Supporting Information

Kröger et al. 10.1073/pnas.1201061109

SI Materials and Methods

***Salmonella* Typhimurium SL1344 Genome Sequence.** *S.* Typhimurium genomic DNA was fragmented by sonication, and several libraries were generated in pUC18 using size fractions ranging from 1.0 to 2.5 kb. The whole genome was sequenced to a depth of 9× coverage from M13mp18 (insert size 1.4–2 kb) and pUC18 (insert size 2.2–4.2 kb) small insert libraries, using dye terminator chemistry on ABI3700 automated sequencers. End sequences from larger insert plasmid (pBACe3.6, 12–30 kb insert size) libraries were used as a scaffold. The sequence was assembled, finished, and annotated as described previously (1). The finished and chromosome and plasmid sequences have been submitted to the European Molecular Biology Laboratory (accessions FQ312003, HE654724, HE654725, and HE654726).

Isolation of Total RNA from *S.* Typhimurium. Total RNA from *S.* Typhimurium SL1344 and ST4/74 was isolated from cells grown aerobically until early stationary phase (ESP) (OD_{600} = 2.0) in Lennox (L)-broth, using hot phenol or TRIzol. The RNA of the multiple growth stages was isolated using the hot phenol method (2), as described earlier (3). Four ODs of *Salmonella* cells were harvested, mixed with 2/5 vol of the cell suspension of ice-cold "stop solution" [19% ethanol, 1% acidic phenol (pH 4.3)], and incubated on ice for 30 min to prevent RNA degradation (4). The cells were then pelleted by centrifugation for 10 min (3,220 × *g*, 4 °C). The major part of the supernatant was discarded, and the cell pellet was resuspended in the remaining liquid and transferred into a 1.5-mL Eppendorf cap. Subsequently, the cell suspension was centrifuged for 1 min at 20,000 × *g* and 4 °C and the supernatant was discarded. The cell pellet was resuspended in 1 mL TRIzol and transferred into a 2-mL heavy phase-lock tube, and 400 μL chloroform was added and immediately mixed for 10 s. After incubation for 2 min at room temperature, the suspension was centrifuged for 15 min at 20,000 × *g*. The upper phase containing the RNA (~500 μL) was then transferred into a 1.5-mL Eppendorf cap, 450 μL isopropanol was added, and the suspension was mixed by inverting and then incubated at room temperature for 30 min. The precipitated RNA was pelleted by centrifugation for 30 min at 20,000 × *g* and the supernatant was discarded. The pellet was then washed in 350 μL ethanol (70 %) and the suspension was again centrifuged for 10 min at 20,000 × *g*. The supernatant was discarded, and the pellet was air-dried, resuspended in RNase-free water by shaking (900 rpm) for 5 min in a heat block (65 °C), and then stored at −80 °C until further use. The RNA quality was analyzed using an Agilent Bioanalyzer 2100. RNA concentrations were measured using the nanodrop or the Qubit fluorometer (Invitrogen) according to the manufacturer's manual.

cDNA Library Construction and Sequencing. Total RNA from two biological replicates was isolated from ESP as described above and digested with DNase. The 454 pyrosequencing was done as described previously (5). Flow cell reverse transcription sequencing (FRT-seq) cDNA libraries were generated from 250 ng of total RNA as in Mamanova et al. (6). All procedures including library preparation, on-flowcell reverse transcription, and clustering were as detailed in the protocol in ref. 6 with the exception of the final library purification step. The dual adapter ligated material was resolved in denaturing acrylamide gels and eluted from gel slices corresponding to 150–350 nt in size. Sequencing was carried out on the GAIIx platform, utilizing a v2 Paired End Cluster kit and two v3 SBS-36 cycle kits (Illumina).

