Comparative genomics boosts target prediction for bacterial small RNAs

Patrick R. Wright[a,b], Andreas S. Richter[b], Kai Papenfort[c,d], Martin Mann[b], Jörg Vogel[c], Wolfgang R. Hess[a,e], Rolf Backofen[b,e,f,g,1], and Jens Georg[a,1]

[a]Genetics and Experimental Bioinformatics, Faculty of Biology, [e]Centre for Biological Systems Analysis, and [f]BIOSS Centre for Biological Signalling Studies, University of Freiburg, D-79104 Freiburg, Germany; [b]Bioinformatics Group, Department of Computer Science, University of Freiburg, D-79110 Freiburg, Germany; [c]Institute for Molecular Infection Biology, University of Würzburg, D-97080 Würzburg, Germany; [d]Department of Molecular Biology, Princeton University, Princeton, NJ 08544; and [g]Center for Non-Coding RNA in Technology and Health, University of Copenhagen, DK-1870 Frederiksberg C, Denmark

Edited by Gisela Storz, National Institutes of Health, Bethesda, MD, and approved July 30, 2013 (received for review February 22, 2013)

Small RNAs (sRNAs) constitute a large and heterogeneous class of bacterial gene expression regulators. Much like eukaryotic microRNAs, these sRNAs typically target multiple mRNAs through short seed pairing, thereby acting as global posttranscriptional regulators. In some bacteria, evidence for hundreds to possibly more than 1,000 different sRNAs has been obtained by transcriptome sequencing. However, the experimental identification of possible targets and, therefore, their confirmation as functional regulators of gene expression has remained laborious. Here, we present a strategy that integrates phylogenetic information to predict sRNA targets at the genomic scale and reconstructs regulatory networks upon functional enrichment and network analysis (CopraRNA, for Comparative Prediction Algorithm for sRNA Targets). Furthermore, CopraRNA precisely predicts the sRNA domains for target recognition and interaction. When applied to several model sRNAs, CopraRNA revealed additional targets and functions for the sRNAs CyaR, FnrS, RybB, RyhB, SgrS, and Spot42. Moreover, the mRNAs *gdhA, lrp, marA, nagZ, ptsI, sdhA*, and *yobF-cspC* were suggested as regulatory hubs targeted by up to seven different sRNAs. The verification of many previously undetected targets by CopraRNA, even for extensively investigated sRNAs, demonstrates its advantages and shows that CopraRNA-based analyses can compete with experimental target prediction approaches. A Web interface allows high-confidence target prediction and efficient classification of bacterial sRNAs.

regulatory RNA | *E. coli* | RNA–RNA interaction

Small RNAs (sRNAs) are ubiquitous and important regulators of gene expression in bacteria. The most common and best investigated *trans*-acting sRNAs regulate their targets posttranscriptionally by RNA–RNA interactions, often depending on the RNA chaperone Hfq (1). Individual functions of model sRNAs have been discovered primarily through extensive experimental work and may be assigned to many different stress responses and signal transduction pathways, covering virtually all aspects of bacterial growth (1, 2) and virulence (3). One of the most intriguing conceptual advances has been the identification of sRNAs as posttranscriptional regulators that act globally within complex regulatory networks. Examples for such sRNAs are GcvB, which is a major regulator of amino acid metabolism and directly controls ∼1% of all *Salmonella enterica* mRNAs (4); MicA and RybB, which together constitute the repressor arm of the Sigma E response (5); and Spot42, a global regulator of catabolite repression (6). With the advent of high-throughput sequencing and comprehensive transcriptome analysis techniques, increasing numbers of new sRNAs have been detected in bacteria belonging to diverse taxa (7, 8). However, the experimental testing and verification of sRNA targets is costly, labor intensive, and may be challenging, even in model organisms. Moreover, for most environmentally and biotechnologically relevant microbes, experimental verification is hindered further by the lack of systems for their genetic manipulation.

The reliable computational prediction of sRNA targets promises a great reduction of required wet-laboratory analyses while enabling large-scale sRNA–mRNA network analyses in genetically intractable species. However, reliable in silico prediction of mRNA targets has been challenging because of the extreme heterogeneity of sRNAs in size, structure, and the typically short and imperfect sRNA–target complementarity (9). The existing tools for the genome-scale prediction of sRNA targets evaluate the strength of a particular sRNA–target interaction by either base pair complementarity (10) or thermodynamic models (11–13). The latter are built on the observed exponential correlation between repression strength and hybridization free energy (14), which can be corrected by an energy term that reflects the accessibility of the interaction sites (11, 12). However, despite continuous improvement of target prediction methods (15), even the most accurate methods integrating interaction site accessibility scoring and additional features, such as seed regions, produce many false positives and, thus, compromise the selection of putative targets for subsequent experimental investigation (16, 17).

Furthermore, the implementation of seed sequence conservation to improve sRNA target prediction has been difficult to achieve for bacterial systems because of the great flexibility of the interaction patterns (16). It is conceivable that the interaction is preserved while the actual interaction site is not. Therefore, to predict conserved interactions, it is necessary to combine evidence for interactions in different species without resorting to a consensus interaction-based approach.

Here, we introduce a computational approach that uses phylogenetic information from an extended model of sRNA–target evolution (CopraRNA, for Comparative Prediction Algorithm

Significance

This study presents a unique approach (CopraRNA, for Comparative Prediction Algorithm for sRNA Targets) towards reliably predicting the targets of bacterial small regulatory RNAs (sRNAs). These molecules are important regulators of gene expression. Their detailed analysis thus far has been hampered by the lack of reliable algorithms to predict their mRNA targets. CopraRNA integrates phylogenetic information to predict sRNA targets at the genomic scale, reconstructs regulatory networks upon functional enrichment and network analysis, and predicts the sRNA domains for target recognition and interaction. Our results demonstrate that CopraRNA substantially improves the bioinformatic prediction of target genes and opens the field for the application to nonmodel bacteria.

Author contributions: A.S.R., R.B., and J.G. designed research; P.R.W., K.P., and J.G. performed research; P.R.W., A.S.R., M.M., and J.G. contributed new reagents/analytic tools; P.R.W., W.R.H., and J.G. analyzed data; and P.R.W., A.S.R., K.P., J.V., W.R.H., R.B., and J.G. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. E-mail: jens.georg@biologie.uni-freiburg.de or backofen@informatik.uni-freiburg.de.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1303248110/-/DCSupplemental.

for sRNA Targets). CopraRNA depends solely on the conservation of target genes (i.e., conservation of target regulation) and does not require conservation of specific interaction sequences (*SI Appendix*, Figs. S1 and S2).

By introducing a generic approach combining predictions for homologous targets in distinct organisms, we reduced the hitherto existing high false positive rate (FPR) of single-organism target prediction. Using this strategy, CopraRNA matches microarray-based experimental sRNA target prediction with respect to the number of correctly identified direct targets (Fig. 1*B* and Table 1) and the characterization of physiological functions of these sRNAs. Thus, it constitutes a significant improvement of in silico sRNA target prediction and enables competitive and functional large-scale initial screening for sRNA targets without experimental effort and costs. Application of CopraRNA to previously characterized sRNAs proposed and partially verified additional targets and functions for the sRNAs cyclic AMP activated sRNA (CyaR), FNR regulated sRNA (FnrS), RybB, RyhB, sugar transport-related sRNA (SgrS), and Spot42. Also, it suggested the *gdhA, lrp, marA, nagZ, ptsI, sdhA*, and *yobF-cspC* mRNAs as hubs targeted by up to seven different sRNAs. A Web interface for CopraRNA has been set up under http://rna.informatik.uni-freiburg.de/CopraRNA/.

Results

Prediction Strategy. CopraRNA begins with a genome-wide target prediction (12) for each considered organism, as summarized in Fig. 1*A*. The interaction energies are fitted to a general extreme value distribution and transformed into *P* values to normalize for organism-specific GC-content and dinucleotide frequency. These *P* values are combined for orthologous genes into a single *P* value per conserved interaction. Orthologous genes are determined based on the respective amino acid sequences (25); genes that are present in less than 50% of the investigated genomes are discarded. Two aspects require specific normalization. First, CopraRNA normalizes for the degree of overall dependency to account for the nonindependent *P* values that result from the

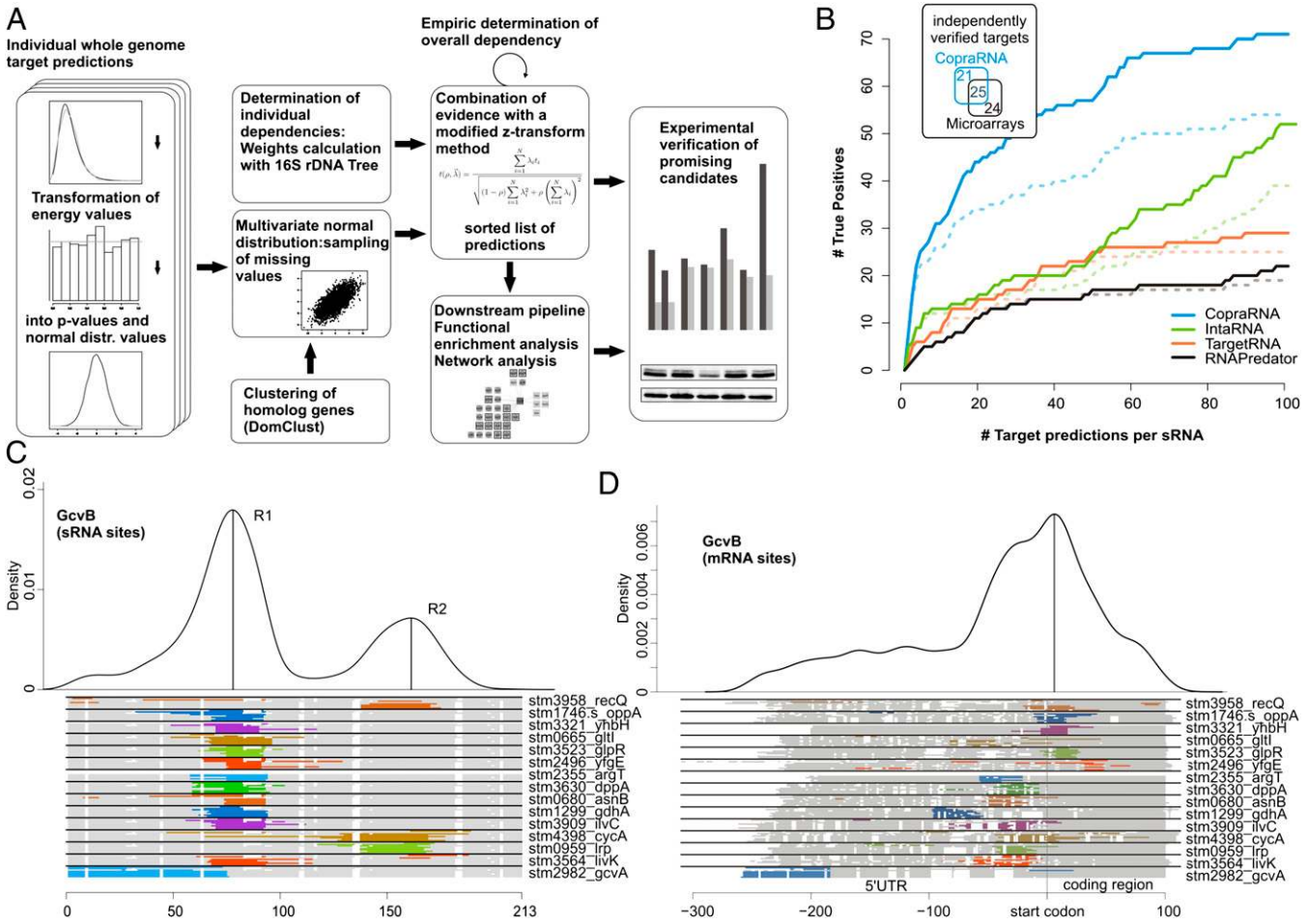

Fig. 1. (*A*) Schematic overview of the CopraRNA pipeline. (*B*) Comparison of CopraRNA predictions with microarray results and other target prediction methods. Genome-wide target predictions for 18 sRNAs in *E. coli* and *S. enterica* with 101 experimentally verified targets from the literature. The plot shows the number of correctly predicted targets (true positive predictions, *y* axis) vs. the number of target predictions per sRNA (*x* axis) for our comparative method CopraRNA and the existing single-organism–based methods IntaRNA, TargetRNA, and RNApredator. The results, including the verifications from this study, are shown with solid lines, and the results based on the benchmark set only are demarcated with a dashed line. (*Inset*) total numbers of independently verified targets detected by either CopraRNA (46 targets) or microarray experiments (49 targets) for the sRNAs CyaR, FnrS, GcvB, MicF, RyhB, SgrS, and Spot42; 25 targets were identified by both methods. The numbers refer to our benchmark dataset (*SI Appendix*, Table S1) and to the table comparing CopraRNA with different microarray experiments (Table 1). Visualization of the predicted interaction domains in GcvB (*C*) and the predicted mRNA targets of GcvB (*D*). The density plots at the top give the relative frequency of a specific sRNA or mRNA nucleotide position in the predicted sRNA–target interactions. The plots combine all predictions with a *P* value ≤0.01 in all included homologs. Local maxima indicate distinct interaction domains and are marked with upright lines. The schematic alignment of homologous sRNAs and targets at the bottom show the predicted interaction domains. The aligned regions are displayed in gray, gaps in white, and predicted interaction regions in color (color differences are for contrast only). The locus tag and gene name (if available) of a representative cluster member are given on the right.

Table 1. Comparison of CopraRNA predictions and published microarray studies

	CopraRNA			Microarray				
sRNA	No. of candidates ($P \leq 0.01$)	No. of candidates after postprocessing*	No. verified†	No. sig. diff. expr. genes‡	No. verified†	Ref.	No. overlap§ verified/ unverified	Overlap genes¶
CyaR	69	55	1 + 3‖	24 genes	4	18	1/1	*fepA*, ***ompX***
				1 gene	1	19	1/0	***ompX***
FnrS	67	41	3 + 4‖	16 genes/11 operons	6 + 1‖	20	3/0	***marA***, ***sodB***, ***yobA***
				31 genes	7 + 1‖	21	4/2	*adhP*, ***marA***, ***sfcA/maeA***, ***sodB***, *ydhD/grxD*, ***yobA***
GcvB	60	34	14	54 genes	16	4	10/3	***argT***, *aroP*, ***brnQ***, ***cycA***, ***dppA***, ***gdhA***, ***gltI***, ***lrp***, ***oppA***, ***serA***, ***sstT***, *trpE*, *yifK*
MicF	50	30	4	5 genes	4	22	2/0	***lrp***, ***ompF***
RyhB	70	37	2 + 5‖	56 genes/18 operons	3 + 1‖	23	3/3	*frdA*, ***fumA***, *msrB*, ***sdhA***, *sdhD*, ***sodB***
SgrS	66	35	2 + 1‖	6 genes	4	24	2/0	***ptsG***, ***yigL***
Spot42	85	48	4 + 3‖	16 genes	7	6	3/0	***galk***, ***gltA***, ***xylF***

The candidates after postprocessing for these sRNAs are given in Table S5.

*Top 15 targets + automatically and manually functionally enriched.

†Verified targets after postprocessing regarding the benchmark list (*SI Appendix*, Table S1), published data, and this study.

‡Significantly differentially expressed genes with regard to the respective publications.

§Genes detected by prediction and microarray (independently verified/unverified).

¶Independently verified targets are in boldface.

‖Verified in this study.

general sequence conservation between related organisms. Second, the individual dependencies have to be calculated because, in most cases, the considered organisms will not be equidistant from each other. Thus, we additionally used species-specific weights that were calculated based on 16S rDNA-based phylogenetic trees. The combination of the *P* values used a modified z-transform method, which permits adjustment for dependency in the data and a weighting based on the phylogenetic relationship (26). We defined significance thresholds either on CopraRNA *P* values or on q-values (27); the latter provide correction for multiple testing by controlling the false discovery rate (FDR). Both methods have proven useful for the analysis of the benchmark dataset. The chosen *P* value threshold of 0.01 allows for the detection of approximately half of all verified benchmark targets (*SI Appendix*, Fig. S3*A*) and was applied for the functional enrichment and network analysis. The q-value gives a measure of how many false positive predictions are expected in the group of targets called significant. True positives are all experimentally verified targets (with regard to our benchmark dataset in *SI Appendix*, Table S1) within the positive predictions, whereas false positives are all positive predictions that are no real targets, i.e., in our case, those that have not been verified experimentally. Positive predictions (also called candidates below) are all targets that match the respective threshold criterion (e.g., a *P* value ≤0.01 or a given rank); they consist of true positive and false positive predictions (statistical terms are defined also in *SI Appendix*). A reliable bioinformatic prediction tool for sRNA targets should not predict more than ∼50% of false positive targets; therefore, we chose a q-value threshold of 0.5. The validity of this approach for CopraRNA was tested with the prediction for GcvB. We assume that GcvB, with its 22 verified targets, is so far the most thoroughly investigated sRNA (4). In the CopraRNA prediction of GcvB, 37 targets are predicted with a q-value ≤0.5. Of these, 35 have homologs in *Escherichia coli* or *S. enterica*, 11 of which have been verified. Fifteen of the 35 homologs are involved in amino acid metabolism or transport, i.e., they fit to the known biological function of GcvB. This corresponds to an FDR of 69% or 57%, respectively, with regard to currently known targets and is not very far from the statistical estimate of 50%. In general, the number of significant predictions with a q-value ≤0.5 is a rough approximation of the expected number of targets and the prediction quality of the tested sRNA. A detailed description of the CopraRNA procedure is provided in *SI Appendix*.

Benchmark with Experimentally Verified Targets. To evaluate the accuracy of CopraRNA, we performed a benchmarking test on a set of 18 conserved enterobacterial sRNAs and their 101 experimentally verified mRNA targets (modified from ref. 16) using homologous sequences from three to eight organisms (*SI Appendix*, Fig. S4). Compared with predictions by the existing approaches IntaRNA (12), TargetRNA (10), and RNApredator (11) (Fig. 1*B*), CopraRNA showed a clear improvement in the sensitivity or true positive rate (sensitivity $= \frac{\#\text{ true positives}}{\#\text{ positives}}$) and positive predictive value (PPV $= \frac{\#\text{ true positives}}{\#\text{ positive predictions}}$). Based on published data, CopraRNA's top 1 target predictions were correct for 8 of 18 sRNAs (PPV: 44%), compared with 5 (PPV: 28%) for IntaRNA, 2 (PPV: 11%) for TargetRNA, and 1 (PPV: 6%) for RNApredator. When considering the top 5 and top 15 target predictions per sRNA, CopraRNA correctly detected 23 and 32, respectively, of all 101 targets (true positive rate: 23% and 32%, respectively), which constitutes a twofold increase in sensitivity compared with IntaRNA and a 2.9-fold and fourfold improvement compared with TargetRNA and RNApredator, respectively (*SI Appendix*, Table S2). In addition, our experimental verification (below) demonstrated that the existing lists of known targets are still incomplete, implying an underestimation of the true positive rate (Fig. 1*B*).

In many cases, the comparative approach resolved the problem of false negatives (i.e., verified targets missed in the prediction) in single-organism–based methods. Prominent examples are the GcvB targets *lrp* (4), *oppA* (4), and *stm3903* (4); the RybB target *ompN* (28); and the Spot42 target *gltA* (6). The ranking of these targets improved from rank 95 to 3, rank 164 to 14, rank 1,297 to 40, rank 69 to 3, and rank 392 to 2, respectively (*E. coli*- or *S. enterica*-specific prediction vs. CopraRNA prediction). The benchmark dataset and the complete ranked list of all predictions are given in *SI Appendix*, Table S1 and Table S3.

Prediction of Interaction Domains. In addition to the ranked list of predicted targets, CopraRNA provides comparative information on the putative interaction sites of the sRNA and its mRNA

targets. These data are summarized in two density plots combining all predictions with a P value ≤0.01 for a specific sRNA (Fig. 1 *C* and *D* shows the GcvB example). Based on multiple sequence alignments, these plots visualize the frequency of single residues participating in the predicted sRNA–mRNA interactions. The plots are complemented by a series of schematic alignments for both sRNAs and mRNAs that highlight organism-specific predicted interactions. From these plots, the interaction domains of the sRNA can be inferred, as they provide the combined information of accessibility, complementarity, and phylogenetic conservation.

This visualization immediately highlights the two previously described interaction regions of GcvB (4) (Fig. 1*C*), the three different interaction regions of Spot42 (6), and the single 5′ located region of RybB (9) (*SI Appendix*, Fig. S5). In agreement with the published data for Spot42, *gltA* is targeted by the first single-stranded region (6) centered at position 6 in the multiple-sequence alignment (*SI Appendix*, Figs. S5 and S6). The newly identified targets *sucC* and *gdhA* base pair with the second and third interaction region of Spot42, respectively. For *galK*, all three regions are predicted to be involved in the interaction for four of the eight investigated organisms (*SI Appendix*, Fig. S5). As previously described (4), GcvB targets *lrp* and *cycA* via region "R2" of the sRNA (Fig. 1*C*), whereas most targets (e.g., *dppA* and *oppA*) interact with region "R1." In the case of RprA, the full-length form appears to have two interaction domains, and only the distal site is retained after processing (29) (*SI Appendix*, Fig. S5), leading to a significant shift in the list of predicted targets. The mRNA plots are useful to obtain a rapid overview on the predicted interaction sites regarding their relative position and their phylogenetic conservation. The density plot also reveals the predominant interaction regions when using target sequences of the same length. For GcvB targets, there is a clear tendency toward the region near the start codon (Fig. 1*D*).

