Microbiol.* **39**, 81–88 (2017).

8. 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).

9. Vanderpool, C. K. et al. Analysis of a Bacteroides thetaiotaomicron CRISPR-Cas system: a tool for automatic detection of CRISPR recognition sites. *Nucleic Acids Res.* **42**, 1551–1566 (2014).

10. Frith, M. C., Saunders, N. F., Cobb, B. & Bailey, T. L. Discovering sequence motifs with arbitrary insertions and deletions. *PLoS Comput. Biol.* **4**, e1000701 (2008).

11. Bobrovskyy, M. & Vanderpool, C. K. Regulation of bacterial metabolism by small RNAs and their regulatory systems modulate gene expression of human commensals in the gut. *Cell Host Microbe* **14**, 683–695 (2013).

12. Inc., P. T. Collaborative Data Science, *plotly.js*. (2019).

13. Buels, R. et al. BUniversity of California, Davis, UC Davis. (2019).

14. Calo, V., Forstner, K. U., Vogel, J. & Smith, C. J. cis-Encoded small RNAs, a conserved mechanism for repression of polysaccharide utilization in Bacteroides thetaiotaomicron. *J. Bacteriol.* **190**, 5228–5236 (2018).

15. Cao, Y., Forstner, K. U., Vogel, J. & Smith, C. J. cis-Encoded small RNAs, a conserved mechanism for repression of polysaccharide utilization in Bacteroides thetaiotaomicron. *J. Bacteriol.* **190**, 2410–2418 (2016).

16. Townsend, G. E. II et al. Dietary sugar silences a colonisation factor in a mammalian gut symbiont. *Proc. Natl Acad. Sci. USA* **116**, 233–238 (2019).

17. Sharma, C. M. & Vogel, J. Differential RNA-seq: the approach behind and the biological insight gained. *Curr. Opin. Microbiol.* **19**, 97–105 (2014).

18. Yu, S. H., Vogel, J. & Forstner, K. U. ANNOgenic: a Swiss army knife for the RNA-seq based annotation of bacterial/archaeal genomes. *GigaScience* https://doi.org/10.1186/s13742-014-0096-0 (2015).

19. 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).

20. Georg, J. & Hess, W. R. Widespread antisense transcription in prokaryotes. *Microbiol. Spectrum* https://doi.org/10.1128/microbiolspec.RWR-0029-2018 (2018).

21. Wade, J. T. & Grainger, D. C. Pervasive transcription: illuminating the dark matter of bacterial transcriptionalities. *Nat. Microbiol.* **12**, 647–653 (2017).

22. Llorens-Rico, V. et al. Bacterial antisense RNAs are mainly the product of transcriptional noise. *Sci. Adv.* https://doi.org/10.1126/sciadv.aad15036 (2016).

23. Ramachandran, S. K., Shearer, N. & Thompson, A. The primary transcriptome of Sal monella enterica Serovar Typhimurium and its dependence on PpGpp during late stationary phase. *PLoS ONE* https://doi.org/10.1371/journal.pone.0092940 (2014).

24. Berger, P. et al. The primary transcriptome of Escherichia coli O104:H4 pPA plasmid and novel insights into its virulence gene expression and regulation. *Sci. Rep.* **6**, 35207 (2016).

25. 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).

26. Kingsford, C. L., Ayanbule, K. & Salzberg, S. L. Rapid, accurate, computational detection of Bho-independent transcript terminators illuminates their relationship to DNA uptake. *Genome Biol.* https://doi.org/10.1186/Gb-2007-8-2-R22 (2007).
62. Bengivena-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. C. & Vogel, J. Activation of gene expression by small RNA. Mol. Microbiol. 90, 1356–1371 (2013).
64. Mishra, S. & Imlay, J. A. An anaerobic bacterium, Bacteroides thetaiotaomicron, uses a consortium of enzymes to scavenge hydrogen from the human colon. Mol. Microbiol. 62, 1705–1718 (2006).
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. Siva, I. J. et al. 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. Spectr. 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-Aran, 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, 876–889 (2013).
79. Pudlo, N. A. et al. Symbiotic human gut bacteria with variable metabolic catabolism by a prominent human gut symbiont is directed by the recognition of homopolymeric G-repeat de

80. Durica-Mitic, S., Gopel, Y. & Gorke, B. Carbohydrate utilization in bacteria: a perspective: sowing the seeds of specificity to claenec joyfulmאותs. DNA homology search with profile HMMs. Bioinformatics 29, 2487–2489 (2013).
81. Lindgreen, S. et al. Robust identification of noncoding RNA from transcriptomes requires phylogenetically-informed sampling. PLoS Comput. Biol. 10, e1003907 (2014).
82. Griffiths-Jones, S. RALEE-RNA Alignment Editor in Emacs. Bioinformatics 21, 257–259 (2005).
83. Silita, 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).
84. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. Nat. Methods 9, 671–675 (2012).
85. 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).
86. Permitzch, S. R.,Tierir, 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).
87. Hopp, R. et al. SQLite v3.22.0 (SQLite Development Team, 2015).
88. Carver, T., Thomson, N., Blesasy, 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 Tasok, 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 group’s support with cDNA normalization. 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 Tasok, 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 group’s support with cDNA normalization.
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.I.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