Illumina [small regulatory (s)RNA-Enriched Libraries, Terminator Exonuclease (TEX)+/−]. cDNA library construction and sequencing was performed by Vertis Biotechnologie AG. In brief, total RNA from two biological replicates was isolated from *S.* Typhimurium strain SL1344 or ST4/74 grown until ESP as described above, digested with DNase I, and enriched for small RNA molecules (<500 nt) using the mirVana kit (Ambion). One of these samples was additionally treated with the 5′-monophosphate–dependent terminator exonuclease to degrade 5′-monophosporylated RNA. Subsequently, the 3′ ends were subjected to poly(A)-tailing using poly(A) polymerase, pyrophosphate groups were removed from the 5′ terminus using tobacco acid pyrophosphatase (TAP), and an RNA adapter was ligated to the 5′ end of the RNA. First-strand synthesis was performed using an oligo(dT) adapter and M-MLV-RNaseH-reverse transcriptase, following PCR amplification of cDNA using high-fidelity DNA polymerase to ~20–30 ng/μL. The cDNAs were purified using the Agencourt AMPure XP kit (Beckman Coulter Genomics) and analyzed by capillary electrophoresis.

Mapping of RNA Sequencing (RNA-Seq) Libraries. Mapping of reads to the reference genome was carried out by the software tool *segemehl* (7) with accuracy set to 100%. To increase coverage, mismatched nucleotides at the lower-quality 3′ end were removed from the reads. The mapping statistics for all libraries are shown in Dataset S1. Uniquely mapped reads (i.e., reads that mapped to a single genomic location) were used for further analysis. The mapped reads were formatted into graph files for visualization in the Integrated Genome Browser (IGB) (8).

Identification of Transcriptional Start Sites. Transcriptional start sites were assigned manually, using the IGB (8). The first criterion for annotating a transcriptional start included enrichment in the TEX+ over the TEX− library, as described earlier (9). In the case of no enrichment, a transcriptional start was assigned if at least three biological replicates agreed on a nucleotide position and if the location was plausible in relation to an adjacent ORF. The start sites were categorized into classes as described previously (9).

Identification of Candidate sRNAs. New small RNA candidates were identified from RNA-seq and Hfq coimmunoprecipitation coupled with RNAseq (Hfq-coIP-seq) as short (usually <250 nt), stable transcripts, present in more than two biological replicates. Enrichment in Hfq-coIP-seq over the relevant control sample indicated association with Hfq.

Northern Blot Analysis. Northern blots using radioactively labeled probes were conducted as described earlier with primers listed in Table S1 (5). Other Northern blots were done using the DIG Northern blot starter kit (Roche) according to the manufacturer's manual. Primers for in vitro transcription to generate DIG-labeled riboprobes are listed in Table S1.

5′-RACE. 5′-RACE experiments were carried out as described before (10), using total RNA extracted from *Salmonella* grown in LB broth to an OD_{600} of 2. Gene-specific amplification was achieved by PCR, using the linker-specific primer JVO-0367 in combination with JVO-1268 (*ompA*), JVO-7120 (*phoP*), JVO-7123 (*prgH*), JVO-7121 (*osmC*), JVO-9037 (*hilD*), JVO-0933 (*sodB*), JVO-8160 (*yfgE*), JVO-8158 (*yibP*), JVO-8157 (*invF*), and JVO-8153 (*slyA*) (Table S1).

Chromatin Immunoprecipitation (ChIP) Analysis. ChIP was done with ESP-grown cultures of SL1344 by the method of Dillon et al. (11). The following monoclonal antibodies were used in this study: RNA polymerase β′-subunit (Neoclone; W0001) and σ^{70} (RpoD; Neoclone; W0004). Chromatin immunoprecipitation coupled with microarray hybridization (ChIP-chip) was done as in Dillon et al. (11), using microarrays that carried 43,453 60-mer oligonucleotides tiled throughout the *S.* Typhimurium SL1344 chromosome and pSLT plasmid. Microarrays were scanned as, and the ChIPOTle algorithm was used to define, peaks of enrichment in all ChIP-chip datasets by using a sliding-window approach and then estimating the significance of enrichment for a genomic region, using a standard Gaussian error function (12). ChIPOTle assigns a *P* value to the average $\log_2$ ratio within each window and then corrects for multiple comparisons, using the Bonferroni correction. The significance *P*-value cutoff assigned was $P = 0.00001$ with a subsequent corrected *P*-value cutoff of $P = 0.05$. A window size of 500 bp and a step size of 125 bp were used for analyzing the datasets, the rationale being that the ChIP procedure produces DNA fragments of ∼500 bp in size. The complete ChIP-on-chip datasets have been submitted to the Gene Expression Omnibus (GEO) database (accession GSE35827).