Functional Enrichment of Predicted Targets. Many well-studied sRNAs control sets of functionally related genes [e.g., RyhB, nonessential iron-binding proteins (30), GcvB, amino acid biosynthesis genes (4)]. Therefore, we analyzed the top-ranked targets of all benchmark sRNAs for functional relationships based on automated functional enrichment using the database for annotation, visualization, and integrated discovery (DAVID) (31). A combination of CopraRNA and functional enrichment provided very clear results for several sRNAs and suggested their potential involvement in diverse cellular networks (Tables S4 and S5). The DAVID Web server clusters related terms and calculates a combined enrichment score. Table 2 shows representative terms for the most strongly enriched clusters of selected sRNAs. The accuracy of this approach is demonstrated exemplarily for GcvB: this sRNA has a broad set of 22 verified target mRNAs (4) and a clearly defined function as a regulator of amino acid metabolism and transport (4). GcvB has 60 positive predictions (P value ≤0.01, *E. coli*). Seven experimentally verified targets are in the top 10 list, which supports the prediction accuracy of our algorithm and represents a PPV of 70%. Among the 60 candidate targets, 19 were annotated with the term "cellular amino acid biosynthetic process" and were significantly enriched (enrichment score ∼6.65) over background (i.e., all genes included in the prediction output). In summary, 26 of the 60 predictions were grouped as amino acid related, including genes for 11 amino acid biosynthesis proteins, 9 amino acid transporters, and 4 peptide transporters. These results are complementary

Table 2. Results of the functional enrichment analysis using the DAVID Web server (31)

sRNA	No. predicted	Enrichment score	Category	Term	No.
CyaR	69	4.95	UP_SEQ_FEATURE	Topological domain:Periplasmic	26
		3.45	SP_PIR_KEYWORDS	Cell inner membrane	32
		2.15	GOTERM_BP_FAT	GO:0005976~polysaccharide metabolic process	11
FnrS	67	2.43	SP_PIR_KEYWORDS	Flavoprotein	6
		1.44	GOTERM_MF_FAT	GO:0005506~iron ion binding	9
		1.41	GOTERM_MF_FAT	GO:0046872~metal ion binding	19
GcvB	60	6.65	GOTERM_BP_FAT	GO:0008652~cellular amino acid biosynthetic process	19
		4.12	GOTERM_BP_FAT	GO:0006865~amino acid transport	9
		2.78	GOTERM_MF_FAT	GO:0015171~amino acid transmembrane transporter activity	5
MicA	46	1.97	GOTERM_CC_FAT	GO:0009279~cell outer membrane	6
		1.12	GOTERM_BP_FAT	GO:0000271~polysaccharide biosynthetic process	6
MicF	50	2.36	GOTERM_CC_FAT	GO:0044462~external encapsulating structure part	7
		2.14	GOTERM_CC_FAT	GO:0030312~external encapsulating structure	16
		1.28	SP_PIR_KEYWORDS	Lipoprotein	5
RyhB	70	3.41	GOTERM_MF_FAT	GO:0005506~iron ion binding	13
		2.86	GOTERM_MF_FAT	GO:0046872~metal ion binding	22
		2.59	GOTERM_MF_FAT	GO:0051536~iron-sulfur cluster binding	9
SgrS	66	1.62	KEGG_PATHWAY	02060: phosphotransferase system (PTS)	5
		1.36	GOTERM_MF_FAT	GO:0046872~metal ion binding	17
		1.35	GOTERM_BP_FAT	GO:0051188~cofactor biosynthetic process	7
Spot42	85	2.96	GOTERM_BP_FAT	GO:0046356~acetyl-CoA catabolic process	7
		2.53	GOTERM_BP_FAT	GO:0006732~coenzyme metabolic process	12
		1.83	KEGG_PATHWAY	00020:Citrate cycle, tricarboxylic acid cycle (TCA cycle)	5
FsrA	54	4.77	GOTERM_MF_FAT	GO:0051536~iron-sulfur cluster binding	8
		3.81	GOTERM_BP_FAT	GO:0022900~electron transport chain	6
		3.69	UP_SEQ_FEATURE	domain:4Fe-4S ferredoxin-type 2	4
PrrF	103	4.47	GOTERM_MF_FAT	GO:0051536~iron-sulfur cluster binding	12
		4.88	GOTERM_MF_FAT	GO:0005506~iron ion binding	20
		3.81	SP_PIR_KEYWORDS	electron transport	7
SR1	50	1.88	GOTERM_BP_FAT	GO:0030435~sporulation resulting in formation of a cellular spore	8

The top 3 significantly enriched terms (DAVID enrichment score ≥1.1) for 11 tested sRNAs are shown. For each sRNA, the number of predicted targets with a P value ≤0.01 (column 2), the score of the enriched functional cluster (column 3), the name and source of a representative term of this cluster (columns 4 and 5), and the number of unique genes in this cluster (column 6) are given. Individual gene members of the enriched terms are given in Table S5.

to the existing experimental findings and add several plausible candidates.

The known functions of GcvB were predicted almost completely by CopraRNA and the subsequent functional enrichment. The top 15 predictions and functionally enriched target candidates are shown in Fig. 2*A*.

CopraRNA also returned the correct functional characterization for several other sRNAs. The predicted targets of MicA (Table 2 and *SI Appendix*, Fig. S7) and MicF were strongly enriched for outer membrane proteins, whereas the most strongly enriched cluster of RyhB targets consists of iron-binding proteins (Table 2 and Fig. 2 *B* and *C*).

Network Analysis of Predicted Targets. Certain genes serve as regulatory hubs and are targeted by several sRNAs. For example, the mRNA encoding the alternative sigma factor RpoS is targeted directly by at least three sRNAs, the Arc-associated sRNA Z (ArcZ), DsrA, and the RpoS regulator RNA (RprA) (1), whereas the *csgD* mRNA is regulated by five different sRNAs, i.e., GcvB (32), the multicellular adhesive sRNA (McaS) (32, 33), the OmpR-regulated sRNA A/B (OmrA/B) (34), and RprA (35). Computational target prediction by CopraRNA allows the analysis of a high number of sRNAs, and the results can be combined to infer the gene regulatory network for a given organism. Indeed, our global network analysis based on the benchmark dataset predicted known and potential hotspots of sRNA-based regulation. In total, 15 mRNAs were predicted to be targeted by four or more sRNAs and ~50 mRNAs by three or more sRNAs (Table S6). A striking example of an mRNA with multiple potential sRNA regulators encodes Lrp (leucine-responsive regulatory protein) and is predicted to be regulated by 7 of the 18 investigated sRNAs, including the previously identified regulators MicF (22, 36) and GcvB (4). The mRNA encoding the succinate dehydrogenase subunit SdhA has six predicted sRNA

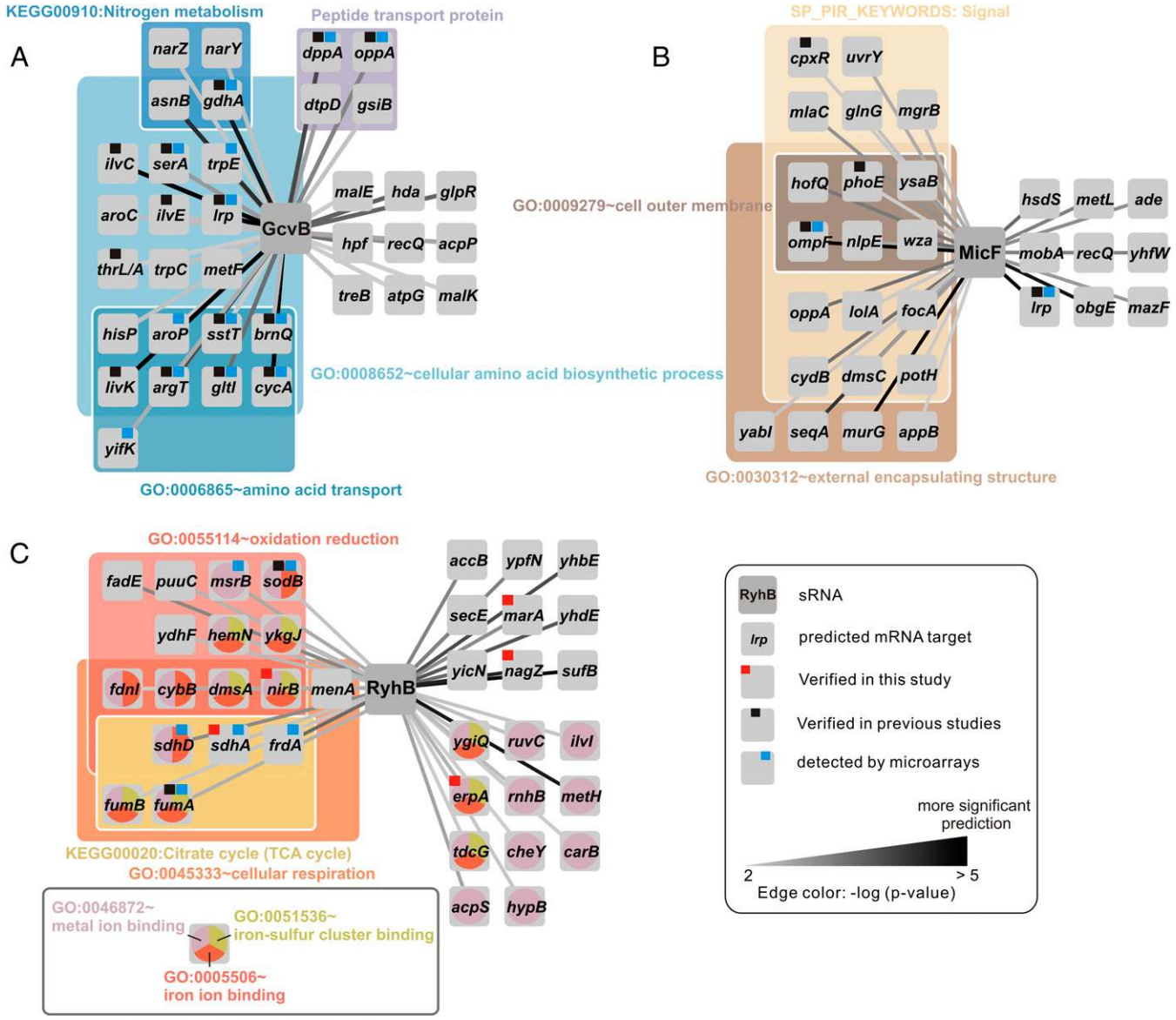

Fig. 2. Visualization of the functional enrichment analysis. All top 15 target predictions are shown plus predictions with a CopraRNA *P* value ≤0.01 that are functionally enriched (selected enriched terms). The edges connecting the sRNAs and targets are color coded according to the CopraRNA prediction *P* value, a darker color indicates a statistically more significant prediction. Previously experimentally verified targets from the literature [with regard to our benchmark list (*SI Appendix*, Table S1)] are marked with a black square, verifications from this study with a red square, and targets detected by microarrays with a blue square. Functionally enriched targets are color coded with respect to the enriched term. Results for (*A*) GcvB, (*B*) MicF, and (*C*) RyhB.

regulators, three of which were verified in this study (see below). We also detected multiple regulators of *csgD* and *rpoS* mRNAs. In addition to OmrA/B (34) and RprA (35), we predicted ChiX as a potential regulator of *csgD*. Another interesting example is the *yobF-cspC* dicistron with four potential regulators (CyaR, OmrA/B, and OxyS). From these, OxyS was previously shown to negatively regulate the *yobF-cspC* mRNA (10). The network obtained for 18 sRNAs and their previously verified and new targets is presented in Fig. 3*A*. In total, when using a *P* value threshold of 0.01, CopraRNA predicted 52 of the 101 benchmark targets. Furthermore, we verified 17 as yet unknown targets, uncovering connections between the regulatory networks of GcvB and Spot42, CyaR, RyhB and FnrS, and CyaR and SgrS. FnrS and RyhB share a dense overlapping regulon of at least four targets (Fig. 3*A*). Additionally, several operons were predicted to be influenced by multiple sRNAs: the *sdhCDAB-sucABCD* operon is targeted by five sRNAs at three different positions (Fig. 3*B*); Spot42 and RyhB each regulate two genes in the operon, *sdhC* (37) and *sucC*, as well as *sdhD* (37) and *sdhA*, respectively. In addition, the *iscRUAB* operon is regulated by both FnrS and RyhB (38) (Fig. 3*C*).

Experimental Verification of Predicted Targets. Based on the benchmark results, we restricted the final set of target candidates for each sRNA to the top 15 predictions plus candidates that belong to the functional-enriched terms (Table S5). This approach provides a reasonable balance between sensitivity and specificity because it uses the high positive predictive value in the topmost predictions (*SI Appendix*, Fig. S3*B*) while allowing investigation of an extended target set. We selected 23 previously uncharacterized potential targets (*SI Appendix*, Table S7) for experimental testing using a GFP reporter system tailored to investigate posttranscriptional regulation (22). We verified 17 additional targets, which equals a success rate of ∼74%, and exemplarily proved the predicted interaction sites of *yobF*-CyaR, *iscR*-FnrS, *nirB*-RyhB, and *gdhA*-Spot42 through the introduction of compensatory mutations and for *marA*-FnrS, *erpA*-RyhB, *marA*-RyhB, and *sucC*-Spot42 by point mutations in their respective 5′UTRs (Fig. 4 *A* and *B* and *SI Appendix*, Fig. S8). Interestingly, the point mutations in the *marA**[1] construct resulted in an increased repression by wild-type RyhB, which indicates an improved RNA–RNA hybrid formation. Posttranscriptional repression of the remaining predicted targets was tested by flow cytometry (Fig. 4*C*) or Western blots (*SI Appendix*, Fig. S9). An overview of the constructs used and the respective mean fluorescence intensities is given in *SI Appendix*, Figs. S9 and S10. Most of the predicted interactions resemble the classic binding proximal to the translational start site. However, the binding sites for Spot42 in *gdhA* and *icd* align with positions +80 and +75 downstream from the start codon, deeply within the

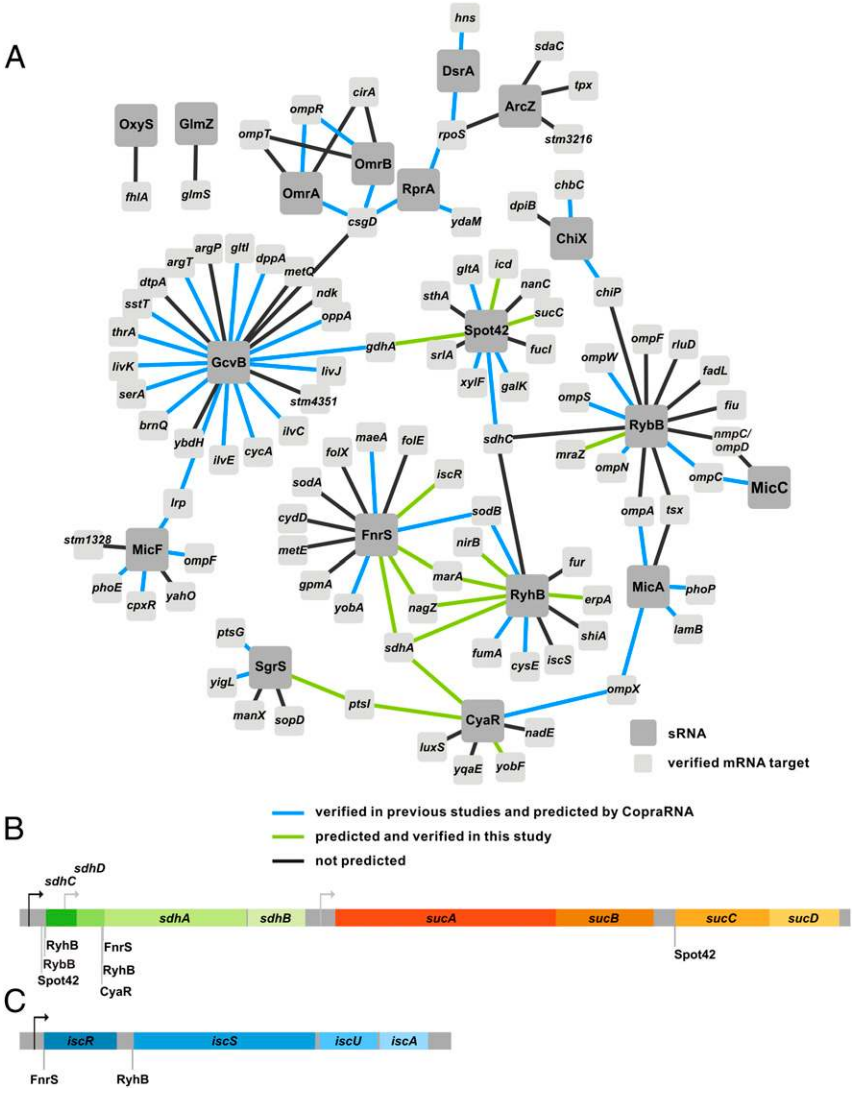

Fig. 3. (*A*) Network of verified targets for the 18 sRNAs of the benchmark dataset. Visualization of the (*B*) *sdhCDABsucABCD* and (*C*) *iscRSUAB* operon with verified interaction sites; the promoters are annotated according to EcoCyc (52).

coding region. A direct inhibition of translation seems unlikely for these targets; rather, we assume a mechanism that reduces the half-life of the mRNAs, as shown for the *ompD*–MicC interaction in *S. enterica* (39, 40).

Performance of CopraRNA for sRNAs from Nonenterobacterial Species. To evaluate the performance of CopraRNA for sRNAs that are not conserved in *E. coli* or *S. enterica*, we extended our benchmark dataset by five additional sRNAs from a wide range of bacterial families and phyla—the Fur-regulated sRNA A (FsrA) and SR1 (Firmicutes, Bacillaceae), LhrA (Firmicutes, Listeriaceae), the inhibitor of hctA translation (IhtA) (Chlamydiae), and PrrF (Proteobacteria, Pseudomonadaceae)—with a total of 17 experimentally verified targets (*SI Appendix*, Table S8). CopraRNA detects 11 of the 17 verified targets in the top 35 predictions, which resembles a true positive rate of ∼65% and a PPV of ∼6.3%. Again, this is at least ∼3.7 times better than the single-organism–specific methods (*SI Appendix*, Fig. S11). We also obtained intriguing functional enrichments for FsrA and PrrF (Table 2 and Table S5). The topmost enriched term for the predicted FsrA and PrrF targets is "GO:0051536∼iron-sulfur cluster binding" followed by other iron-related terms. This is in agreement with the known roles of these sRNAs in the iron stress response (30) and may hint at additional yet-unknown target genes of those sRNAs. The complete prediction dataset is given in Table S9.

Discussion

Comparison with Other Target Identification Strategies. In this study, we present a comparative method for sRNA target identification in bacteria. The method is superior to existing bioinformatics tools (Fig. 1*B*) and works for a wide range of bacterial organisms. For seven tested benchmark sRNAs, CopraRNA can compete with microarray-based experiments for target detection (Table 1). CopraRNA is available as an easy-to-use Web interface (http://rna.informatik.uni-freiburg.de/CopraRNA/). True positive predictions are enriched by the downstream refinement of the prediction results through integration of existing data.

Using CopraRNA, we detected 17 as yet unknown targets for six sRNAs (Fig. 4 and *SI Appendix*, Fig S9). For the sRNAs

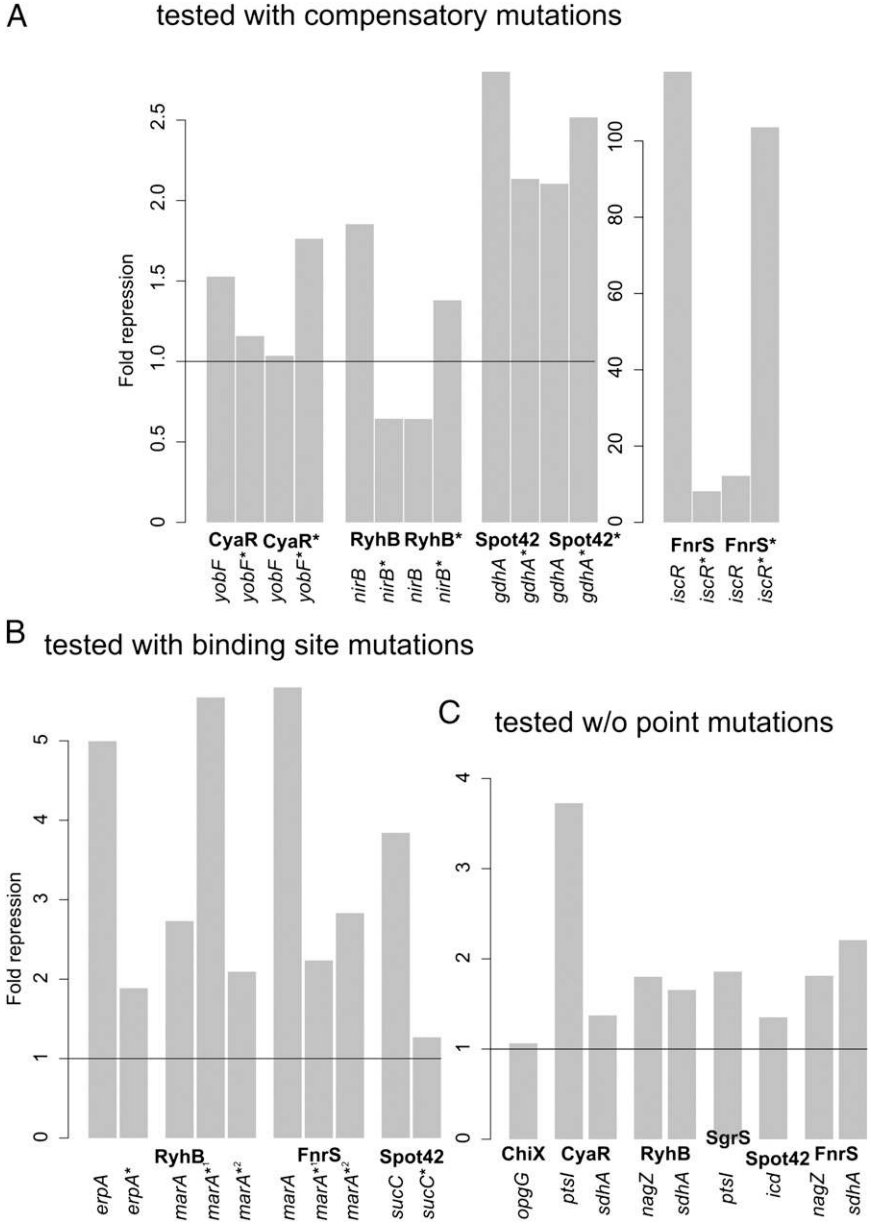

Fig. 4. Verification of sRNA target candidates. Translational repression of 5′ UTR–*gfp* fusions when overexpressing the sRNA. The fold repression is the ratio of the GFP fluorescence of the respective translational 5′ UTR–GFP fusion in the presence of the control plasmid pJV300 and a plasmid for the overexpression of the respective sRNA, after subtraction of the background fluorescence. Compensatory point mutations in the UTR and sRNA are indicated with an asterisk. (*A*) Verification of the *yobF*–CyaR, *nirB*–RyhB, *gdhA*–Spot42, and *iscR*–FnrS interactions with compensatory point mutations. (*B*) Verification of the *erpA*–RyhB, *marA*–RyhB, *marA*–FnrS, and *sucC*–Spot42 interactions with point mutations in the 5′UTR. (*C*) Verification of the *ptsI*–CyaR, *sdhA*–CyaR, *nagZ*–RyhB, *sdhA*–RyhB, *ptsI*–SgrS, *icd*–Spot42, *nagZ*–FnrS, and *sdhA*–FnrS interactions without point mutations.