Identification of RNAP and σ^{70} Binding Sites and Promoter Consensus Motifs. DNA sequences were extracted from the *S.* Typhimurium SL1344 genome sequence, using Artemis release 11 (13). Significantly overrepresented DNA sequence motifs were identified using unbiased motif-searching algorithms MEME 4.4.0 (14) and BioProspector (15). σ^{70} (RpoD) binding sites are composed of two discrete sequence blocks (−10 and −35) with variable spacing between blocks; thus preliminary searches for motifs were conducted using BioProspector, which is particularly suited to identifying two-block motifs. BioProspector was used to search for 6- to 12-bp motifs separated by 10–25 bp, and this search consistently identified a very strong σ^{70} −10 motif in DNA bound by σ^{70}, but failed to detect a consensus −35 motif. The σ^{70} −35 motif is highly degenerate, necessitating a further refinement of motif searching to test for its presence. To this end, promoter regions were further divided into segments ranging from −2 to −35 (for the −10 motif) and from −25 to −50 (for the −35 motif) relative to transcription start sites, and these were searched using MEME with the following parameters: Motifs could range in size from 4 to 15 bp, each DNA sequence was expected to contain one sequence matching the motif, and only the forward DNA strand was analyzed. To determine the *Escherichia coli* σ^{70} binding motif, the same MEME parameters were used to search DNA sequences containing the 857 annotated *E. coli* σ^{70} sites that were retrieved from RegulonDB 7.0 (16). Sequence logos were generated using WebLogo (17), and base positions within the motifs are numbered using the coordinate system of Shultzaberger et al. (18).

Determination of sRNA Conservation. To determine the sRNAs conservation, the set of 113 experimentally verified sRNAs (Dataset S1) was aligned against 29 bacterial whole genomes, using Exonerate version 2.2.0 with the ungapped alignment option (19). Exonerate software was chosen because it can precisely identify the homology of the short sRNA sequences across entire genomes (20, 21) and the algorithm behind this tool allows rapid implementation of heuristics to approximate many complex alignment models (19). The resulting percentage of identity, on the basis of the full sRNA sequence length, provided a distance matrix expressing a percentage of identity for each sRNA against each genome. The matrix was clustered using hierarchical clustering and then rendered in a heat map using R software.

1. Parkhill J, et al. (2001) Complete genome sequence of a multiple drug resistant Salmonella enterica serovar Typhi CT18. *Nature* 413:848–852.
2. Mattatall NR, Sanderson KE (1996) Salmonella typhimurium LT2 possesses three distinct 23S rRNA intervening sequences. *J Bacteriol* 178:2272–2278.
3. Sittka A, Pfeiffer V, Tedin K, Vogel J (2007) The RNA chaperone Hfq is essential for the virulence of Salmonella typhimurium. *Mol Microbiol* 63:193–217.
4. Tedin K, Bläsi U (1996) The RNA chain elongation rate of the lambda late mRNA is unaffected by high levels of ppGpp in the absence of amino acid starvation. *J Biol Chem* 271:17675–17686.
5. Sittka A, et al. (2008) Deep sequencing analysis of small noncoding RNA and mRNA targets of the global post-transcriptional regulator, Hfq. *PLoS Genet* 4:e1000163.
6. Mamanova L, et al. (2010) FRT-seq: Amplification-free, strand-specific transcriptome sequencing. *Nat Methods* 7:130–132.
7. Hoffmann S, et al. (2009) Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput Biol* 5:e1000502.
8. Nicol JW, Helt GA, Blanchard SG, Jr., Raja A, Loraine AE (2009) The Integrated Genome Browser: Free software for distribution and exploration of genome-scale datasets. *Bioinformatics* 25:2730–2731.
9. Sharma CM, et al. (2010) The primary transcriptome of the major human pathogen Helicobacter pylori. *Nature* 464:250–255.
10. Vogel J, et al. (2003) RNomics in Escherichia coli detects new sRNA species and indicates parallel transcriptional output in bacteria. *Nucleic Acids Res* 31:6435–6443.
11. Dillon SC, et al. (2010) Genome-wide analysis of the H-NS and Sfh regulatory networks in Salmonella Typhimurium identifies a plasmid-encoded transcription silencing mechanism. *Mol Microbiol* 76:1250–1265.
12. Buck MJ, Nobel AB, Lieb JD (2005) ChIPOTle: A user-friendly tool for the analysis of ChIP-chip data. *Genome Biol* 6:R97.
13. Rutherford K, et al. (2000) Artemis: Sequence visualization and annotation. *Bioinformatics* 16:944–945.
14. Bailey TL, Williams N, Misleh C, Li WW (2006) MEME: Discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res* 34(Web Server issue):W369–W373.
15. Liu X, Brutlag DL, Liu JS (2001) BioProspector: Discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput* 127–138.
16. Gama-Castro S, et al. (2011) RegulonDB version 7.0: Transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Res* 39(Database issue):D98–D105.
17. Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: A sequence logo generator. *Genome Res* 14:1188–1190.
18. Shultzaberger RK, Chen Z, Lewis KA, Schneider TD (2007) Anatomy of Escherichia coli sigma70 promoters. *Nucleic Acids Res* 35:771–788.
19. Slater GS, Birney E (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6:31.
20. Bolduc F, Hoareau C, St-Pierre P, Perreault JP (2010) In-depth sequencing of the siRNAs associated with peach latent mosaic viroid infection. *BMC Mol Biol* 11:16.
21. Esteve-Codina A, et al. (2011) Exploring the gonad transcriptome of two extreme male pigs with RNA-seq. *BMC Genomics* 12:552.
22. Hallin PF, Binnewies TT, Ussery DW (2008) The genome BLASTatlas-A GeneWiz extension for visualization of whole-genome homology. *Mol Biosyst* 4:363–371.

Fig. S1. (*A*) Circular maps of *S*. Typhimurium SL1344 and plasmids pSLT^{SL1344}, pCol1B9^{SL1344} (p2), and pRSF1010^{SL1344}. For the chromosome, from the outermost circle in, the outer circle 1 shows the size in base pairs. Circles 2 and 3 show the position of coding sequences (CDS) transcribed in a clockwise and an anticlockwise direction, respectively. Genes in circles 2 and 3 are color coded according to the function of their gene products: dark green, membrane or surface structures; yellow, central or intermediary metabolism; cyan, degradation of macromolecules; red, information transfer/cell division; cerise, degradation of small molecules; pale blue, regulators; salmon pink, pathogenicity, adaptation or prophage; black, energy metabolism; orange, conserved hypothetical; pale green, unknown; and brown, pseudogenes. Circle 4 shows a plot of G + C content (in a 10-kb window). Circle 5 shows a plot of GC skew [(G − C)/(G + C), in a 10-kb window]. For the plasmids, the outer circle 1 indicates the size in base pairs. Circles 2 and 3 show the position of CDS transcribed in a clockwise and an anticlockwise direction. Circles 4 and 5 are as above for G + C content and GC skew. Genes are color coded accordingly: salmon pink, pathogenicity/adaptation;