FnrS, FsrA, GcvB, MicA, MicF, PrrF, RyhB, SgrS, and Spot42, bona fide physiological functions could be predicted accurately on our in silico results (Table 2). Compared with microarrays, CopraRNA has an advantage in that genetic modifications and time-consuming, expensive wet-laboratory experiments are not required for initial target screening. Additionally, CopraRNA is not biased by secondary effects, which might be picked up by experimental screening, and allows detection of targets not expressed under the tested conditions. Consequently, the predicted targets verify but also extend the existing microarray data.

However, CopraRNA also comes with certain limitations. The primary limitation of bioinformatic target prediction methods is that most predictions correspond to false positive predictions. The comparative approach of CopraRNA reduces this problem to the extent that further experimental analysis becomes much more reasonable than with existing tools, but it does not solve this problem completely. In our benchmark assay, half of the 101 known targets are detected with a P value threshold of 0.01 (*SI Appendix*, Fig S3*A*). At this threshold, an average of 65 targets is predicted for each sRNA and the FPR is ∼95% (*SI Appendix*, Fig S3*B*). Thus, a reasonable sensitivity of 50% comes with a low specificity of 5%. In fact, this is a strong improvement, as the other tools tested reach a maximum sensitivity of 25% (IntaRNA) at 65 predictions per sRNA, and e.g., IntaRNA needs 226 predictions per sRNA to reach a sensitivity of 50%. Nevertheless, a low specificity challenges investigators to follow up on the predictions. For that reason, we do not stick to the P value threshold strictly, but focus on the top 15 list and on the predictions ($P \leq 0.01$) suggested by further postprocessing steps. These steps may include automatic and manual functional enrichment (Fig. 2), network analysis (Fig. 3), overlaps with transcription factor regulons (Fig. 5 and *SI Appendix*, Fig S13), or correlation patterns coming from microarray data (41, 42). This combined strategy was very successful in retaining sensitivity while enhancing specificity. We demonstrated this by the experimental verification of 73% of the selected 23 predicted targets that were not characterized previously. These results also show that the FPR is at least slightly overestimated because of previously unknown targets (*SI Appendix*, Fig S3*B*; compare dashed and solid blue lines). Another challenge is a prediction without a meaningful postprocessing result, caused, e.g., by the lack of additional data or lower prediction quality. For these cases, we control the FDR statistically by calculating a q-value. The average q-value at prediction rank 65 is ∼0.54 and therefore judged by the current benchmark data, rather too optimistic. Nevertheless, the q-value distribution is valuable to roughly estimate the general prediction quality for a given sRNA. For example, we could not predict known targets for ArcZ. This less informative prediction is accompanied correctly by a rapidly growing q-value and only 10 predictions with $q \leq 0.5$. On the other side, the good prediction for GcvB has 38 predictions with $q \leq 0.5$, and as described above, the q-value fits well to the benchmark dataset. CopraRNA generally requires the conservation of an sRNA and also a substantial level of target conservation in the selected species. Therefore, single-organism–specific targets are likely to be missed, as are interactions that generally are not predictable by the underlying IntaRNA algorithm (e.g., double-kissing hairpin complexes). For example, the *metE*–FnrS interaction [verified in *E. coli* (20)] seems to be conserved or detectable only in three of the eight included species (*SI Appendix*, Fig. S12). This results in a high combined P value of 0.54 and a rank of 1,969 in the combined prediction and shows the importance of carefully selecting species. A small evolutionary distance favors sensitivity, and a large distance favors specificity. The downstream

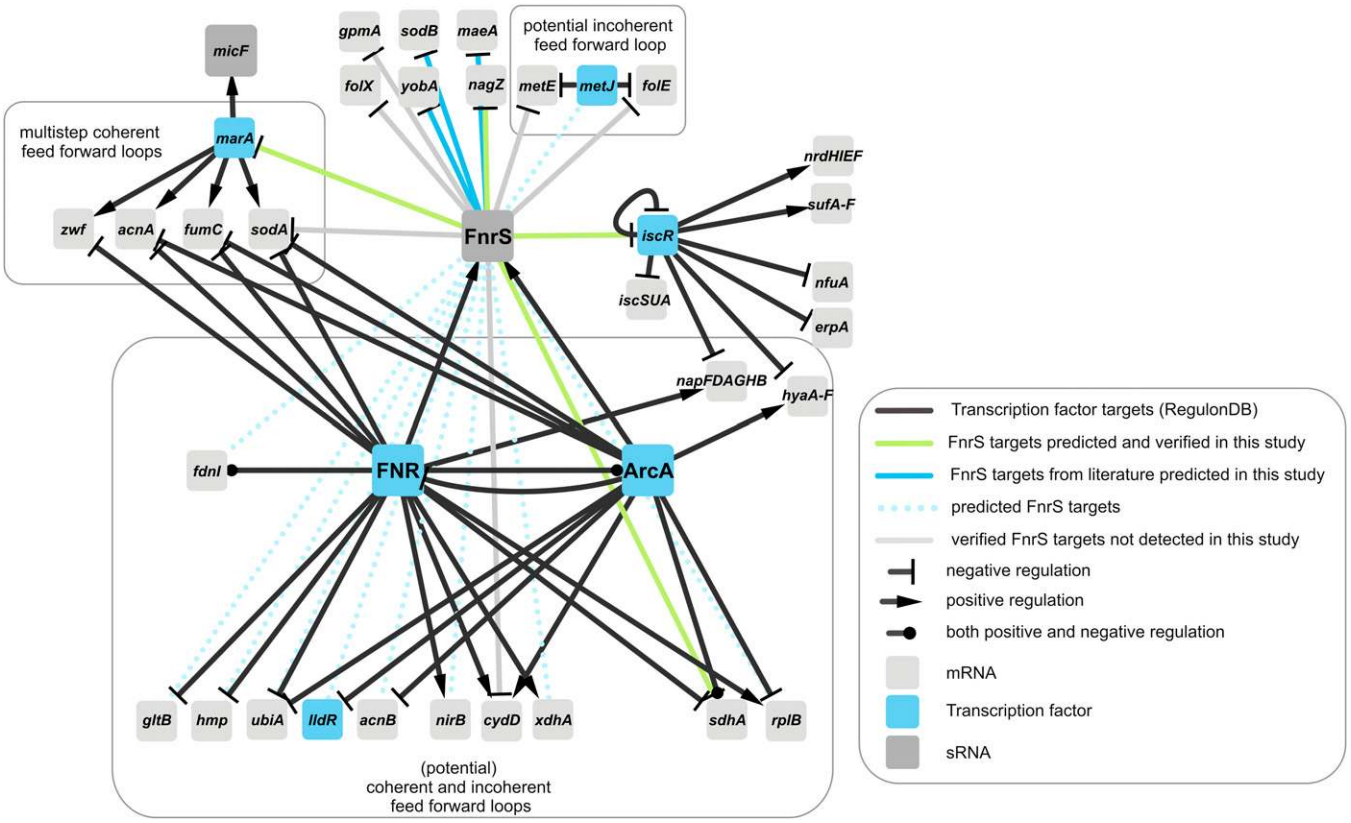

Fig. 5. Partial regulatory network around FNR, ArcA, and FnrS. The figure shows verified FnrS targets, as well as predicted targets (CopraRNA P value ≤0.01) regulated by FNR or ArcA. For the transcription factors, only selected targets are displayed.

functional enrichment analysis relies on the availability of the organism in the DAVID database (31), and the results depend on the annotation quality of the genome of interest. Of note, CopraRNA is a target prediction tool for sRNAs that are expected to act in *trans*; it is not suitable for the differentiation of a *trans*-acting RNA from other types of transcripts. However, the functional enrichment analysis, the conservation plots, and the q-value distribution provided by CopraRNA might provide a hint as to whether a given conserved RNA is a functional *trans*-acting sRNA.

Additional Targets and Functions of Previously Characterized sRNAs. The inspection of the benchmark dataset revealed additional targets and functions, even for sRNAs extensively characterized in the past. For the cAMP receptor protein (CRP)-regulated sRNA CyaR (18, 19), we detected as yet unidentified targets in primary metabolism (*sdhA*) and the phosphotransferase system (*ptsI*), constituting previously unreported links of the CyaR regulon to carbon metabolism. Furthermore, with regard to the *yobF-cspC* operon, we found a potential explanation for the indirect negative effect of CyaR on the *rpoS* mRNA, which was detected in a screen with 26 sRNAs (43). The *yobF* gene is organized together with *cspC* in a dicistronic operon, and the RNA chaperone CspC is a posttranscriptional stabilizer of the *rpoS* message (44).

FnrS is involved in gene regulation after the shift from aerobic to anaerobic conditions, and its expression is activated by the transcription factors FNR and ArcA (20, 21). The combination of existing information (45) with our predictions and verifications for FnrS results in a remarkable complex regulatory network (Fig. 5): (*i*) FnrS transduces the signal to several non-FNR and -ArcA targets. These include the target *nagZ* and the two transcription factor mRNAs *iscR* and *marA*. (*ii*) The prediction also revealed several target candidates, which are controlled simultaneously by FNR and ArcA, which would establish multi-output feed-forward loops. Although the transcription factor MarA is not directly regulated by FNR or ArcA, four genes that are activated by MarA (*acnA*, *fumC*, *sodA*, *zwf*) are repressed by ArcA and/or FNR. These four genes are involved in the resistance to superoxide (46) and provide a reasonable explanation for the repression of *marA* by FnrS at anaerobic conditions. The repression of the transcription factor IscR may be part of the observed O_2-dependent expression of the *iscR* regulon (47).

FnrS shares three targets with RyhB. Both sRNAs regulate the mRNA encoding MarA, which is involved in the response to antimicrobial compounds and oxidative stress (46), and of the mRNA for the β-*N*-acetylglucosaminidase NagZ, which permits resistance to β-lactams in *Pseudomonas aeruginosa* (48). Interestingly, both MarA and NagZ are not obviously involved in iron homeostasis. For the iron stress-induced sRNA RyhB, we predicted mRNAs for 13 iron-containing proteins as targets and verified the posttranscriptional regulation of *erpA*, the mRNA of an A-type carrier (ATC) protein involved in iron–sulfur cluster biogenesis (49), and of *nirB*, which codes for a subunit of nitrite reductase.

Regarding the dual-function RNA SgrS, we predicted interactions with mRNAs of additional components of the phosphotransferase system (*chhB, cmtB* and *fruA*) and verified the posttranscriptional regulation of *ptsI* (Fig. 4), which codes for the non–sugar-specific enzyme I component of the PTS. Furthermore, we detected the recently described positive regulated sugar phosphatase mRNA *yigL* (50) as a direct target.

We also predicted and verified targets for the CRP-repressed Spot42 sRNA which is involved in catabolite repression and controls a range of genes in central and secondary metabolism and sugar transport (6). Our predictions show a large, 18-gene overlap with the CRP regulon and point to an even broader regulatory role for Spot42 in primary metabolism involving the citrate cycle and acetyl-CoA–dependent processes (Table 2, Tables S4 and S5, and *SI Appendix*, Fig. S13). Our successful experimental validation of the targets *gdhA*, *icd*, and *sucC* proves the accuracy of our predictions.

In sum, CopraRNA allows for an efficient screening of large numbers of sRNAs and has proven superior compared with existing methods. Using this tool, we obtained compelling evidence that sRNAs are global regulators of large sets of mRNAs, comparable to protein transcription factors and eukaryotic microRNAs. We also show that it is a common concept that mRNAs are targeted by multiple sRNAs and correctly predicted the regulatory hubs *csgD* and *rpoS*. Furthermore, we proposed and partially verified *gdhA*, *lrp*, *marA*, *nagZ*, *ptsI*, *sdhA*, and *yobF-cspC* as hubs targeted by up to seven different sRNAs. Finally, we present examples for complex posttranscriptional events at the operon level, including multiple targeting by the same, as well as different, sRNAs.

Methods

Experimental Methods. ***Bacterial strains and growth.*** Cells were grown in Luria–Bertani (LB) broth or on LB plates at 37 °C. Antibiotics (where appropriate) were applied at the following concentrations: 100 $mg{\cdot}mL^{-1}$ ampicillin and 25 $mg{\cdot}mL^{-1}$ chloramphenicol.

Plasmid construction. The plasmids for the overexpression of FnrS and CyaR and those for the translational superfolder–GFP fusions were constructed as described previously (22).

Oligonucleotides and plasmids. Oligonucleotides and plasmids are listed in *SI Appendix*, Tables S10 and S11.

Fluorescence measurements. Overnight cultures were used to inoculate (1:100) fresh cultures, and cultivation was continued to OD_{600} = 2.0. Culture samples equivalent to 1 OD were harvested by centrifugation and resuspended in PBS. Aliquots of 100 μL were transferred to a 96-well microtiter plate, and relative GFP levels were measured in a Victor3 fluorimeter (Perkin-Elmer). A wild-type strain was measured in parallel to subtract autofluorescence levels. All samples were measured in biological triplicates. This method was used to analyze the RyhB–*nirB* and the CyaR–*yobF* interactions.

Flow cytometry-based fluorescence measurements. Single bacterial colonies were inoculated in 200 μL LB medium in 96-well microtiter plates containing ampicillin and chloramphenicol and grown at 37 °C, 100 rpm for 12–15 h. Cells were diluted 1/5 in LB and fixed with formaldehyde (Roti-Histofix 10%; Carl Roth GmbH) to an final concentration of 1% (wt/vol) and measured directly on an Accuri C6 flow cytometer (BD Biosciences). The mean fluorescence of 50,000 events was averaged for 6–12 independent biological replicates. The fold repression was calculated as the ratio of the mean GFP fluorescence of the respective translational UTR–GFP fusion in the presence of the control plasmid pJV300 and a plasmid for the overexpression of the respective sRNA, after subtraction of the background fluorescence. Background fluorescence was measured with the control plasmids pXG-0 and pJV300 (22):

$$\mathrm{Fold}_{\mathrm{rep}} = \frac{\mathrm{Fluorescence\ UTR}_{\mathrm{pJV300}} - \mathrm{pXG-0}_{\mathrm{pJV300}}}{\mathrm{Fluorescence\ UTR}_{\mathrm{sRNA}} - \mathrm{pXG-0}_{\mathrm{pJV300}}}.$$

The respective mean fluorescences after subtraction of the background fluorescence are shown in *SI Appendix*, Fig. S9. Western blots were performed as described in ref. 9.

Theoretical Methods. ***Benchmark analysis.*** For the benchmark analysis, we conducted whole-genome target predictions for *E. coli* (NC_000913) and *S. enterica* (NC_003197, NC_003277) based on the sequences 200 nt upstream and 100 nt downstream of the annotated start codons as the input (the first nucleotide of the start codon corresponds to position 201). The Web server of RNApredator used the whole gene for target prediction. Otherwise, all the tools were used with the given standard parameters. The *P* value threshold of TargetRNA was set to 0.99 to obtain the top 100 predictions. The benchmark dataset included 18 sRNAs and a total of 101 previously published targets (*SI Appendix*, Table S1). Some targets were verified in both *E. coli* and *S. enterica*; the total number of verified sRNA–target pairs is 113, but we used only the nonredundant dataset. We included only targets for which a direct posttranscriptional regulation by an sRNA was verified experimentally. Targets detected only by RT-PCR, microarrays, or Northern blots and not verified further were excluded.

Functional enrichment. Functional enrichments (functional annotation clustering) were performed on the DAVID Web server (31) for all benchmark sRNA predictions. For each sRNA, the target candidates ($P \leq 0.01$) were tested against all the genes on the list as background. Obvious artifacts,

i.e., predicted interactions with the complementary strand of the genomic coding region of the respective sRNA, were excluded. Enrichments were performed for *E. coli*. The standard parameters were changed to a "Similarity Threshold" of 0.85 and an "Initial Group Membership" and "Final Group Membership" of 2. Our threshold for a functional-enriched term was a DAVID enrichment score of ≥1.1. Networks were visualized using Cytoscape (51).

CopraRNA algorithm. To reduce the number of false positive hits in the interaction predictions, we searched for interactions that are conserved in various species. However, for several reasons, it is conceivable that the interaction is preserved whereas the actual interaction site is not. To be able to still predict conserved interactions, it is necessary to combine the evidence for interactions in the different species without resorting to a consensus-based approach. In addition to the Web server version, a stand-alone version of CopraRNA is available (www.bioinf.uni-freiburg.de/Software/). A more detailed description of CopraRNA, with a focus on the calculation of *P* values, may be found in *SI Appendix*.

ACKNOWLEDGMENTS. This work was supported by the Deutsche Forschungsgemeinschaft Focus Program "Sensory and Regulatory RNAs in Prokaryotes" (SPP1258); Bundesministerium für Bildung und Forschung (BMBF) Grant 0316165 (to W.R.H, R.B., and J.V.); and the Excellence Initiative of the German Federal and State Governments (EXC 294 to R.B.). K.P. was supported by a postdoctoral fellowship from the Human Frontiers in Science Program.