Legend continued on following page

dark gray, partitioning and maintenance; red, plasmid replication; dark green, membrane proteins, secreted proteins; dark blue, resistance to inorganic compounds/metal ions/UV stress; sky blue, conjugal transfer, mobilization; dark pink, antibiotic resistance, restriction/modification systems; yellow, host metabolism; light green, unknown functions; mid-blue, regulators; orange, conserved hypothetical proteins; brown, pseudogenes and partial genes (remnants); light pink, IS elements; light gray, miscellaneous functions. (*B*) BLASTatlas (22) of the 4.878 Mb *Salmonella enterica* serovar Typhimurium SL1344 chromosome. The DNA parameters percentage of AT, GC skew, global direct repeats, and global inverted repeats are included in the innermost lanes. The results of BLAST analysis of the proteins encoded by 29 bacterial strains are included in the outermost lanes. The bacterial strains were chosen to show diversity across the *Salmonella* genus and to include more distantly related members of the *Enterobacteriaceae*, namely *Serratia marcescens*, *Pectobacterium atrosepticum*, *Shigella flexneri*, and *E. coli* strains. Regions of interest are indicated outside the BLAST lanes and include *Salmonella* pathogenicity islands (SPI) 1, 2, 3, 4, 5, 6, 9, 11, and 12. In addition, the Gifsy and Fels phages are shown, as well as other important regions. As expected, there is full homology of the coding regions of SPI1–5 across most *Salmonella* genomes, with the exception of *Salmonella bongori*, which lacks SPI2. The SPI6 island is conserved in only a minority of the *S. enterica* serovars. There is almost no conservation of these pathogenicity islands throughout the genomes of non-*Salmonella* strains.

Fig. S2. Confirmation of transcriptional start sites by 5′-RACE and comparison of published start sites in *S*. Typhimurium with results obtained in this study. (*A–J*) Agarose gels of RT-PCR products of tobacco acid pyrophosphate treated (TAP; T+) and mock-treated (T−) and of a control PCR reaction with genomic DNA as template (ctr.). The arrowheads mark the enriched band in TAP-treated samples, indicating the cDNA of the respective primary RNA species. A DNA size marker is shown on the left [M, sizes in base pairs (bp)]. (*K*) Locations of published transcriptional starts in *S*. Typhimurium are compared with the transcriptional start sites (TSS) identified in this study (Dataset S2). The *x* axis depicts the difference of the published TSS relative to the TSS obtained in this study in nucleotides.