1. Storz G, Vogel J, Wassarman KM (2011) Regulation by small RNAs in bacteria: Expanding frontiers. *Mol Cell* 43(6):880–891.
2. Gottesman S, Storz G (2011) Bacterial small RNA regulators: Versatile roles and rapidly evolving variations. *Cold Spring Harb Perspect Biol* 3(12):a003798.
3. Papenfort K, Vogel J (2010) Regulatory RNA in bacterial pathogens. *Cell Host Microbe* 8(1):116–127.
4. Sharma CM, et al. (2011) Pervasive post-transcriptional control of genes involved in amino acid metabolism by the Hfq-dependent GcvB small RNA. *Mol Microbiol* 81(5): 1144–1165.
5. Gogol EB, Rhodius VA, Papenfort K, Vogel J, Gross CA (2011) Small RNAs endow a transcriptional activator with essential repressor functions for single-tier control of a global stress regulon. *Proc Natl Acad Sci USA* 108(31):12875–12880.
6. Beisel CL, Storz G (2011) The base-pairing RNA spot 42 participates in a multioutput feedforward loop to help enact catabolite repression in *Escherichia coli*. *Mol Cell* 41(3):286–297.
7. Sharma CM, et al. (2010) The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature* 464(7286):250–255.
8. Mitschke J, et al. (2011) An experimentally anchored map of transcriptional start sites in the model cyanobacterium *Synechocystis* sp. PCC6803. *Proc Natl Acad Sci USA* 108(5):2124–2129.
9. Papenfort K, Bouvier M, Mika F, Sharma CM, Vogel J (2010) Evidence for an autonomous 5′ target recognition domain in an Hfq-associated small RNA. *Proc Natl Acad Sci USA* 107(47):20435–20440.
10. Tjaden B, et al. (2006) Target prediction for small, noncoding RNAs in bacteria. *Nucleic Acids Res* 34(9):2791–2802.
11. Eggenhofer F, Tafer H, Stadler PF, Hofacker IL (2011) RNApredator: Fast accessibility-based prediction of sRNA targets. *Nucleic Acids Res* 39(Web Server issue):W149–W154.
12. Busch A, Richter AS, Backofen R (2008) IntaRNA: Efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics* 24(24): 2849–2856.
13. Rehmsmeier M, Steffen P, Höchsmann M, Giegerich R (2004) Fast and effective prediction of microRNA/target duplexes. *RNA* 10(10):1507–1517.
14. Hao Y, et al. (2011) Quantifying the sequence-function relation in gene silencing by bacterial small RNAs. *Proc Natl Acad Sci USA* 108(30):12473–12478.
15. Backofen R, Hess WR (2010) Computational prediction of sRNAs and their targets in bacteria. *RNA Biol* 7(1):33–42.
16. Richter AS, Backofen R (2012) Accessibility and conservation: General features of bacterial small RNA-mRNA interactions? *RNA Biol* 9(7):954–965.
17. Beisel CL, Updegrove TB, Janson BJ, Storz G (2012) Multiple factors dictate target selection by Hfq-binding small RNAs. *EMBO J* 31(8):1961–1974.
18. De Lay N, Gottesman S (2009) The Crp-activated small noncoding regulatory RNA CyaR (RyeE) links nutritional status to group behavior. *J Bacteriol* 191(2):461–476.
19. Papenfort K, et al. (2008) Systematic deletion of *Salmonella* small RNA genes identifies CyaR, a conserved CRP-dependent riboregulator of OmpX synthesis. *Mol Microbiol* 68(4):890–906.
20. Boysen A, Møller-Jensen J, Kallipolitis B, Valentin-Hansen P, Overgaard M (2010) Translational regulation of gene expression by an anaerobically induced small non-coding RNA in *Escherichia coli*. *J Biol Chem* 285(14):10690–10702.
21. Durand S, Storz G (2010) Reprogramming of anaerobic metabolism by the FnrS small RNA. *Mol Microbiol* 75(5):1215–1231.
22. Corcoran CP, et al. (2012) Superfolder GFP reporters validate diverse new mRNA targets of the classic porin regulator, MicF RNA. *Mol Microbiol* 84(3):428–445.
23. Massé E, Vanderpool CK, Gottesman S (2005) Effect of RyhB small RNA on global iron use in *Escherichia coli*. *J Bacteriol* 187(20):6962–6971.
24. Papenfort K, Podkaminski D, Hinton JCD, Vogel J (2012) The ancestral SgrS RNA discriminates horizontally acquired *Salmonella* mRNAs through a single G-U wobble pair. *Proc Natl Acad Sci USA* 109(13):E757–E764.
25. Uchiyama I (2007) MBGD: A platform for microbial comparative genomics based on the automated construction of orthologous groups. *Nucleic Acids Res* 35(Database issue):D343–D346.
26. Hartung J (1999) A note on combining dependent tests of significance. *Biom J* 41:849–855.
27. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* 100(16):9440–9445.
28. Bouvier M, Sharma CM, Mika F, Nierhaus KH, Vogel J (2008) Small RNA binding to 5′ mRNA coding region inhibits translational initiation. *Mol Cell* 32(6):827–837.
29. Argaman L, et al. (2001) Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. *Curr Biol* 11(12):941–950.
30. Salvail H, Massé E (2012) Regulating iron storage and metabolism with RNA: An overview of posttranscriptional controls of intracellular iron homeostasis. *Wiley Interdiscip Rev RNA* 3(1):26–36.
31. Huang W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4(1):44–57.
32. Jørgensen MG, et al. (2012) Small regulatory RNAs control the multi-cellular adhesive lifestyle of *Escherichia coli*. *Mol Microbiol* 84(1):36–50.
33. Thomason MK, Fontaine F, De Lay N, Storz G (2012) A small RNA that regulates motility and biofilm formation in response to changes in nutrient availability in *Escherichia coli*. *Mol Microbiol* 84(1):17–35.
34. Holmqvist E, et al. (2010) Two antisense RNAs target the transcriptional regulator CsgD to inhibit curli synthesis. *EMBO J* 29(11):1840–1850.
35. Mika F, et al. (2012) Targeting of *csgD* by the small regulatory RNA RprA links stationary phase, biofilm formation and cell envelope stress in *Escherichia coli*. *Mol Microbiol* 84(1):51–65.
36. Holmqvist E, Unoson C, Reimegård J, Wagner EGH (2012) A mixed double negative feedback loop between the sRNA MicF and the global regulator Lrp. *Mol Microbiol* 84(3):414–427.
37. Desnoyers G, Massé E (2012) Noncanonical repression of translation initiation through small RNA recruitment of the RNA chaperone Hfq. *Genes Dev* 26(7):726–739.
38. Desnoyers G, Morissette A, Prévost K, Massé E (2009) Small RNA-induced differential degradation of the polycistronic mRNA *iscRSUA*. *EMBO J* 28(11):1551–1561.
39. Pfeiffer V, Papenfort K, Lucchini S, Hinton JCD, Vogel J (2009) Coding sequence targeting by MicC RNA reveals bacterial mRNA silencing downstream of translational initiation. *Nat Struct Mol Biol* 16(8):840–846.
40. Bandyra KJ, et al. (2012) The seed region of a small RNA drives the controlled destruction of the target mRNA by the endoribonuclease RNase E. *Mol Cell* 47(6): 943–953.
41. Hernández-Prieto MA, et al. (2012) Iron deprivation in *Synechocystis*: Inference of pathways, non-coding RNAs, and regulatory elements from comprehensive expression profiling. *G3 (Bethesda)* 2(12):1475–1495.
42. Modi SR, Camacho DM, Kohanski MA, Walker GC, Collins JJ (2011) Functional characterization of bacterial sRNAs using a network biology approach. *Proc Natl Acad Sci USA* 108(37):15522–15527.
43. Mandin P, Gottesman S (2010) Integrating anaerobic/aerobic sensing and the general stress response through the ArcZ small RNA. *EMBO J* 29(18):3094–3107.
44. Cohen-Or I, Shenhar Y, Biran D, Ron EZ (2010) CspC regulates *rpoS* transcript levels and complements *hfq* deletions. *Res Microbiol* 161(8):694–700.
45. 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.
46. Martin RG, Rosner JL (2011) Promoter discrimination at class I MarA regulon promoters mediated by glutamic acid 89 of the MarA transcriptional activator of *Escherichia coli*. *J Bacteriol* 193(2):506–515.
47. Giel JL, Rodionov D, Liu M, Blattner FR, Kiley PJ (2006) IscR-dependent gene expression links iron-sulphur cluster assembly to the control of O_2-regulated genes in *Escherichia coli*. *Mol Microbiol* 60(4):1058–1075.
48. Zamorano L, et al. (2010) NagZ inactivation prevents and reverts β-lactam resistance, driven by AmpD and PBP 4 mutations, in *Pseudomonas aeruginosa*. *Antimicrob Agents Chemother* 54(9):3557–3563.
49. Pinske C, Sawers RG (2012) A-type carrier protein ErpA is essential for formation of an active formate-nitrate respiratory pathway in *Escherichia coli* K-12. *J Bacteriol* 194(2): 346–353.
50. Papenfort K, Sun Y, Miyakoshi M, Vanderpool CK, Vogel J (2013) Small RNA-mediated activation of sugar phosphatase mRNA regulates glucose homeostasis. *Cell* 153(2): 426–437.
51. Cline MS, et al. (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc* 2(10):2366–2382.
52. Keseler IM, et al. (2013) EcoCyc: Fusing model organism databases with systems biology. *Nucleic Acids Res* 41(Database issue):D605–D612.

Comparative genomics boosts target prediction for bacterial small RNAs

Patrick R. Wright[a,b], Andreas S. Richter[b], Kai Papenfort[c,d], Martin Mann[b], Jörg Vogel[c], Wolfgang R. Hess[a,e], Rolf Backofen[b,d,f,g,1] and Jens Georg[a,1]

[a]Genetics and Experimental Bioinformatics, Faculty of Biology, University of Freiburg, Schänzlestr. 1, D-79104 Freiburg, Germany;

[b]Bioinformatics Group, Department of Computer Science, University of Freiburg, Georges-Köhler-Al 106, D-79110 Freiburg, Germany;

[c]Institute for Molecular Infection Biology, University of Würzburg, Josef-Schneider-Str. 2/D15,D-97080 Würzburg, Germany;

[d]Department of Molecular Biology, Princeton University, Washington Road, 08544 Princeton, NJ, USA

[e]Centre for Biological Systems Analysis (ZBSA), University of Freiburg, Habsburgerstr. 49, D-79104 Freiburg, Germany;

[f]BIOSS Centre for Biological Signalling Studies, University of Freiburg, Schänzlestr. 18, D-79104 Freiburg, Germany;

[g]Center for non-coding RNA in Technology and Health, University of Copenhagen, Grønnegårdsvej 3, DK-1870 Frederiksberg C, Denmark

[1]To whom correspondence should be addressed. E-mail: jens.georg@biologie.uni-freiburg.de, backofen@informatik.uni-freiburg.de

Table of contents:

Definition of statistical terms

Positives: Positives are in our case all experimentally verified benchmark targets (given in **Table S1**)**.**

Positive predictions (candidates): Positive predictions are all targets that match the respective threshold criterion (e.g. p-value ≤ 0.01 or a given prediction rank). These targets are actually predicted by CopraRNA to be a target. The positive predictions consist of true positive and false positive predictions.

True positives: True positives are all experimentally verified targets (with regard to our benchmark dataset in **Table S1**) within the positive predictions.

False positives: False positives are all positive predictions that are no real targets, i.e. in our case that have not been experimentally verified.

True negatives: True negatives are all genes that do not match the respective threshold criterion (i.e. they are not predicted by CopraRNA) and are actually no real targets.

False negatives: False negatives are those positives that are not detected by CopraRNA.

True positive rate (sensitivity): The true positive rate defines how many real targets (positives) are found in the positive predictions, it is calculated by the following formula: Sensitivity $=\frac{\#\ True\ positives}{\#\ Positives}$.

Positive predictive value (PPV): The proportion of positive predictions that are true positives. The PPV is calculated by the following formula: $PPV = \frac{\#\ \text{True Positives}}{\#\ \text{Positive Predictions}}$.

Theoretical methods

Benchmark analysis

For the benchmark analysis, we conducted whole-genome target predictions for *Escherichia coli* (*E. coli*, NC_000913) and *Salmonella enterica* (*S. enterica*, NC_003197, NC_003277) using our new method CopraRNA and the web-based tools IntaRNA (1), RNApredator (2) and TargetRNA (3). For CopraRNA, IntaRNA and TargetRNA, we used the sequences 200 nt upstream and 100 nt downstream of the annotated start codons as the input (the first nucleotide of the start codon corresponds to position 201). The webserver of RNApredator did not allow a specification of the input UTR sequence. Otherwise, all the tools were used with the given standard parameters. The p-value threshold of TargetRNA was set to 0.99 to obtain the top 100 predictions. The benchmark dataset included 18 sRNAs and a total of 101 previously published targets (**Table S1**). Some targets were verified both in *E. coli* and *Salmonella*; the total number of verified sRNA-target pairs is 113, but

we used only the non-redundant dataset. We only included targets for which a direct post-transcriptional regulation by an sRNA was experimentally verified. Targets detected only by RT-PCR, microarrays or northern blots and not further verified were excluded.

Functional enrichment

Functional enrichments (Functional Annotation Clustering) were performed on the DAVID webserver (4) for all the benchmark sRNA predictions. For each sRNA, the target candidates ($p \leq 0.01$) were tested against all the genes in the list as background. Obvious artifacts, i.e., predicted interactions with the complementary strand of the genomic coding region of the respective sRNA were excluded. Enrichments were performed for *E. coli*. The standard parameters were changed to a "Similarity Threshold" of 0.85 and an "Initial Group Membership" and "Final Group Membership" of 2. Our threshold for a functional-enriched term was an EASE score of ≥1.1.

Networks

Networks were visualized using Cytoscape (5).

CopraRNA algorithm

To reduce the number of false-positive hits in the interaction predictions, we searched for interactions that are conserved in various species. One approach, as used in Petcofold (6) and ripalign (7), is to perform a combined consensus prediction for all the species together, thereby predicting the conserved interaction sites. However, due to several reasons, it is conceivable that the interaction is preserved while the actual interaction site is not. To be able to still predict conserved interactions, it is necessary to combine the evidence for interactions in the different species without resorting to a consensus-based approach.

Determining organism-specific p-values

The standard approach to quantify the evidence for an interaction prediction is to determine its significance, i.e., to predict p-values for the probability of finding an interaction in random sequences with a score greater than or equal to the score of the observed interaction. This makes the scores for

different organisms comparable, particularly if they have vastly different GC-contents. If only energy scores were to be combined, organisms with higher GC-contents will be weighted inappropriately stronger compared to organisms with low GC-contents due to the stronger binding of GC-rich duplexes.

We tested two different methods to deduce p-values. As the p-value for a single CopraRNA energy score describes the probability of a score of this quality being acquired by chance in front of a background model, it appears sensible to create a background model by shuffling all the putative target sequences to attain a dataset of random interactions. For this purpose, we executed IntaRNA predictions for 92 sRNA/species pairs. The target sequences were shuffled 10 times each while maintaining the di-nucleotide frequencies. The shuffling was performed with the shuffle program from Shawn Eddy's SQUID package (http://selab.janelia.org/software.html).

From the similarity of interaction prediction with local sequence alignment, it was already concluded by Rehmsmeier et al. (2004) (8) that interaction scores follow an extreme value distribution. Hence, one possibility we explored was to estimate the scale and location parameters for an extreme value distribution (Gumbel distribution) from the above-described sampled background data. We are, however, more interested in the tail of the distribution that is associated with significant interactions. Thus, we finally chose the generalized extreme value distribution over the Gumbel distribution, which has an additional shape parameter governing the tail behavior. As a second method, we directly used the empiric p-values determined from the above sampling.

This sampling method has the disadvantage of high computational complexity and would have to be redone for every new genome. For that reason, we also considered the possibility of approximating p-values from the un-shuffled data, i.e., the scores from the whole-genome interaction predictions. Although these un-shuffled scores do not give a correct background model because they also contain true positives, we were surprised to find that the approximated p-values work in practice. In **Fig. S14**, we display the fits of the extreme value distribution to the shuffled and un-shuffled data for GcvB.

This is an extreme case because GcvB has many targets (9), which amounts to many true positive predictions in the un-shuffled data. Nevertheless, the distributions are similar enough such that the error can be compensated by our combination method described below.

Combining p-values from different species

To combine evidence from different species, we employed a pair of a non-coding RNA r_1 and an mRNA r_2 and all the homologs of both RNAs in N species. We then calculated the score for the interaction of the associated homologs of r_1 and r_2 in each species. Using the previously described approach, we have a test for the significance of this interaction in each species, resulting in N p-values $p_i \dots p_N$. Now a test statistic for the combination of the evidence indicated by this specific vector of p-values is required.

One flexible and robust approach, known as the inverse normal method, transforms the p-values into so-called probits $t_i = \Phi^{-1}(p_i)$, where Φ^{-1} is the inverse of the standard normal distribution. Again, because each p_i is uniform within [0..1], we obtain $t_i \sim N(0,1)$. Under the assumption of independence, the combined value

$$t = \frac{\sum_{i=1}^{N} t_i}{\sqrt{N}} \qquad (1)$$

follows a normal distribution (i.e., $t \sim N(0,1)$).

The main problem, however, is that the independence assumption does not hold in many cases and clearly not in our setting in which we include even close homologs that share extensive sequence similarity. Thus, in particular, the test for closely related species will be strongly correlated. In more detail, the degree of correlation depends on the evolutionary relationship between the species, and we have to consider this when combining the p-values.

Hartung (10) introduced a modification of the inverse normal method that accounts for correlations between p-values and allows one to weight the difference test. He assumed a constant correlation

$\rho = corr(t_i, t_j)$ between each pair $i \neq j$ of probits. When ρ is known, then the Eq (1) can be rewritten as follows:

$$t(\rho) = \frac{\sum_{i=1}^{N} t_i}{\sqrt{N+N(N-1)\rho}} \text{ with } t(\rho) \sim N(0,1). \qquad (2)$$

Correcting for correlation

The remaining problems are to estimate the correlation, and to weight the different tests due to their different similarities. Concerning the latter, one approach would be to include the weighting in the correlation correction by deviating from the assumption of a constant correlation, i.e., by assuming a known correlation matrix R. The extension of Hartung's method to the case of non-constant correlation has been investigated in detail by (11), who used the copula approach to determine conditions for the correlation matrix that still allow for (asymptotic) normality. Another example is the work by (12), who also considered an extension of the inverse normal method, allowing for a non-constant correlation, with an application to the significance analysis of GO terms. There are several reasons for not applying these techniques in our case. First, we typically use between 5 and 8 species, which would imply that we have to estimate between 10 and 45 parameters for the correlation matrix, which causes problems. Second, for a pair of sRNA r_1 and mRNA r_2, we do not want to identify only interactions between r_1 and r_2 that are conserved in all species. Instead, we are also interested in interactions that display a conserved regulon in a "core group" of species, particularly if the set of species considered contains also distantly related species. Consider the two trees in **Fig. S15**. In the first tree, we will most likely find interactions that are conserved in all species $S_1 \ldots S_2$. In the second tree, however, although there are interactions that might be conserved in all species, it is very likely that we also will find important interactions that are conserved only in $S_1 \ldots S_3$ due to the huge evolutionary distance of S_4. The latter aspect cannot be modeled with correlation alone because it is very likely that the correlation (due to sequence similarity) between S_4 and the group $S_1 \ldots S_3$ is close to zero. Thus, we are likely to overestimate the

importance of S_4 because correlation 0 would imply independence of the test associated with S_4, which, thus, receives a high weight.

For that reason, we decided to introduce a weighting that reflects evolutionary distances and, in addition, to correct for an overall correlation between the different p-values. Thus, we compute weights $\lambda_1 \ldots \lambda_N$ for each p-value that are derived from a phylogenetic tree (based on 16S rDNA) for the species, as described later, and an overall correlation ρ that has to be estimated from the data. Following (10), this resorts to the modification of the weighted inverse normal method:

$$t\left(\rho, \vec{\lambda}\right) = \frac{\sum_{i=1}^{N} \lambda_i t_i}{\sqrt{(1-\rho)\sum_{i=1}^{N} \lambda_i^2 + \rho\left(\sum_{i=1}^{N} \lambda_i\right)^2}}, \qquad (3)$$

which approximately follows a normal distribution. Thus, by using the probability integral transformation, we can derive a p-value for the combined test, as follows:

$$P = \Phi(\mathrm{t}\left(\rho, \vec{\lambda}\right)) \qquad (4)$$

The last problem is to estimate ρ. Hartung (10) provided a method to estimate ρ from the t_i. We decided, however, to follow the approach that was already successfully employed in the interaction prediction tool RNAhybrid (8), albeit on a different combined test statistic that did not allow for weighting. The idea is simply that if we correctly estimated ρ and thus $\mathrm{t}\left(\rho, \vec{\lambda}\right)$ follows a normal distribution under the null hypothesis, then the value P calculated in equation (4) should be uniformly distributed under the combined null hypothesis. Thus, for determining the ρ empirically, we evaluate each possible value of ρ in the interval [0..1] (in steps of 0.1). For every such ρ, the distribution of the p-values according Eq. (4) is compared to the uniform distribution. For the optimal $\hat{\rho}$ according to this comparison (using the least square error measurement), the same procedure is repeated in the interval $[\hat{\rho} - 0.1 \,..\, \hat{\rho} + 0.1]$ in steps of 0.01.

Tree-based weighting

Finally, we have to estimate the weights of the different sequences in the definition of the combined p-values. The problem is related to the weights used in the calculation of a multiple sequence alignment. Different publications agreed on the fact that these kinds of weights have to be introduced. However, to the best of our knowledge there is no accepted theory of how to determine them. An overview of different methods to determine such weights is given in Wallace et al. (13). One popular method is the one introduced by Thompson et al. (14). The basic idea here is to add up the weights for each edge from the organism to the root. The weight of each edge, however, is divided by the number of organisms below the edge, thus distributing the weight of each edge to the associated organisms. In the example tree given in **Fig. S16A**, the species O_1 gets the absolute weight $w_3^a = 1 + \frac{9}{1} + \frac{9}{3} = 8.5$. Analogously, we get $w_2^a = 8.5$ and $w_1^a = 13$. This gives rise to the relative weights of $w_1 = 43.\overline{3}\%$, $w_2 = 28.\overline{3}\%$ and $w_3 = 28.\overline{3}\%$.

If we now compare the relative weight of species O_1 in the tree in **Fig. S16A** with the weight in **Fig. S16B**, one can conclude that the situation is not very different. In both cases, the weight of O_1 should be around 50%. The fact that the first tree contains two close homologs O_2 and O_3 should not change much in the relative weight of O_1. For that reason, we introduce a weighting scheme that recursively splits the weights of the edges according to the weight induced by the complete subtree. The relative weight of the subtree is defined as the sum of all edges, this defining the relative weights for each subtree. The basic idea is explained in **Fig. S17**. Using this recursive scheme, the final weight for each organism is given by

$$w(O_x) = \prod_{i \in ancestors(x)} relweight(i)$$

where ancestors(x) are all ancestors of x(i.e., all internal nodes on the path between the root and x), and relweight(i) is the relative weight of node iw.r.t. its sibling. Given the parent $p(i)$ of node i, relweight(i) is defined by

$$relweight(i) = \frac{\frac{w_{e(p(i),i)}}{2} + weight(i)}{weight(p(i))}$$

Here, $w_e(p(i), i)$ is the weight of the edge between i and the parent of i, and for any node j, weight (j) is the sum of all edges' weights for nodes below j. Finally, the weights are modified by a root function to limit the influence of outlier species while maintaining a high resolution for closely related species.

CopraRNA implementation

CopraRNA is implemented in Perl and R. Several Perl modules, R libraries and bio software packages are incorporated. The general design is depicted in **Fig. 1A.** As input, CopraRNA initially requires the homologous sRNA sequences of each participating organism in FASTA format and the affiliated RefSeq Ids of the genomes (i.e. NC_000913 for *E. coli*). Only one RefSeq Id is needed per sequence. Additional replicons are automatically retrieved and included in the analysis. Furthermore, the regions that shall be subjected to the computation must be specified. Regions upstream and downstream of either start or stop codon can be selected. In the benchmark analysis for example, we applied regions of -200 and +100 with respect to the start codon.

The putative target and 16S rDNA sequences are parsed from all replicons of the participating organisms. Sequence retrieval is aided by BioPerl. Then the clusters of homologous genes amongst all organisms are calculated by application of DomClust, the algorithm behind the homology calculations on MBGD (15). In the following, IntaRNA (1) (options -p 7 -w 140 -L 70 -o) predicts the whole genome interactions for each individual organism. This step is aided by the Perl module Parallel::ForkManager. This greatly reduces runtime, as IntaRNA tasks run in parallel instead of successively.

After the single predictions' completion, the combination of results commences. Statistical operations are implemented in R and evaluated within Perl, assisted by the Perl module Statistics::R, while the R evir library simplifies handling of extreme value distributions. The EMBOSS programs

emma (clustalw wrapper), distmat (Jukes-Cantor method) and fneighbor (neighbor-joining) are combined in order to calculate the phylogenetic tree from which the weights for the individual organisms are derived. These weights are subjected to a root function (i.e. $weight^{-2.5}$) in order to reduce inappropriately strong influence of outliers. Given the homologous gene clusters, the weights and the IntaRNA predictions, the combined p-value is calculated for each cluster that contains genes from at least 50% of all participating organisms. Missing p-values in clusters $\geq 50\%$ are sampled, using a multivariate normal distribution, in order to maintain the original correlation of the data. The multivariate normal distribution is calculated on the clusters containing genes from every organism entered in the analysis.

Finally, the results are annotated, the interaction regions plots for mRNAs and sRNAs are generated (R script using the seqinr library and clustalw) and automatic functional enrichment is assessed by utilizing the Perl interface supplied by DAVID (16). Further Perl modules, which were used but not contextually explained, are List::MoreUtils, SOAP::Lite, HTTP::Cookies and Getopt::Long.

Table S1. Benchmark set of experimentally verified sRNA targets. Published verified targets which were used to benchmark CopraRNA against the single organism specific prediction tools IntaRNA, TargetRNA and RNApredator. Obvious artifacts, i.e. predicted interactions with the complementary strand of the genomic region coding for the respective sRNA were excluded. If a target was verified for more than one organism we included the respective data, but for our calculations we used the non-redundant set of 101 experimentally verified targets.

Rank							
IntaRNA	**TargetRNA**	**RNApredator**	**CopraRNA**	**Target**	**sRNA**	**Reference**	**Evidence**
285	>100	>100	644	b2741(rpoS)	ArcZ	(17)	L, CM, GAL
3480	>100	>100	2393	stm3216	ArcZ	(18)	G, CM, W
4420	>100	>100	2454	stm2970(sdaC)	ArcZ	(18)	G, CM, W
1890	>100	>100	1376	stm1682(tpx)	ArcZ	(18)	G, CM, W
7	>100	11	3	stm0687(ybfM)	ChiX	(19)	L, IM, GAL
3	13	2	2	stm1313(celB)	ChiX	(19)	N, IM, indirect GAL
2	11	2	2	b1737(chbC)	ChiX	(20)	N, **
7	>100	8	171	b0619(dpiB)	ChiX	(21)	L, CM, GAL
3	>100	3	3	b0681(chiP)	ChiX	(22)	PE, CM, GS
615	>100	>100	126	b2687(luxS)	CyaR	(23)	L, CM, GAL
1861	>100	>100	1742	b1740(nadE)	CyaR	(23)	L, CM, GAL
92	>100	>100	4	b0814(ompX)	CyaR	(23)	L, IM, GAL
570	>100	>100	935	b2666(yqaE)	CyaR	(23)	L, CM, GAL
44	>100	>100	4	stm0833(ompX)	CyaR	(24)	G, CM, W
5	18	16	3	b1237(hns)	DsrA	(25)	L, CM, indirect GAL
1	3	20	2	b2741(rpoS)	DsrA	(25)	L, IM, GAL
428	>100	>100	1448	b3908(sodA)	FnrS	(26)	G, W
470	>100	>100	74	b1656(sodB)	FnrS	(26)	G, W
23	>100	>100	1969	b3829(metE)	FnrS	(26)	G, W
584	>100	>100	1950	b0887(cydD)	FnrS	(26)	PE
502	>100	>100	403	b2153(folE)	FnrS	(27)	L, IM, GAL
876	>100	>100	1781	b2303(folX)	FnrS	(27)	L, IM, GAL
1341	>100	>100	389	b0755(gpmA)	FnrS	(27)	L, CM, GAL
316	19	>100	55	b1479(maeA)	FnrS	(27)	L, CM, GAL
3	2	>100	49	b1841(yobA)	FnrS	(26)	G, W
42	42	>100	58	b3089(sstT)	GcvB	(28)	L, CM, GAL
58	12	>100	4	b4208(cycA)	GcvB	(29)	L, IM, GAL
52	26	84	9	stm2355(argT)	GcvB	(30)	G, IM, W
16	2	91	8	stm3630(dppA)	GcvB	(30)	G, IM, W
90	>100	>100	12	stm0665(gltI)	GcvB	(30)	G, IM, W
84	>100	>100	2	stm3567(livJ)	GcvB	(30)	G, IM, W
47	>100	46	2	stm3564(livK)	GcvB	(30)	G, IM, W
164	47	>100	14	stm1746.s (oppA)	GcvB	(30)	G, IM, W
219	>100	>100	*	stm4351	GcvB	(30)	G, IM, W

234	>100	>100	840	stm3064(iciA)	GcvB	(9)	G, IM, FL
3	>100	5	5	stm3909(ilvC)	GcvB	(9)	G, IM, FL
1297	>100	>100	40	stm3903(ilvE)	GcvB	(9)	G, IM, FL
93	>100	>100	6	stm1299(gdhA)	GcvB	(9)	G, IM, FL
79	48	>100	19	stm3062(serA)	GcvB	(9)	G, IM, FL
95	>100	>100	3	stm0959(lrp)	GcvB	(9)	G, IM, FL
49	>100	>100	34	stm0399(brnQ)	GcvB	(9)	G, IM, W
12	5	>100	4	stm4398(cycA)	GcvB	(9)	G, IM, FL
1191	>100	>100	144	stm1452(tppB)	GcvB	(9)	G, IM, FL
67	>100	>100	612	stm2526(ndk)	GcvB	(9)	G, IM, FL
747	51	>100	85	stm0602(ybdH)	GcvB	(9)	G, IM, FL
403	>100	>100	58	stm3225(ygjU)	GcvB	(9)	G, IM, FL
239	>100	>100	124	csgD (b1040)	GcvB	(31)	FLA, W
359	>100	>100	916	stm0245 (metQ)	GcvB	(9)	G, IM, FL
284	31	>100	62	stm0001 /stm0002 (thrL/thrA)	GcvB	(9)	G, IM, FL
603	>100	>100	90	b3729(glmS)	GlmZ	(32)	G, CM, W
181	>100	>100	29	stm4231(lamB)	MicA	(33)	L, CM, GAL
87	12	>100	27	b1130(phoP)	MicA	(34)	L, CM, GAL
419	>100	>100	13	b0814(ompX)	MicA	(35)	G, CM, FL
550	>100	>100	526	b0411(tsx)	MicA	(35)	G, CM, FL
78	67	>100	9	b0957(ompA)	MicA	(36)	TP, MS, 2D, SP, N
1	1	22	1	b2215(ompC)	MicC	(37)	LU, CM
168	>100	>100	239	stm1572 (nmpC/ompD)	MicC	(38)	CM, W
1	>100	>100	1	stm0959(lrp)	MicF	(39)	G, CM, FL
1	>100	>100	1	b0889(lrp)	MicF	(40)	G, IM, PR
359	>100	>100	*	stm1328(lpxR)	MicF	(39)	G, CM, FL
259	>100	>100	219	stm0366(yahO)	MicF	(39)	G, CM, FL
242	>100	>100	44	b3912(cpxR)	MicF	(40)	G, IM, PR
165	7	>100	52	b0241(phoE)	MicF	(40)	G, IM, PR
12	6	4	3	b0929(ompF)	MicF	(41)	L, GAL
60	>100	60	13	b1040(csgD)	OmrA	(42)	G, FLA, CM, W
950	>100	>100	1110	b2155(cirA)	OmrA	(43)	L, CM, GAL
59	>100	>100	133	b0565(ompT)	OmrA	(43)	L, CM, GAL
83	>100	>100	22	b3405(ompR)	OmrA	(43)	L, CM, GAL
28	>100	11	2	b1040(csgD)	OmrB	(42)	G ,FLA, CM, W
1185	>100	>100	867	b2155(cirA)	OmrB	(43)	L, CM, GAL
162	>100	>100	318	b0565(ompT)	OmrB	(43)	L, CM, GAL
275	>100	>100	16	b3405(ompR)	OmrB	(43)	L, CM, GAL
1829	>100	>100	3384	b2731(fhlA)	OxyS	(44)	L, CM, GAL

19	>100	>100	1	b2741(rpoS)	RprA	(45)	L, CM, GAL
520	>100	30	414	b1341(ydaM)	RprA	(46)	Gs, CM, W, L, GAL
12	18	8	71	b1040(csgD)	RprA	(46)	Gs, CM, W
414	>100	>100	50	b1341(ydaM)	RprA-S	(46)	Gs, CM, W, L, GAL
1	2	1	2	b1040(csgD)	RprA-S	(46)	Gs, CM, W
2627	>100	>100	386	b2741(rpoS)	RprA-S	(45)	L, CM, GAL
541	>100	>100	511	stm2391(fadL)	RybB	(47)	G, CM, FL, W
1017	>100	>100	212	stm1070(ompA)	RybB	(47)	G, CM, FL, W
98	>100	>100	12	stm2267(ompC)	RybB	(47)	G, CM, FL, W
618	>100	>100	1674	stm1572(ompD)	RybB	(47)	G, CM, W
226	>100	>100	*	stm0999(ompF)	RybB	(47)	G, CM, FL, W
96	32	>100	3	stm1473(ompN)	RybB	(48)	G, CM, W
556	>100	>100	3	stm1995(ompS)	RybB	(47)	G, CM, W
931	>100	>100	132	stm1732 (ompW)	RybB	(47)	G, CM, FL, W
491	>100	>100	431	stm0413(tsx)	RybB	(47)	G, CM, FL, W
894	>100	>100	165	stm0687 (ybfM, chiP)	RybB	(49)	L, CM, GAL
377	>100	>100	*	b0805 (fiu)	RybB	(35)	G, CM, W
2441	>100	>100	1848	b0721 (sdhC)	RybB	(50)	L, CM, GAL
74	>100	>100	12	b2215 (ompC)	RybB	(51)	W
636	>100	>100	132	b1256 (ompW)	RybB	(51)	W
652	>100	>100	902	b2594 (rluD)	RybB	(35)	G, CM, W
112	24	>100	53	b3607 (cysE)	RyhB	(52)	L, CM, GAL
239	>100	>100	86	b2530 (iscS)	RyhB	(53)	W
3129	>100	>100	1274	b0683 (fur)	RyhB	(54)	IV, IM
721	>100	13	360	b1981(shiA)	RyhB	(55)	L, CM, GAL
508	34	>100	58	b1656 (sodB)	RyhB	(56)	IV, CM
193	35	27	125	b0721 (sdhC)	RyhB	(50)	L, CM, GAL
164	>100	>100	33	b1612 (fumA)	RyhB	(57)	L, CM, W
5	>100	17	5	b1101 (ptsG)	SgrS	(58)	N, CM
1602	>100	>100	270	b1817 (manX)	SgrS	(59)	L, CM, GAL
1163	>100	>100	*	stm2945 (sopD)	SgrS	(60)	G, CM, W
118	>100	>100	3	stm3962 (yigL)	SgrS	(61)	G, W, CM
1	1	>100	1	b0757(galK)	Spot42	(62)	TP, 2D, W, MS
392	>100	>100	2	b0720(gltA)	Spot42	(63)	L, IM, GAL, N
50	35	32	332	b4311(nanC)	Spot42	(63)	L, CM, GAL
300	>100	>100	488	b2702(slrA)	Spot42	(63)	L, CM, GAL, N
441	>100	>100	2784	b3962(sthA)	Spot42	(63)	L, CM, GAL, N
72	>100	>100	51	b3566(xylF)	Spot42	(63)	L, IM, GAL
1607	>100	>100	583	b2802(fucI)	Spot42	(63)	L, IM, GAL, N
95	>100	18	14	b0721(sdhC)	Spot42	(50)	L, CM, GAL

84	>100	>100	15	b2416 (ptsI)	CyaR	this study	G, FL
52	>100	>100	7	b1824 (yobF)	CyaR	this study	G, CM, FL
60	>100	>100	18	b0723 (sdhA)	CyaR	this study	G, FL
83	>100	>100	34	b1531 (marA)	FnrS	this study	G, FL
5	10	>100	26	b2531 (iscR)	FnrS	this study	G, FL
755	>100	>100	10	b1107 (nagZ)	FnrS	this study	G, FL
22	>100	82	15	b0723 (sdhA)	FnrS	this study	G, FL
76	>100	95	2	b0081 (mraZ)	RybB	this study	G, W
55	>100	48	7	b3365 (nirB)	RyhB	this study	G, CM, FL
128	>100	>100	37	b0156 (erpA)	RyhB	this study	G, FL
1369	>100	>100	15	b1531 (marA)	RyhB	this study	G, FL
53	>100	>100	4	b1107 (nagZ)	RyhB	this study	G, FL
74	>100	>100	56	b0723 (sdhA)	RyhB	this study	G, FL
25	10	>100	21	b2416 (ptsI)	SgrS	this study	G, FL
43	81	>100	30	b1761 (gdhA)	Spot42	this study	G, FL
168	>100	>100	3	b0728 (sucC)	Spot42	this study	G, FL
48	87	>100	24	b1136 (icd)	Spot42	this study	G, FL

L) LacZ-fusion, G) GFP-fusion, LU) Luciferase fusion, FLA) Flag-fusion, W) Western blot, FL) Flow cytometry, GAL) b-Galactosidase activity assay, N) Northern blot, PE) Primer extension, GS) Gel shift, PR) Plate reader, TP) Toeprint, 2D) 2D electrophoresis, MS) mobility shift assay, SP) structural probing, CM) compensatory mutations, IM) interaction site mutations, IV) in vitro translation assay, RT) real time PCR

* not enough homologs for calculation

**overexpression of a chbB-chbC IGR ChiX trap

Table S2. Comparison of CopraRNA with other target prediction tools. Total number of verified targets for the 18 benchmark sRNAs which were detected by the indicated prediction method and varying cut-offs for the prediction rank. We show the results based on a set of 101 published targets with (right side of the slash) and without (left side of the slash) the results of the 17 verifications made in this study.

Prediction rank cut-off	CopraRNA	IntaRNA	targetRNA	RNApredator
1	8 / 8	5 / 5	2 / 2	1 / 1
2	15 / 16	6 / 6	5 / 5	2 / 2
3	19 / 22	9 / 9	6 / 6	3 / 3
5	23 / 26	11 / 12	6 / 6	5 / 5
15	32 / 41	13 / 14	11 / 13	8 / 8

Table S3. Complete CopraRNA result list for the 18 benchmark sRNAs (additional multisheet Excel file). Results of the CopraRNA predictions for all benchmark sRNAs. Each worksheet contains the whole ranked prediction list for one sRNA. The final lists only contain genes which are conserved in at least 50% of the investigated species, with regard to the MGDB cluster table and a predicted IntaRNA energy score of < 0 kcal/mol in at least 50% of the investigated species. Each sheet contains the CopraRNA p-value (column1), the q-value (column2) and the annotation of the gene cluster according to MGDB (column3). If an MGDB cluster contains more than one gene in the organism of interest (i.e., a homolog), the respective locus tag or tags are given in column 4. For the calculation of the CopraRNA p-value always the homolog with the lowest predicted IntaRNA interaction energy is used. The following columns give the organism-specific prediction results and gene information, beginning with the locus tag. Within brackets the following information is quoted: Gene name, single organism specific prediction energy [kcal/mol], single organism specific p-value, start of the interaction in the target RNA input sequence, end of the interaction in the target sequence, start of the interaction in the sRNA input sequence, end of interaction in the sRNA sequence, Entrez GeneID of the target.

Table S4. Functional enrichment for the 18 benchmark sRNAs (additional multisheet Excel file). Functional enrichments (Functional Annotation Clustering) as provided from the DAVID webserver (16) for all benchmark sRNAs. For each sRNA the target candidates ($p \leq 0.01$) were tested against all genes in the list as background. Obvious artifacts, i.e. predicted interactions with the complementary strand of the respective sRNA genomic coding region were excluded. Enrichments were done for *E. coli*. The standard parameters were changed to a "Similarity Threshold" of 0.85 and an "Initial Group Membership" and "Final Group Membership" of 2. A help text for the table is given at the DAVID website under the URL:
http://david.abcc.ncifcrf.gov/helps/functional_annotation.html#E4.

Table S5. Predicted targets after post-processing (additional multisheet Excel file). The candidates after post-processing include the top 15 predictions for each sRNA and those targets with

a p-value ≤ 0.01 which are significantly functional enriched by the DAVID webserver (4) (i.e. they belong to a cluster with an DAVID score of ≥ 1.1). For each predicted target the CopraRNA p-value and information about the IntaRNA results for the reference organism are given (locus tag, gene name, IntaRNA energy, IntaRNA p-value, interaction coordinates in the mRNA and the sRNA, Entrez GeneID). Furthermore, the table shows the enriched terms with the corresponding DAVID score, and the membership of a predicted target to the first term in the respective cluster (compare Table S4). Membership is indicated by a "1" and a yellow background in the respective table element.

Table S6. List of potentially multiple targeted mRNAs (additional Excel file). List of all predicted targets with a CopraRNA p-value ≤ 0.01 for at least one of the benchmark sRNAs. Columns 1 - 3 give the Entrez GeneID, the gene name and the locus tag of the target candidate for *E. coli*. Column 4 gives the number of sRNAs which are predicted to interact with the target. Columns 5-23 indicate for each target/sRNA combination if the prediction meets the threshold criteria (1) or not (0).

Table S7. List of experimentally tested targets. List of the 23 selected predictions which were tested with the GFP-fusion system(39, 64) as part of this study.

sRNA	predicted target	Method	Post-transcriptional regulation
ChiX	opgG (b1048)	Flow cytometry	no
CyaR	ptsI (b2416)	Flow cytometry	yes
CyaR	yobF (b1824)	Plate reader	yes
CyaR	sdhA (b0723)	Flow cytometry	yes
FnrS	marA (b1531)	Flow cytometry	yes
FnrS	iscR (b2531)	Flow cytometry	yes
FnrS	sdhA (b0723)	Flow cytometry	yes
FnrS	nagZ (b1107)	Flow cytometry	yes
GcvB	mraZ (b0081)	Western blot	no
GlmZ	mraZ (b0081)	Western blot	no
MicA	ftsB (b2748)	Western blot	no
MicC	mraZ (b0081)	Western blot	no
RprA	phoU (b3724)	Western blot	no
RybB	mraZ (b0081)	Western blot	yes
RyhB	nirB (b3365)	Plate reader	yes
RyhB	erpA (b0156)	Flow cytometry	yes
RyhB	marA (b1531)	Flow cytometry	yes
RyhB	nagZ (b1107)	Flow cytometry	yes
RyhB	sdhA (b0723)	Flow cytometry	yes
SgrS	ptsI (b2416)	Flow cytometry	yes
Spot42	gdhA (b1761)	Flow cytometry	yes
Spot42	sucC (b0728)	Flow cytometry	yes
Spot42	icd (b1136)	Flow cytometry	yes

Table S8. Benchmark set for non enterobacterial organisms. Published verified targets which were used to benchmark CopraRNA against the single organism specific prediction tools IntaRNA, TargetRNA and RNApredator. TargetRNA predictions have been done on the TargetRNA2 server without usage of the "sRNA conservation and accessibility" option. Targets in this list may have a lower degree of experimental evidence for an actual post-transcriptional regulation than the targets listed in **Table S1.**

Rank							
IntaRNA	**TargetRNA**	**RNApredator**	**CopraRNA**	**Target**	**sRNA**	**Reference**	**Evidence**
109	83	149	2	hctA (ct743)	IhtA	(65)	***
5200	3694	2728	282	sodB (pa4366)	PrrF	(66)	N
186	35	689	10	pa4880	PrrF	(66)	L, GAL
64	1423	120	2	sdhC (pa1581)	PrrF	(67)	RT
2995	4459	3214	166	acnA (pa1562)	PrrF	(67)	RT
410	2650	119	14	acnB (pa1787)	PrrF	(67)	RT
296	2479	915	1747	antR (pa2511)	PrrF	(67)	RT
1022	>100	1351	*	antA (pa2512)	PrrF	(67)	RT
5	14	51	1	chiA (lmo1883)	LhrA	(68)	N, TP, MS
95	>100	69	161	lmo302	LhrA	(68)	L, N, GAL, TP
45	>100	49	30	lmo0850	LhrA	(69)	L, GAL, CM, N, MS
70	103	1556	23	lutA (bsu34050)	FsrA	(70)	W, N
187	88	854	31	lutB (bsu34040)	FsrA	(70)(71)	W, N, 2D, RT
273	117	902	19	sdhC (bsu28450)	FsrA	(71)	MS
9	27	10	4	citB (bsu18000)	FsrA	(71)	2D
118	206	460	35	leuC (bsu28260)	FsrA	(71)	2D
3855	2213	1442	1571	ahrC (bsu24250)	SR1	(72)	IV, MS

L) LacZ-fusion, G) GFP-fusion, LU) Luciferase fusion, FLA) Flag-fusion, W) Western blot, FL) Flow cytometry, GAL) b-Galactosidase activity assay, N) Northern blot, PE) Primer extension, GS) Gel shift, PR) Plate reader, TP) Toeprint, 2D) 2D electrophoresis, MS) mobility shift assay, SP) structural probing, CM) compensatory mutations, IM) interaction site mutations, IV) in vitro translation assay, RT) real time PCR

* not enough homologs for calculation

*** indirect, rescue of hctA dependent growth phenotype in *E. coli*

Table S9. CopraRNA results and DAVID outputs for non enterobacterial sRNAs (additional multisheet Excel file). Results of the CopraRNA predictions for the non enterobacterial sRNAs. Each worksheet contains the whole ranked prediction list for one sRNA. The final lists contain only genes which are conserved in at least 50% of the investigated species, regarding to the MGDB cluster table and have a predicted IntaRNA energy score of < 0 kcal/mol in at least 50% of the investigated species. Each sheet contains the CopraRNA p-value (column1), the q-value (column2) and the

annotation of the gene cluster according to MGDB (column3). If an MGDB cluster contains more than one gene in the organism of interest (i.e., a homolog), the respective locus tag or tags are given in column 4. For the calculation of the CopraRNA p-value always the homolog with the lowest predicted IntaRNA interaction energy is used. The following columns give the organism-specific prediction results and gene information, beginning with the locus tag. Within brackets the following information is quoted: Gene name, single organism specific prediction energy [kcal/mol], single organism specific p-value, start of the interaction in the target RNA input sequence, end of the interaction in the target sequence, start of the interaction in the sRNA input sequence, end of interaction in the sRNA sequence, Entrez GeneID of the target. The CopraRNA output is accompanied by the DAVID webserver functional enrichment output in the adjacent sheet if there was a significant functional enrichment. Please compare with the **Table S5** (Predicted targets after post-processing).