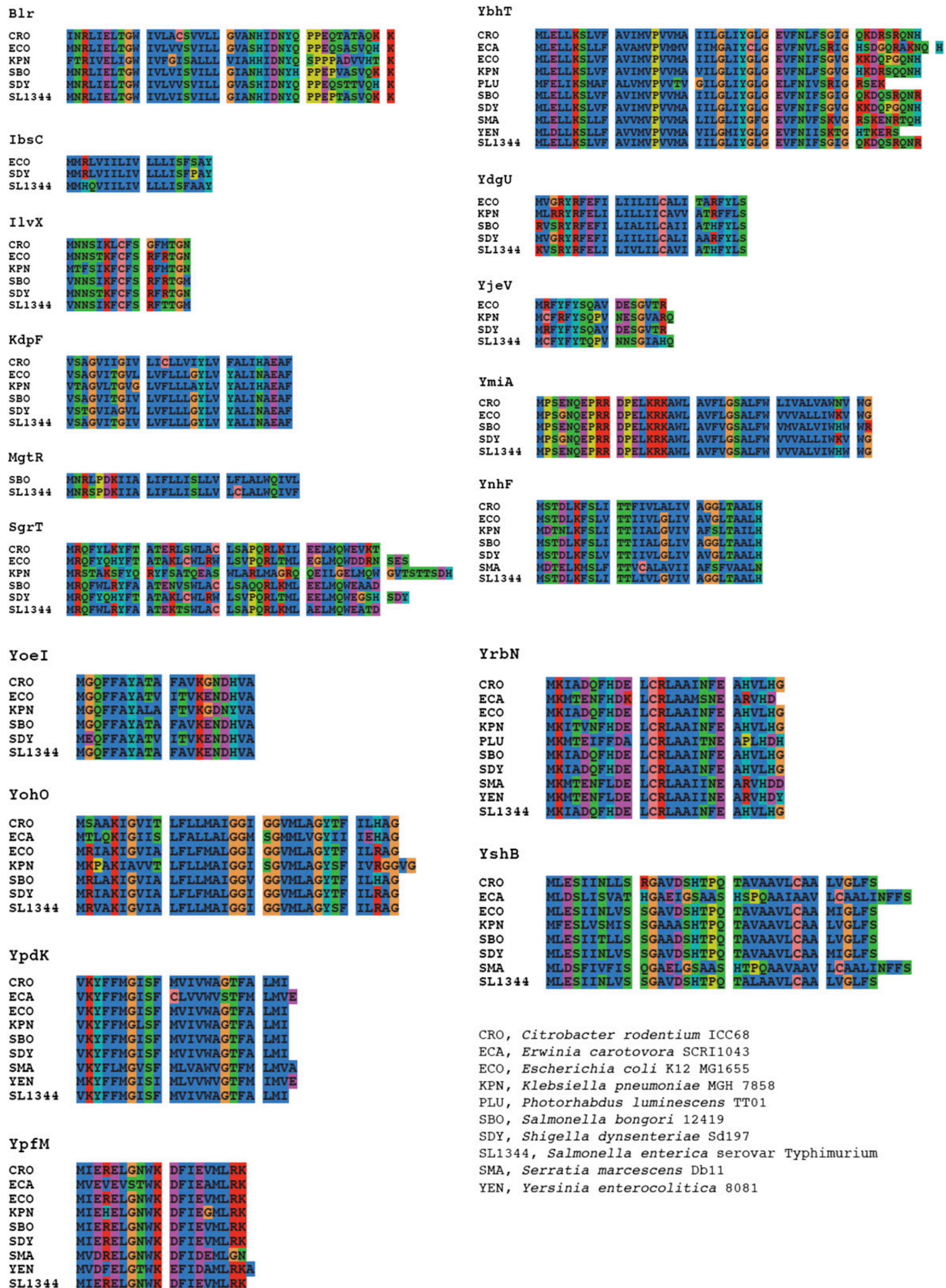

Fig. S3. Amino acid sequence alignment of small ORF-encoded proteins present in *S.* Typhimurium SL1344 with other enteric bacteria (sORFs are detailed in Dataset S1).

Fig. S4. (Continued)

Fig. S4. (Continued)

Fig. S4. Identification of new small RNAs by RNA-seq and validation by Northern blot. New small RNAs were identified by RNA-seq and subsequently validated by Northern blot, using primers listed in Table S1. *Left* (of each section) shows the Northern blot, and *Right* (of each section) shows the RNA-seq data visualized in the Integrated Genome Browser (IGB). The scaling for the RNA-seq data is as follows: For the samples TEX_1, RNA-seq_1, TEX_2, RNA_seq_2, FRT-seq, FRT-seq_dep, Hfq-coIP, and coIP_control 0–10 mapped sequence reads are shown; for TEX_3, RNA-seq_3, and sRNA 0–100 mapped sequence reads are shown.

Fig. S5. Conservation of *S.* Typhimurium sRNAs within enteric bacteria. Heat map shows the conservation of *S.* Typhimurium SL1344 sRNAs in 29 genome sequences of bacteria belonging to the family Enterobacteriaceae (details in Fig. 7 of the main text).