Table S10. List of oligonucleotides used in this study

Name	Sequence	Used for
JVO-9422	agctCtaccaggaaccacc	pKP-299-1
JVO-9423	ggtaGagctagcatttatgg	pKP-299-1
JVO-9424	attgGtcacattgcttcca	pKP-296-2
JVO-9425	GTGACCAATGTCGTGCTTT	pKP-296-2
JVO-9261	gtttttATGCATAGTGGGAAATTGTGGGGC	pKP-287-1
JVO-9262	GTTTTTTGCTAGCGAGATTGACTAACGTTGCTCC	pKP-287-1
JVO-9357	GTTTTTTGCTAGCGAAGCGGTATTCAACGTCA	pKP-295-1
JVO-9356	gtttttATGCATACGCCAGTTTAAGTATCTGC	pKP-295-1
JVO-9350	gtttttATGCATAATAGAAAAGAAATCGAGGC	pKP-293-1
JVO-9351	GTTTTTTGCTAGCTTCGATAAAGCGATGGCC	pKP-293-1
JVO-9416	ggtaGagttctgtttatgtgtg	pKP-300-1
JVO-9417	AACTCTACCTCGTTTAACCC	pKP-300-1
JVO-9420	atgaCcaaagtcagactcg	pKP-297-1
JVO-9421	TTTGGTCATTTTTGCCTC	pKP-297-1
JVO-9348	gtttttATGCATGTTGTCGCGGTATCCCCA	pkp292-1
JVO-9349	GTTTTTTGCTAGCCCACAGCGAATACTGTAGCC	pkp292-1
JVO-9352	gtttttATGCATCAACACGGACGATCTGTTC	pkp294-1
JVO-9353	GTTTTTTGCTAGCACTTTCCAGTTCGGCGTT	pkp294-1
erpA5‘	TTAATGCATATTATTGGGTTAGAATTTGCCCAATTG	pJG1
erpA3‘	TTAGCTAGCTTTAACTTTGTTGGCTGCTGCGTC	pJG1
gdhA5‘	TTAATGCATGCAAAAGCACATGACATAAACAACATA	pJG3

gdhA3‘	TTAGCTAGCATATTTTGGATTTTGTTCAAGAAAAGGC	pJG3
icd5‘	TTAATGCATGCCAATTACAAATCATTAACAAAAAATTGC	pJG6
icd3‘	TTAGCTAGCTTTATAGGCTTTCTCGACTGCAGC	pJG6
iscR5‘	TTAATGCATGCTATGCAATACCCCCACTTTTAC	pJG8
iscR3‘	TTAGCTAGCCGGGCCCGCTTCAGAGTTGA	pJG8
marA5’	TTAATGCATAACTAATTACTTGCCAGGGCAACT	pJG10
marA3‘	TTAGCTAGCTGGCGATTCCAGGTTGTCCTC	pJG10
nagZ5‘	TTAATGCATTGGCTGCTGATGCTCAAAGCA	pJG12
nagZ3‘	TTAGCTAGCCACCAGCGGATGCGCCAGTA	pJG12
ptsI5‘	TTAATGCATAATTATTTTGATGCGCGAAATTAATCGTTAC	pJG14
ptsI3‘	TTAGCTAGCCTGGTCGGCAGAAATTTTTTTCCG	pJG14
sdhA5‘	TTAATGCATTGGCAGGTGTTGACCGACTAC	pJG16
sdhA3‘	TTAGCTAGCGCCGCTCTGGGAAATTTGCAG	pJG16
CyaR5‘	GCTGAAAAACATAACCCATAAAATGCTA	pJG19
CyaR3‘	GTTTTTTCTAGATGGACGTGACCAGAAATAAATCC	pJG19
FnrS5‘	GCAGGTGAATGCAACGTCAAGCG	pJG20
FnrS3‘	GTTTTTTCTAGAGTGGACTCTTAAAGGGTAGACGC	pJG20
ChiX5‘	ACACCGTCGCTTAAAGTGACG	pJG21
ChiX3‘	GTTTTTTCTAGAGAGAAGGGAATTTGCCGCAAATG	pJG21
Spot42*5‘	cttGaGgtaatcggatttggctgaatattttag	pJG22
Spot42*3‘	gattacCtCaagtaaaaggtctgaaagatagaac	pJG22
gdhA*5’	ccgttcGtCaagtaatgaccacactc	pJG23
gdhA*3’	cttGaGgaacggcttgcgcgaac	pJG23
FnrS*5‘	tgtGttacttccttttttgaattactgcatagc	pJG24
FnrS*3‘	GGAAGTAAcACAATATGGAGCGCAACG	pJG24
iscR*5’	GAAGTAAcACATGAGACTGACATCTAAAGGG	pJG25
iscR*3’	catgtGttacttcacctcaaactcgcc	pJG25
erpA*5’	TATGAcTGATGACGTAGCACTGCCG	pJG26
erpA*3’	catcagtcatattttgctccaaacgacatc	pJG26
marA1*5’	gtatgaGCatgtccagacgcaatactga	pJG27
marA1*3’	gacatGCtcatacctctttttttgtttacgg	pJG27
marA2*5’	gtatAaAgatgtAcagacgcaatactga	pJG28
marA2*3’	gTcatcTtTatacctctttttttgtttacgg	pJG28

sucC*5′	ttactgaaAAatggacagaacacatgaacttac	pJG29
sucC*3′	ctgtccatTTttcagtaatcgttatcttttaaacc	pJG29

Table S11. List of plasmids used in this study.

Plasmid trivial name	Plasmid stock name	Relevant fragment	Comment	Origin, marker	Reference
control	pJV300		Control plasmid, expresses a ~50 nt nonsense transcript.	ColE1, Amp^R	(73)
pP_L-RybB	pFM1-1	RybB	RybB expression plasmid	ColE1, Amp^R	(48)
pP_L-CyaR	pKP39-3	CyaR	CyaR expression plasmid used to test the yobF-CyaR interaction	ColE1, Amp^R	(24)
pP_L-RyhB	pJU-002	RyhB	RyhB expression plasmid	ColE1, Amp^R	(64)
pP_L-CyaR*	pKP299-1	CyaR*	Derivative of pKP39-3. Point mutant in *yobF* binding site.	ColE1, Amp^R	this study
pP_L-RyhB*	pKP296-2	RyhB*	Derivative of pJU-002. Point mutant in *nirB* binding site.	ColE1, Amp^R	this study
pP_L-CyaR	pJG19	CyaR	CyaR expression plasmid used to test the interaction with sdhA and ptsI	ColE1, Amp^R	this study
pP_L-FnrS	pJG20	FnrS	FnrS expression plasmid	ColE1, Amp^R	this study
pP_L-ChiX	pJG21	ChiX	ChiX expression plasmid	ColE1, Amp^R	this study
pSpot42	pISpf	Spot42	Spot42 expression plasmid	pMB1, Amp^R	(64)
pGlmZ:	pJV103IH	GlmZ:	GlmZ expression plasmid	ColE1, Amp^R	(32)
pGcvB	pJU-014	GcvB	GcvB expression plasmid	p15A, Amp^R	(64)
pMicA	pJV150IG-34	MicA	MicA expression plasmid	ColE1, Amp^R	(36)
pMicC	pSK-017	MicC	MicC expression plasmid	ColE1, Amp^R	(64)
pRprA	pJV100IA-T4	RprA	RprA expression plasmid	ColE1, Amp^R	(64)
P*opgG:: gfp*	pJU-126	*opgG*	GFP reporter plasmid. Carries the *opgG/mdoG* 5'UTR and ORF.	pSC101*, Cm^R	(64)
P*mraZ:: gfp*	pKP287-1	*mraZ*	GFP reporter plasmid. Carries the *mraZ* 5'UTR and ORF (60bp).	pSC101*, Cm^R	this study
P*yobF::sfgfp*	pKP295-1	*yobF*	GFP reporter plasmid. Carries the *yobF* 5'UTR and ORF (60bp).	pSC101*, Cm^R	this study
P*nirB::sfgfp*	pKP293-1	*nirB*	GFP reporter plasmid. Carries the *nirB* 5'UTR and ORF (60bp).	pSC101*, Cm^R	this study
P*yobF*::sfgfp*	pKP300-1	*yobF**	Derivative of pKP295-1. Point mutant in CyaR binding site.	pSC101*,	this study

				Cm^R	
P*nirB*::sfgfp*	pKP297-1	*nirB**	Derivative of pKP293-1. Point mutant in RyhB binding site	pSC101*, Cm^R	this study
P*ftsB*::*gfp*	pkp292-1	*ftsB*	GFP reporter plasmid. Carries the *ftsB* 5'UTR and ORF (60bp)	pSC101*, Cm^R	this study
P*phoU*::*gfp*	pkp294-1	*phoU*	GFP reporter plasmid. Carries the *phoU* 5'UTR and ORF (60bp)	pSC101*, Cm^R	this study
P*ptsI*::*sfgfp*	pJG14	*ptsI*	GFP reporter plasmid. the complete *ptsH* ORF, the intergenic region between *ptsH* and *ptsI*, and parts of the *ptsI* ORF (105bp)	pSC101*, Cm^R	this study
P*sdhA*::*sfgfp*	pJG16	*sdhA*	GFP reporter plasmid. Carries the sdhA 5'UTR and ORF (90bp) and parts of the sdhD ORF (120nt).	pSC101*, Cm^R	this study
P*iscR*::*sfgfp*	pJG8	*iscR*	GFP reporter plasmid. Carries the iscR 5'UTR and ORF (75bp).	pSC101*, Cm^R	this study
P*marA*::*sfgfp*	pJG10	*marA*	GFP reporter plasmid. Carries the complete marR ORF, the intergenic region between marR and marA, and parts of the marA ORF (75bp)	pSC101*, Cm^R	this study
P*nagZ*::*sfgfp*	pJG12	*nagZ*	GFP reporter plasmid. Carries the last 114bp of the thiK ORF, the intergenic region between thiK and nagZ, and parts of the nagZ ORF (78bp)	pSC101*, CmR	this study
P*erpA*::*sfgfp*	pJG1	*erpA*	GFP reporter plasmid. Carries the erpA 5'UTR and ORF (60bp).	pSC101*, CmR	this study
P*gdhA*::*sfgfp*	pJG3	*gdhA*	GFP reporter plasmid. Carries the gdhA 5'UTR and ORF (135bp).	pSC101*, CmR	this study
P*icd*::*sfgfp*	JG6	*icd*	GFP reporter plasmid. Carries the icd 5'UTR and ORF (114bp).	pSC101*, CmR	this study
PsucC::gfp	pJU-159	*sucC*	GFP reporter plasmid. Carries the last 138bp of the sucB ORF, the intergenic region between sucB and sucC, and parts of the sucC ORF (150bp)	pSC101*, CmR	(64)
pSpot42*	pJG22	Spot42*	Derivative of pISpf. Point mutant in *gdhA* binding site.	pMB1, Amp^R	this study
P*gdhA*::sfgfp*	pJG23	*gdhA**	Derivative of pJG3. Point mutant in Spot42 binding site.	pSC101*, CmR	this study
pP_L-FnrS*	pJG24	FnrS*	Derivative of pJG20. Point mutant in *iscR* binding site.	ColE1, Amp^R	this study
P*iscR*::sfgfp*	pJG25	*iscR**	Derivative of pJG8. Point mutant in FnrS binding site.	pSC101*, Cm^R	this study
P*erpA*::sfgfp*	pJG26	*erpA**	Derivative of pJG1. Point mutant in RyhB binding site	pSC101*, CmR	this study
P*marA*1*::*sfgfp*	pJG27	*marA*1*	Derivative of pJG10. Point mutant in RyhB and FnrS binding site	pSC101*, Cm^R	this study
P*marA*2*::*sfgfp*	pJG28	*marA*2*	Derivative of pJG10. Point mutant in RyhB and FnrS binding site	pSC101*, Cm^R	this study
PsucC*::gfp	pJU-159	*sucC**	Derivative of pJU-159. Point mutant in Spot42 binding site	pSC101*, CmR	this study

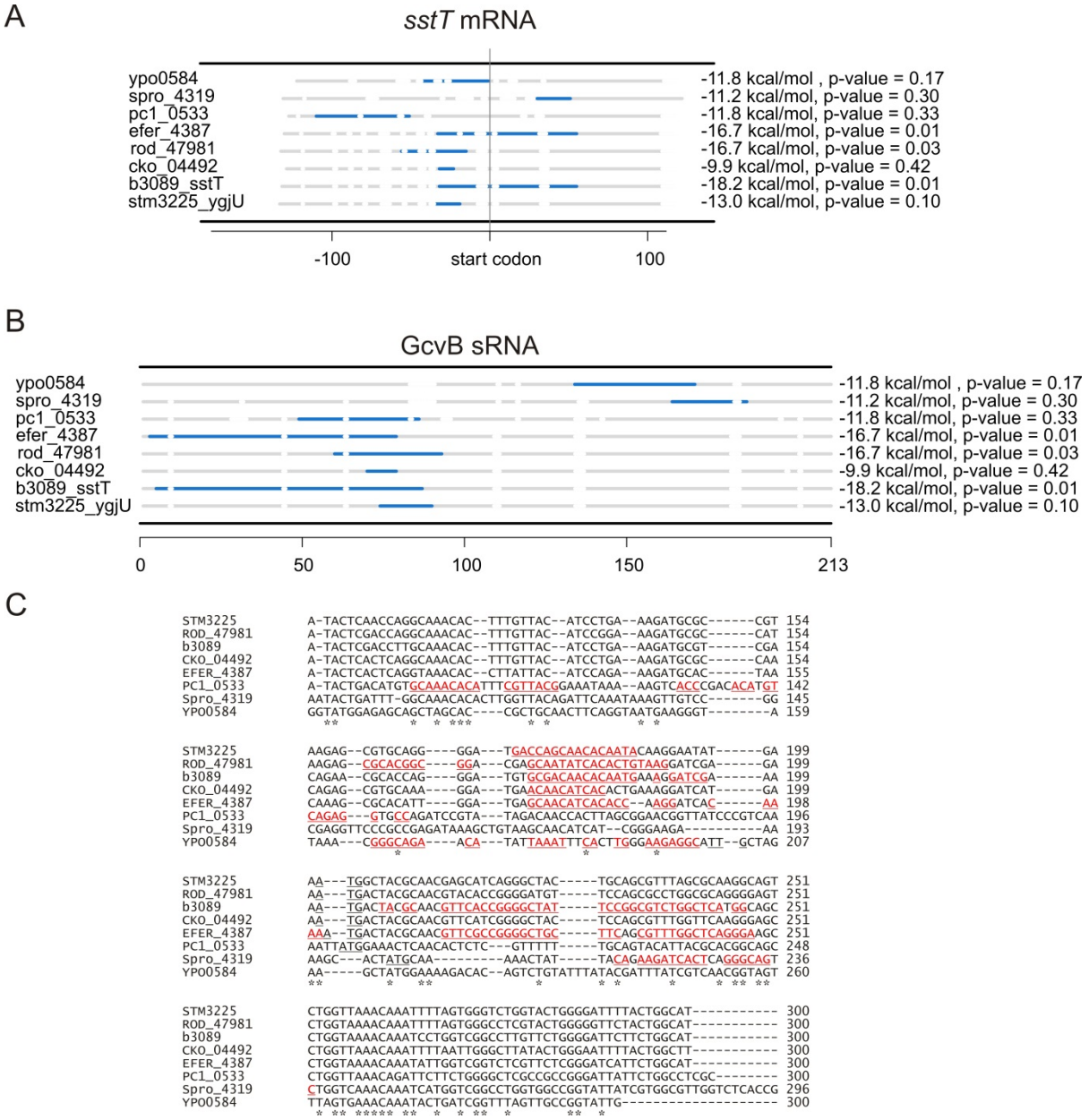

Fig. S1. Comparison of species specific *sstT*-GcvB interactions. A previous study (74) investigated how sRNA-mRNA interactions are conserved. By the comparison of a positive and a negative dataset, it was found that only interaction sites in sRNAs, but not in targets, displayed significant sequence conservation. As a result of the missing target site conservation, there is no general conservation of complementarity between sRNAs and targets. Consequently, predicting a consensus interaction, i.e. base pairs shared by the majority of the homologous sequences, would be too conservative. With our p-value combination strategy we also detect more remote related interactions as they appear e.g. in the *in silico* prediction for the *sstT*-GcvB interaction which was experimentally verified for *E. coli* and *S. enterica*. Panel A) shows a schematic alignment of the interaction sites in the *ssT* mRNA for all investigated organism together with the predicted interaction energy and the single organism specific p-value. Panel B) shows the interaction sites in the repective GcvB homologs. Panel C) shows an alignment of the homologous mRNA sequences used. The bases which are predicted by IntaRNA to participate in base pairing with GcvB are shown in red. There is clearly a relatively strong heterogeneity of the interaction sites on the nucleotide level in the mRNAs and the sRNAs.

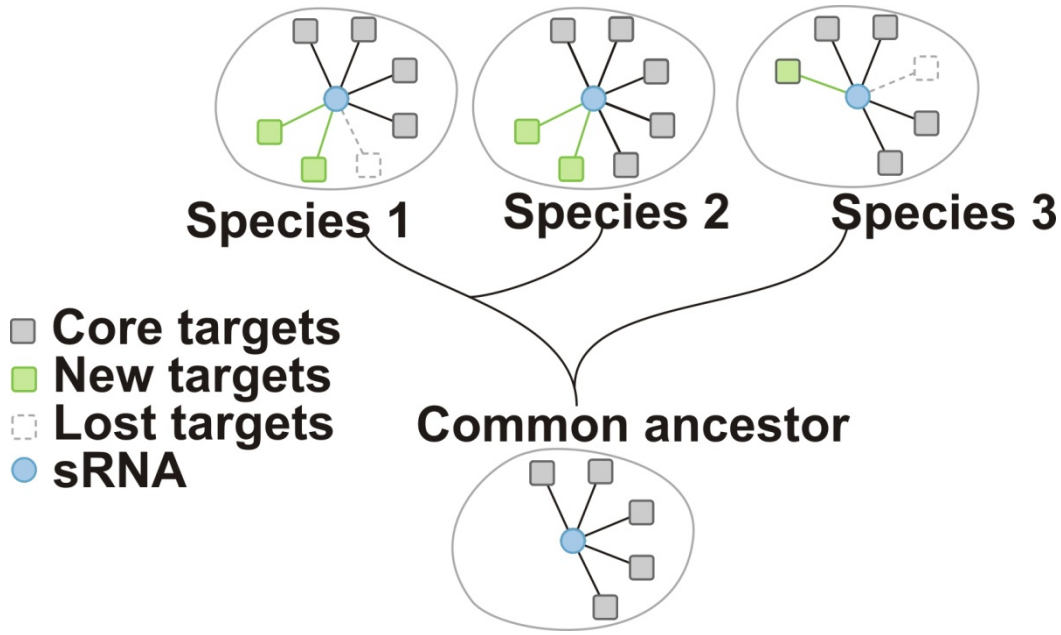

Fig. S2. Evolutionary model. Scheme of the proposed evolutionary model for the conservation of an sRNA and its target set. In our strategy, we consider for a particular sRNA fully conserved core targets that were passed from a common ancestor (grey boxes), targets that were lost in individual species (empty boxes with grey broken lines) and targets that were newly acquired in individual species or branches during evolution (green boxes). The two latter effects lead to targets that are conserved only in specific sub-groups of all species in which the sRNA is evolutionary conserved.