Table S1. Oligonucleotides used in this study

Name	Sequence	Purpose
JVO-0367	ACTGACATGGAGGAGGGA	5′-RACE; linker-specific primer
JVO-0933	CGATCTCTTCCAGCGATT	5′-RACE; *sodB*
JVO-1268	GCCATCATTGTGAATGAAG	5′-RACE; *ompA*
JVO-7120	GTAGTAATCAGCTTCCCTGG	5′-RACE; *phoP*
JVO-7121	AACCGGGTATTGAATCCA	5′-RACE; *osmC*
JVO-7123	TACCACAAAGAGTGTTCGG	5′-RACE; *prgH*
JVO-8153	CTAAAGCATCGATGTCTGACG	5′-RACE; *slyA*
JVO-8157	GCGGAACACGTCTGTATAAA	5′-RACE; *invF*
JVO-8158	CTAAGTACGCTGGCGTAGAAC	5′-RACE; *yibP*
JVO-8160	CCAGAAACTTGCGAAAGTTT	5′-RACE; *yfgE*
JVO-9037	CTTTTAATATGTGTTGAGACGC	5′-RACE; *hilD*
ASO-2	GATTACCACAACCACATCACG	Antisense oligo to probe for STnc1020
ASO-7	CGGTATTCACCGGCAACA	Antisense oligo to probe for STnc1060
ASO-9	AGGGGTCACCAACGGCT	Antisense oligo to probe for STnc1080
ASO-12	AAGCAAACCAGCCGAATCTC	Antisense oligo to probe for STnc1110
ASO-13	CCCTGCAGAGAGCGATAAC	Antisense oligo to probe for STnc1120
ASO-16	GCAGTGTGTTAATTCCGCC	Antisense oligo to probe for STnc1150
ASO-17	GGTTGGCGTAACCAGATCA	Antisense oligo to probe for STnc1160
ASO-19	TTTTGCAGCAATACTCAGGG	Antisense oligo to probe for STnc1180
rp_1380_f	CGGTACGGGGAATTGGTGC	Riboprobe template for STnc1380
rp_1380_r	GAATTAATACGACTCACTATAGTAGATCACTTTGATTGAACTC	Riboprobe template for STnc1380
rp_470_f	TGAAATAATAAGAGAGCGCC	Riboprobe template for STnc470
rp_470_r	GAATTAATACGACTCACTATAAAGAGGGCCAGCATCCG	Riboprobe template for STnc470
rp_1480_f	GCGAAAGGAGTGTGAAGTGG	Riboprobe template for STnc1480
rp_1480_r	GAATTAATACGACTCACTATACGCCAATCAAAGTGCATAACC	Riboprobe template for STnc1480
rp_880_f	AGGTAACGAATTCATCGCGC	Riboprobe template for STnc880
rp_880_r	GAATTAATACGACTCACTATAGGCTTCCACCCGGGAAGC	Riboprobe template for STnc880
rp_630_f	TCCATACCTTTGGAGGAGAG	Riboprobe template for STnc630
rp_630_r	GAATTAATACGACTCACTATAGCACCCGTGCCTGGGTG	Riboprobe template for STnc630
rp_1300_f	GTTGGTTTTCTCCATAAAATTTG	Riboprobe template for STnc1300
rp_1300_r	GAATTAATACGACTCACTATAGGGTGACGCCCGCGAGT	Riboprobe template for STnc1300
rp_1650_f	CCTGCTTTATAGCAAAAATAGC	Riboprobe template for STnc1650
rp_1650_r	GAATTAATACGACTCACTATATGAAAGAGCGGACGGATAAT	Riboprobe template for STnc1650
rp_1640_f	TGCGGCAGAATAACTTTCCG	Riboprobe template for STnc1640
rp_1640_r	GAATTAATACGACTCACTATAAACCGCCGATCCTGTCCAC	Riboprobe template for STnc1640
rp_1600_f	GAAAAAATGACGAGGATGACA	Riboprobe template for STnc1600
rp_1600_r	GAATTAATACGACTCACTATAGAGAGCCGCTCTCACGCA	Riboprobe template for STnc1600
rp_1220_f	ATATGATGCGCTGCGTGATC	Riboprobe template for STnc1220
rp_1220_r	GAATTAATACGACTCACTATAGGAACAGCATGTAAAACATTTAC	Riboprobe template for STnc1220
rp_1460_f	TCGTGATGCGTCAAGGCGG	Riboprobe template for STnc1460
rp_1460_r	GAATTAATACGACTCACTATACTGATCAGGTTCCGCGGAT	Riboprobe template for STnc1460

Other Supporting Information Files

Dataset S1 (xlsx)
Dataset S2 (xlsx)
Dataset S3 (xlsx)