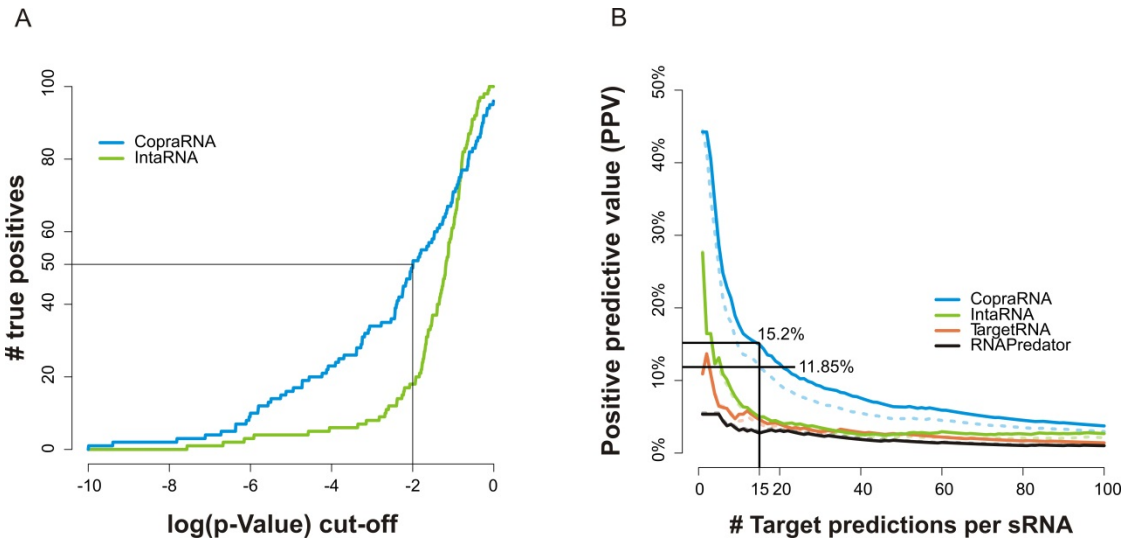

Fig. S3. Benchmark results. a) The plot shows the number of true positive predictions vs. the log transformed p-value cut-off for our comparative target prediction method CopraRNA and the single organism-based method IntaRNA. At a p-value cut-off of 0.01 (i.e. log (0.01) = -2) 50 of the experimentally verified 101 benchmark targets are detected. b) The plot shows the positive predictive value (PPV = (# true positive predictions)/(# positive predictions)) vs. the prediction rank cut-off for our comparative target prediction method CopraRNA and the existing single organism-based methods IntaRNA, TargetRNA and RNAPredator, respectively. These results (including our own verifications) are shown with solid lines while results based only on the benchmark set are indicated with dashed lines. The number of positive predictions is defined in this case as the prediction rank cut-off multiplied by the number of sRNAs in the benchmark dataset. The number of true positives is the number of experimentally verified targets among the positive predictions. To give an example: For the set of 18 sRNAs, at a prediction rank cut-off of 15 a total of 270 targets (= 18*15) is predicted (i.e. positive predictions) and 32 of those are actually experimentally verified (i.e. true positives). The resulting PPV is 11.85 % (32/270*100), i.e. we yield ~1.8 true targets in the top 15 predictions on average per sRNA. If the newly verified targets are included, the number of true positives increases to 41 and the new PPV is 15.2%.

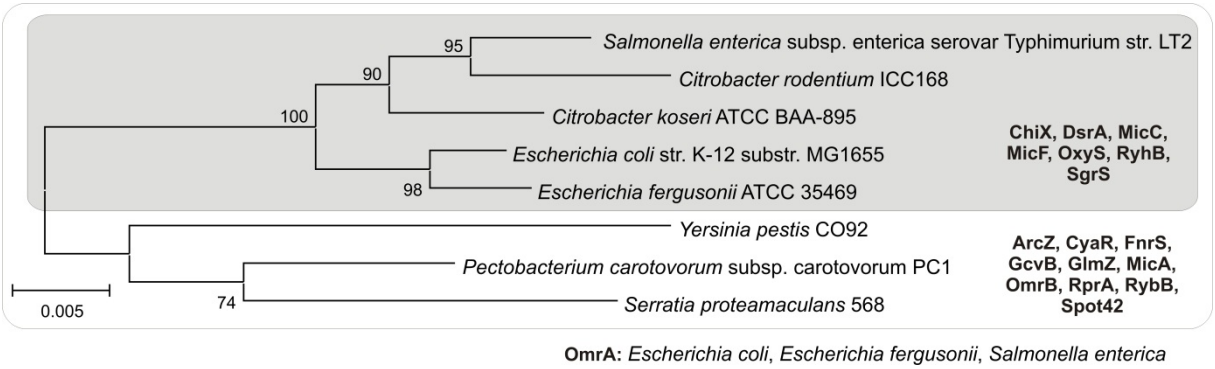

Fig. S4. Phylogenetic tree of enterobacerial benchmark species. The neighbor joining tree is based on 16S rDNA sequences. One group of benchmark sRNAs was only tested in the 5 organisms which are highlighted by the grey box. OmrA is only conserved in 3 of the 8 benchmark organisms.

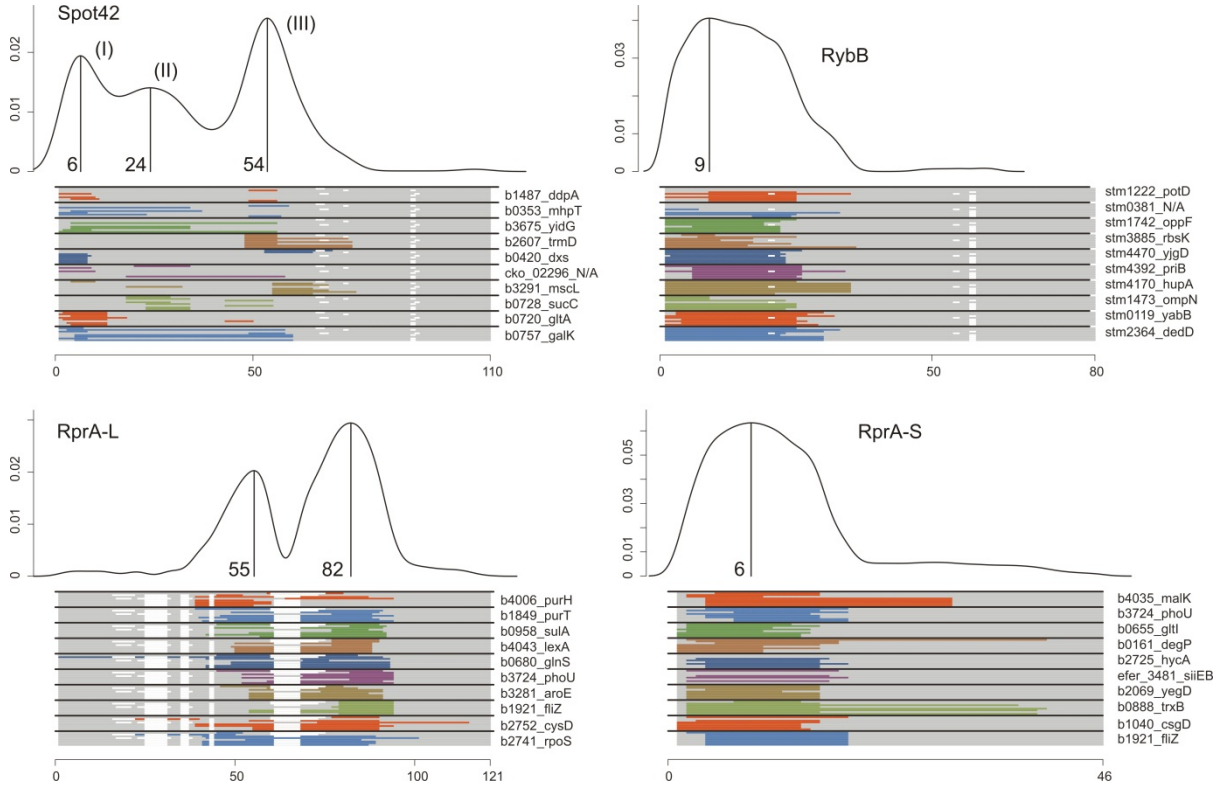

Fig. S5. Visualization of predicted interaction domains. Visualization of the predicted interaction domains in Spot 42, RybB, RprA-L and RprA-S. In the upper part, the density plot gives the relative frequency of the involvement of a specific sRNA or mRNA nucleotide position in predicted target interactions. The plot is a combination of all predicted interactions with a p-value ≤0.01 in all homologs used. Local maxima indicate distinct interaction domains and are marked with upright lines. The number of the central nucleotide of the interaction domain regarding to the multiple sequence alignment is indicated. The lower part shows the predicted interaction domains for the top 10 predicted mRNA targets in a schematic alignment of the homologous sRNAs. Aligned regions are shown in grey, gaps are indicated in white and the predicted interaction regions are indicated as colored lines (color differences are for contrast only). The locus tag and the gene name (if available) of a representative cluster member is given at the right side.

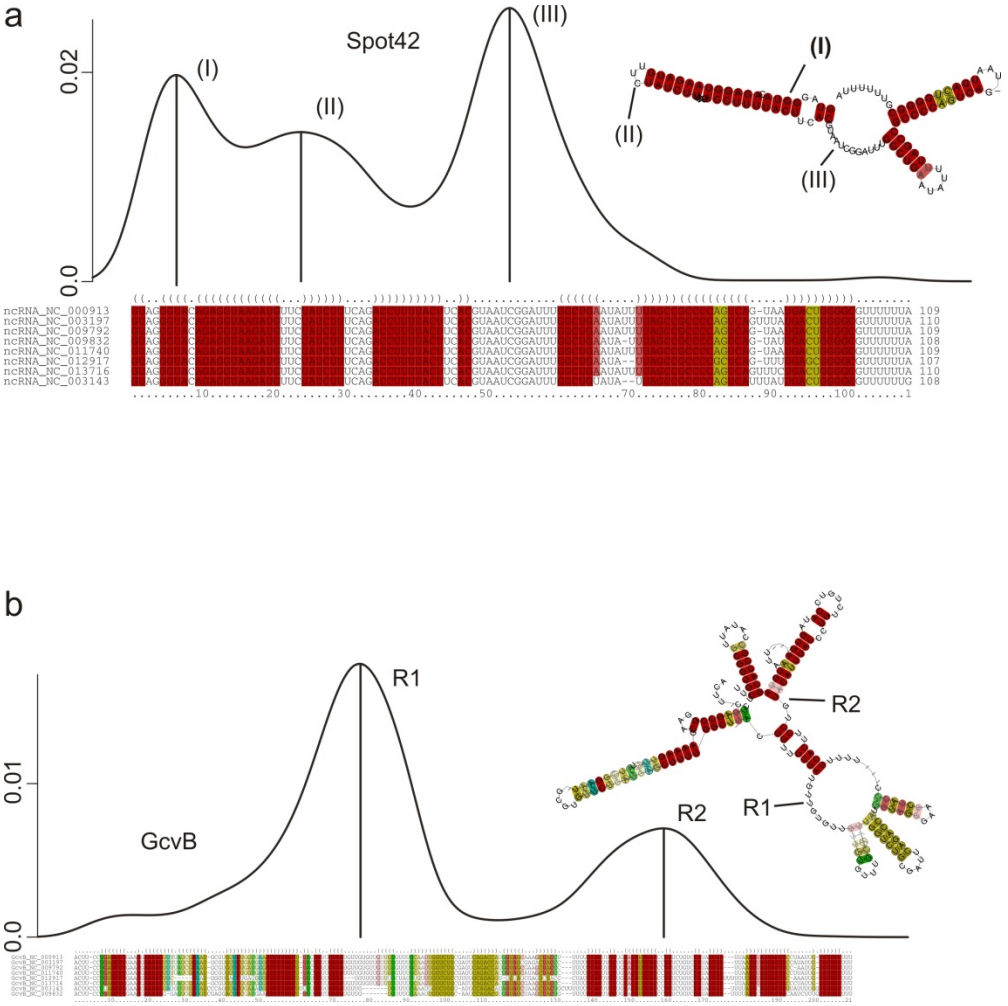

Fig. S6. Comparison of LocARNA alignment with CopraRNA plots. Visualization of the predicted interaction domains in Spot42 (a) and GcvB (b). In the upper part, the density plot gives the relative frequency of the involvement of a specific sRNA or mRNA nucleotide position in predicted target interactions. The plot is a combination of all predicted interactions with a p-value ≤0.01 in all homologs used. Local maxima indicate distinct interaction domains and are marked with upright lines. The plots are combined with a structural LocARNA (75) alignment from all involved sRNA homologs involved and a consensus secondary structure obtained by RNAalifold (76) based on the LocARNA alignment.

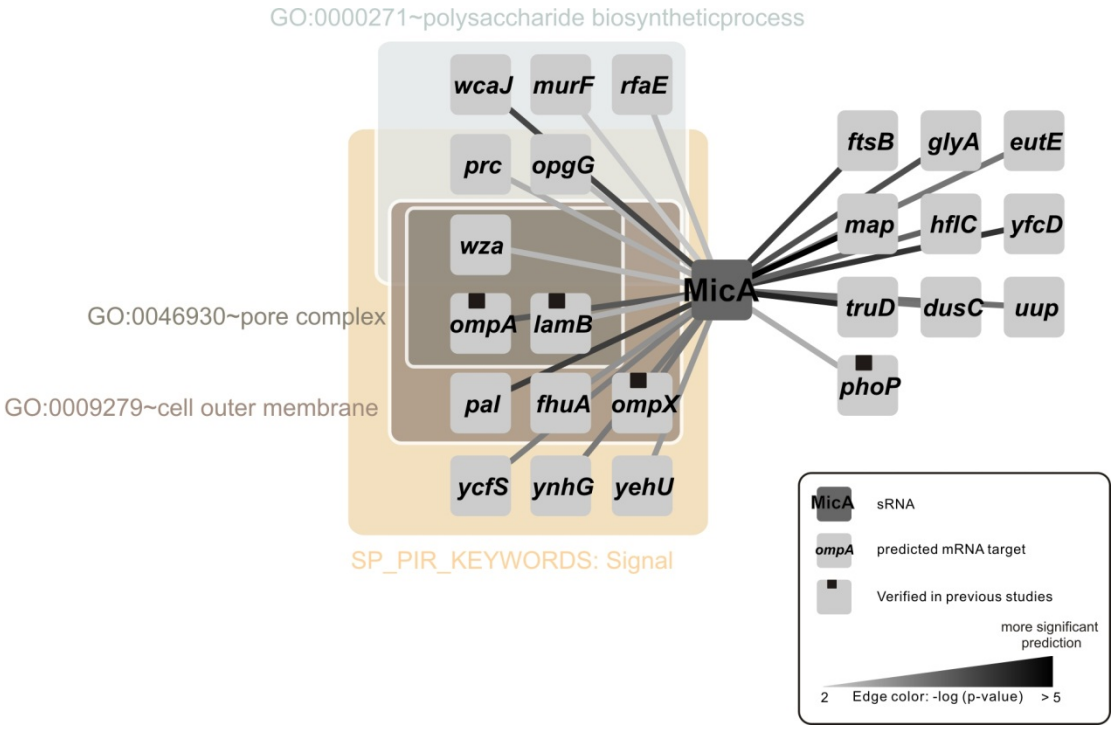

Fig. S7. Functional enrichment of the MicA prediction. Visualization of the functional enrichment analysis for MicA. The figure shows all top 15 predictions and those targets with a p-value ≤ 0.01 which are significantly functional enriched (automatic or by visual inspection). The edges connecting the sRNAs and targets are color coded according to the CopraRNA prediction p-value, a darker color indicates a statistically more significant prediction. Previously experimentally verified targets from the literature (with regard to our benchmark list, **Table S1)** are marked with a black square, verifications from this study with a red square and targets detected by microarrays with a blue square. Functionally enriched targets are color coded with respect to the enriched term.

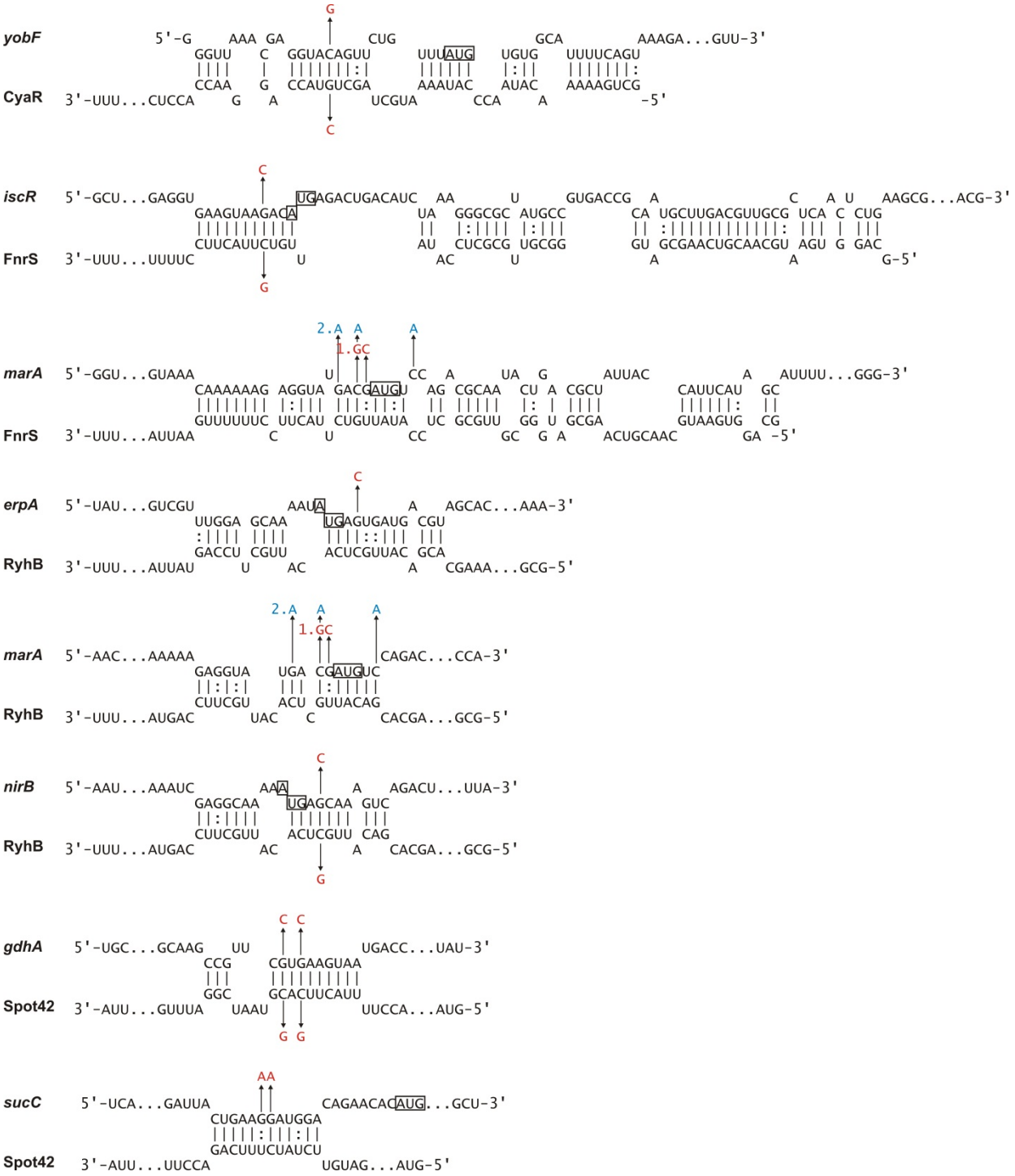

Fig. S8. Predicted base pairings in interactions tested by point mutations. Start codons are marked by a box and introduced point mutations are indicated by an arrow and the changed base in red or blue.

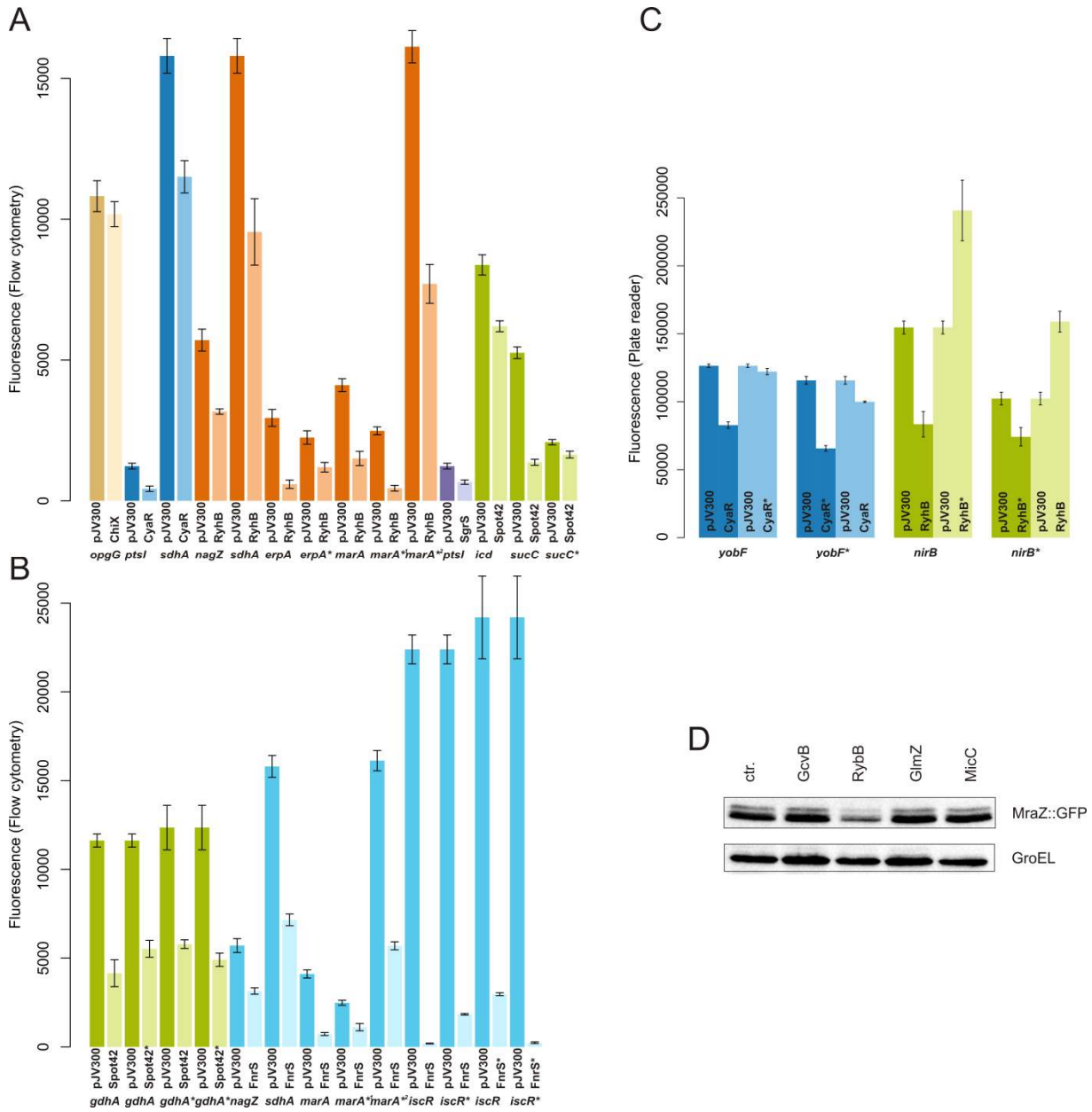

Fig. S9. Results of experimental target verifications (A-C) Mean fluorescence of all constructs in presence of the plasmid carrying the respective sRNA or the control plasmid pJV300 after the subtraction of the background fluorescence in the flow cytometer measurement. **(D)** Western blot with GFP antibody for the *mraZ*-UTR GFP fusion in presence of different sRNAs.

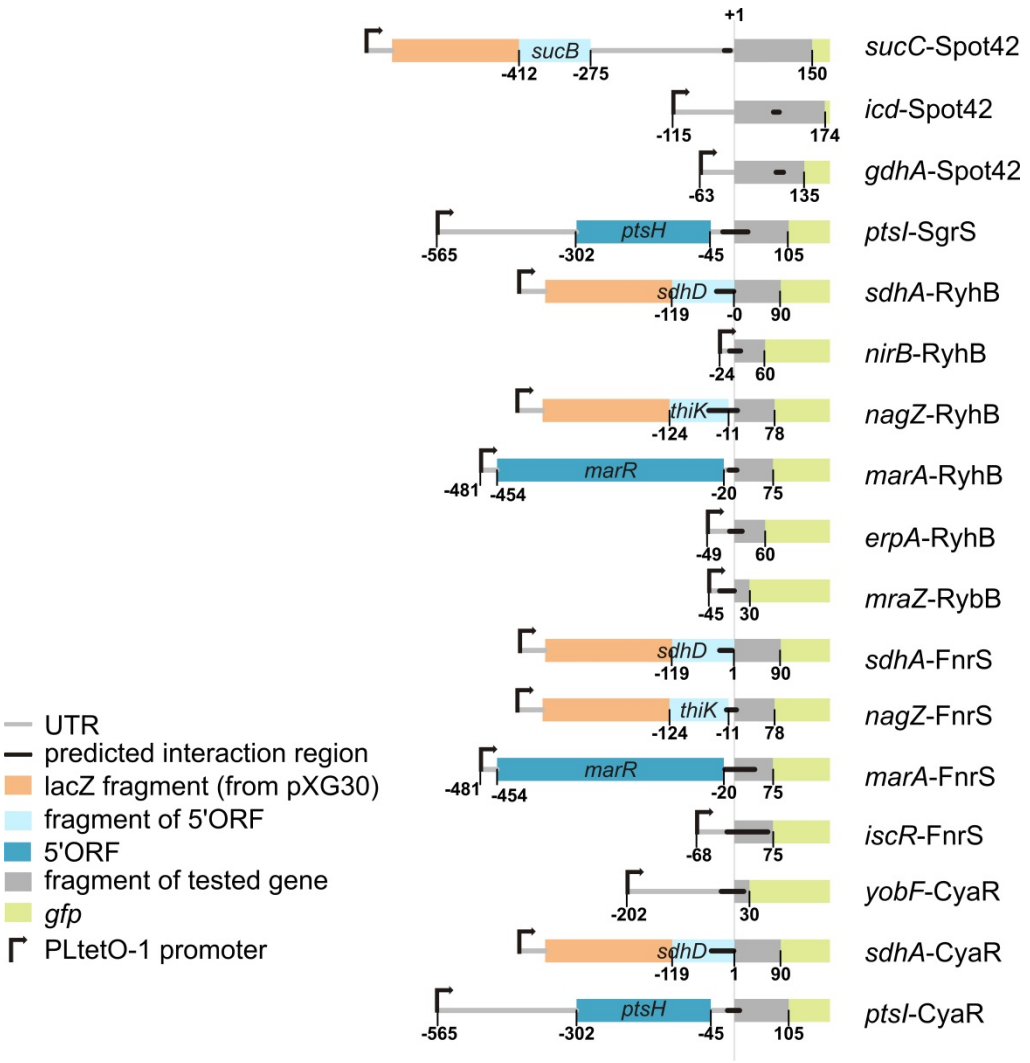

Fig. S10. Scheme of verification constructs. Scheme of the constructs which showed post-transcriptional regulation. The predicted interaction sites are shown as black lines. The color code for other elements is given in the figure. The coordinates of the constructs are given relative to the first nucleotide of the start codon (+1) of the UTR of interest. The pXG10 and pxG30 plasmids that have been used to construct the UTR-GFP fusions are described in (39, 64). The used primers and further information to the resulting plasmids are given in **Table S10** and **Table S11**. The pXG30 plasmid was constructed to mimic a polycistronic operon with an artificial translated gene consistent of a *lacZ* fragment and the 3' part of the gene in front of the gene of interest. In cases where the gene of interest was the second gene in operon we used the complete 5' part and pXG10 instead. All fusions are transcribed from the constitutive PLtetO-1 promoter.

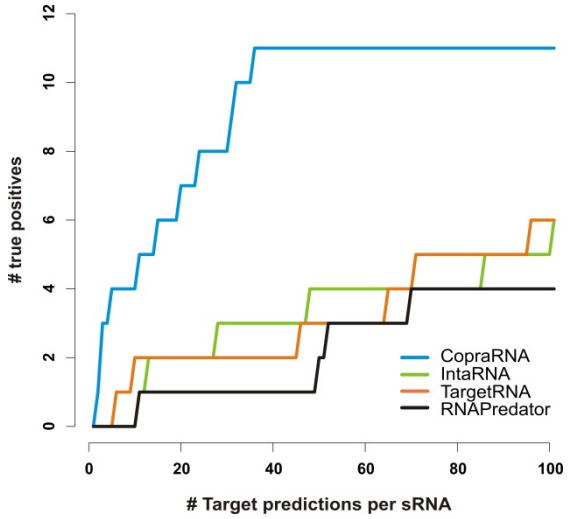

Fig. S11. CopraRNA results for non enterobacterial sRNAs. The plot shows the number of true positive predictions vs. the number of target predictions per sRNA for our comparative method CopraRNA and the existing single organism-based methods IntaRNA, TargetRNA and RNAPredator. The plot was created with the data from **Table S8**.

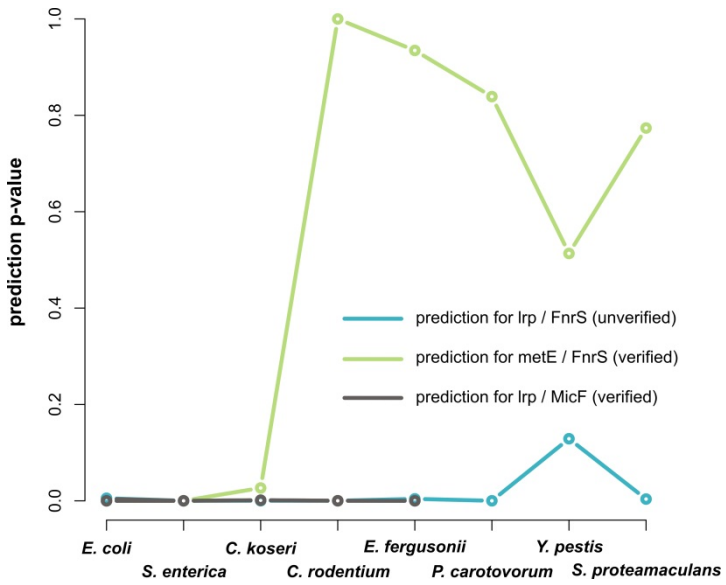

Fig. S12. Example for interaction conservation in the benchmark organisms. The figure shows the p-values of the single organism specific IntaRNA predictions for the *lrp*-FnrS, *lrp*-MicF, and the *metE*-FnrS mRNA-sRNA pairs. The predictions for the *lrp*-FnrS and *lrp*-MicF have low p-values in all benchmark organisms, whereas the *metE*-FnrS interaction has only in *E. coli*, *S. enterica* und *C. koseri* low p-values. For that reason the verified *metE*-FnrS interaction (77) is not detectable by CopraRNA with the given benchmark organism set.

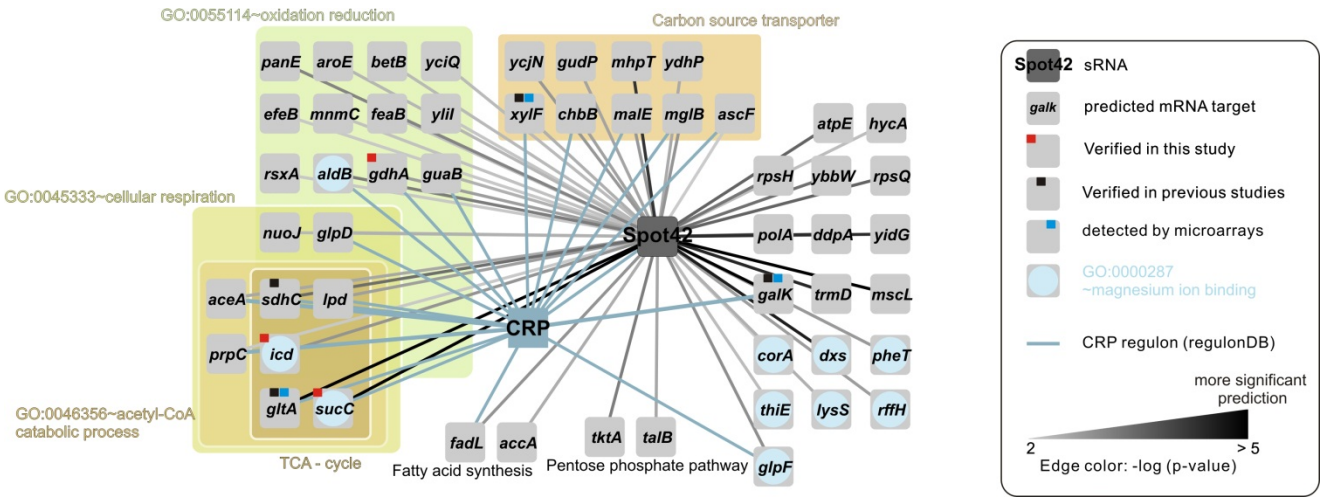

Fig. S13. Functional enrichment of the Spot42 prediction. Visualization of the functional enrichment analysis for Spot42. The network shows also the overlap between the predicted Spot42 regulon and the CRP regulon by blue edges. For further information see **Fig. S7**.

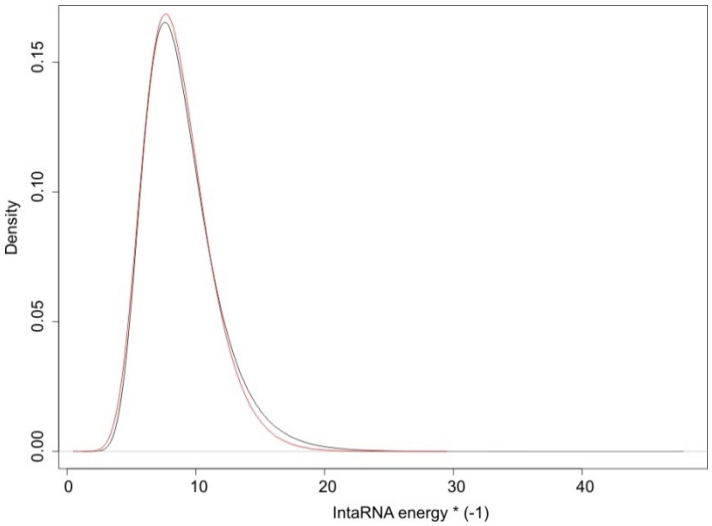

Fig. S14. Fit to extreme value distribution. Fit of IntaRNA energies scores from whole genome target predictions for GcvB on 10 times di-nucleotide shuffled UTRs (red) and un-shuffled UTRs (black) to an extreme value distribution.

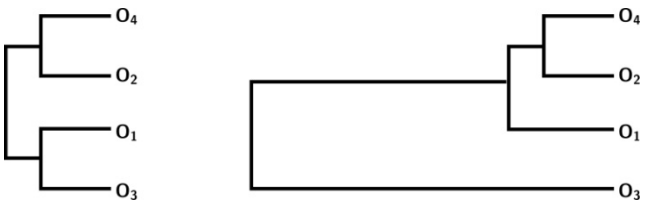

Fig. S15. Phylogenetic trees with different evolutionary distances.

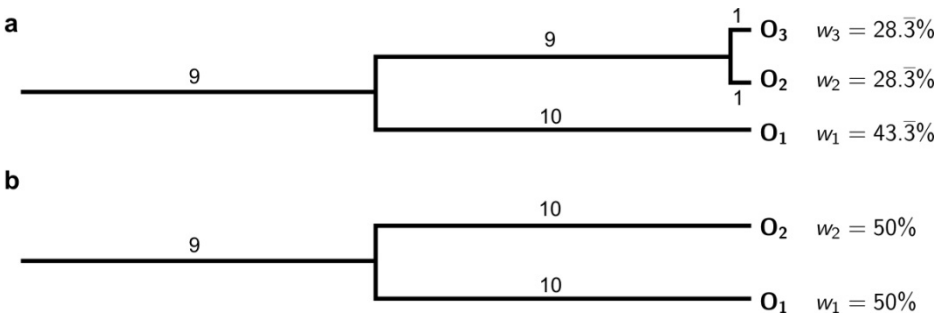

Fig. S16. Different phylogenetic trees and associated weights according to Thompson et al.(60).

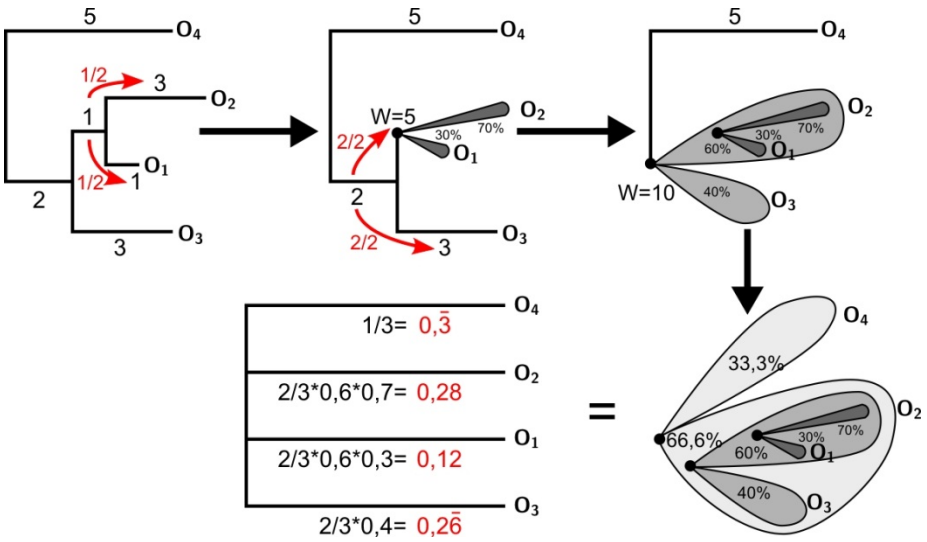

Fig. S17. Recursive weighting of subtrees. Each edge is distributed equally to its subtrees. Thus, the combination of common distance (edge above the subtrees) and the weight of subtrees (sum of weights in the subtree) is defining the relative weight of each subtree.

Help for the CopraRNA webserver

CopraRNA is a tool for sRNA target prediction. It computes whole genome predictions by combination of distinct whole genome IntaRNA predictions. As input, CopraRNA requires at least 3 homologous sRNA sequences from 3 distinct organisms in FASTA format. Furthermore each organisms' genome has to be part of the NCBI Reference Sequence (RefSeq) database (i.e. it should have exactly this format NC_XXXXXX where X stands for a digit between 0 and 9). Depending on sequence length (target and sRNA), amount of input organisms and genome sizes, CopraRNA can take up to 24h to compute (in most cases it is significantly faster). It is suggested you supply your email and return when the job has finished. As output CopraRNA produces a CopraRNA p-value sorted list of putative targets. Results can be viewed in the browser, but closer examination of the downloadable data is suggested.

Input Parameters

Organism selection

The major CopraRNA parameter is the selection of the species, which definitely has an impact on the prediction results. A small evolutionary distance between the species favors sensitivity and a high distance favors specificity. Hence, we suggest selecting 8 sRNA homologs (maximal number of organisms allowed by the webserver) from species with varying evolutionary distance, if there is no availability constraint by the species in which respective sRNA is conserved. For the benchmark we used a blend of close, medium and more remotely conserved species (based on the 16S rDNA sequence, see Fig. S4 in the accompanied publication). In general the maximal evolutionary distance is given by the conservation of the sRNA that is often restricted to a phylum or a class.

sRNA sequences (3-8)

CopraRNA accepts input in form of a multiple FASTA file. A simple example looks like this:

>NC_000913

cccagagguauugauuggugaagucucucaugcgcagguuuuuuuu

>NC_011740

cccagagguauugauucggcacccgcggaugcgcagguuuuuuuu

>NC_003197

cccagagguauugauuggugagauuaggaugcgcagguuuuuuuu

Note the FASTA headers have to represent a RefSeq ID of the according organism! In order to be CopraRNA compatible, an entered organism must be part of the NCBI Reference Sequence (RefSeq) database. This given, an organism has one, or several (depending on the existence of further replicons such as plasmids) RefSeq ID(s) in the following format:

NC_XXXXXX where X stands for a digit between 0 and 9 (NC_000913 for E. coli)

Only one RefSeq ID has to be supplied for each organism. If you supply the ID for a plasmid, the prediction will also be executed on all other replicons of the organism. Vice versa, if you supply the ID of the major replicon, the prediction will also be taken out on all additionally available replicons. IDs such as these NZ_CM001165, NS_000191 are not valid.

To check if the organisms you selected are CopraRNA compatible, check the list of available RefSeq IDs on our homepage. Currently, more than 2400 organisms are CopraRNA compatible. The list is regularly updated.

Please contact us if you know your organism is part of the RefSeq database and has an ID in the NC_XXXXXX format but is not present in the list, or is missing IDs. Then we can run an update.

Input can be given either as direct text input or by uploading a FASTA file. The sequences you upload should be homologous to each other. If you have an sRNA sequence and are trying to find homologs, then you can start by using BLAST. If you don't find anything with BLAST there are more sophisticated methods for this task, such as GotohScan. Furthermore it is also possible that there are no homologs for your sequence. In this case we suggest you resort to the IntaRNA whole genome target prediction webserver.

Extract sequences around

This option allows you to select from which region of the mRNAs you would like to retrieve your putative target sequences. Selecting "start codon" selects regions upstream and downstream (see nt up, nt down) relative to the start codon. The same logic holds if you select "stop codon".

nt up (0-300)

This parameter specifies the number of nucleotides (nt) upstream of your start or stop codon (depending which one you selected). If you selected start codon, and have prior knowledge about average 5'UTR lengths in your input organisms then it is sensible to set nt up to this number in order to increase prediction quality. The sum of nt up and nt down must be at least 140.

nt down (0-300)

This parameter specifies the number of nucleotides (nt) downstream of your start or stop codon (depending which one you selected). If you selected stop codon, and have prior knowledge about average 3'UTR lengths in your input organisms then it is sensible to set nt down to this number in order to increase prediction quality. The sum of nt up and nt down must be at least 140.

Putative target sequences

These are the putative target sequences, extracted from the organism of interest's RefSeq file(s). In most cases their length is nt up + nt down.

Organism of interest

Usually a user has a specific organism he is especially interested in. The organism of interest which is finally selected takes a prime position in the output display and post processing. However it does not change the internal computations of the core CopraRNA algorithm. The online output can only be viewed for the organism of interest. In the downloadable data, all organisms are incorporated, but the functional enrichment of the top candidates is only computed for the organism of interest.

Output Description

Main result:

The main CopraRNA result is a CopraRNA p-value sorted table, of target candidates for the entered homologous sRNAs. The data displayed on the output page of the webserver is comparatively limited, when compared to the downloadable data. For this reason we suggest you download the results for closer inspection.

Positions of interactions:

The positions of the interactions are not relative to start or stop codon, but rather absolute positions with respect to the lengths of your sRNA/mRNA sequence. For example, if you were to extract sequences 200 upstream and 100 downstream of the start codon, the location of your start codon is 201,202,203.

Annotation:

The annotation is retrieved from the RefSeq genome files.

Additional homologs:

In some cases, genes from the same organism can be part of the same cluster of targets. In these cases only the sequence with the best IntaRNA energy score participates in the calculation of the CopraRNA p-value. To secure that no potential targets are lost because of this, the additional homologs are added for the organism of interest.

Regions plots:

These plots are meant to give you an overview of the regions in the target and sRNA sequences that play predominant roles in the statistically significant interactions. The density plot in the top of the image, is calculated from all predicted interactions with a CopraRNA-p-value <= 0.01, while the interactions displayed in the bottom of the image are shown for the top 20 predicted targets. The

different coloring contains no information and is purely intended to increase contrast between different genes.

Interactions:

The interaction you see on the webserver, is the interaction calculated by IntaRNA for the specific candidate you are viewing (the highlighted line in the table). Single interactions can be downloaded for further use. For additional information on how the RNA interactions are computed, please resort to the IntaRNA publication.

Downloadable files:

Main CopraRNA result:

This is a CopraRNA p-value sorted, comma separated table (*.csv), containing all the results for all organisms entered in the analysis. Each column, named by a RefSeq ID, represents the prediction for one organism. The other columns should be self explanatory. See explanation of additional homologs further up in this help. Each line represents one cluster of homologous genes within the organisms entered in the analysis. The content of the cells follows this scheme:

locus_tag(gene name|IntaRNA energy score|IntaRNA p-value|pos. start mRNA|pos. stop mRNA |pos. start sRNA|pos. stop sRNA|Entrez GeneID)

Functional Enrichment:

This file contains the DAVID functional enrichment result for the target candidates up to CopraRNA p-values <= 0.01. A certain term appears as enriched, if it is significantly overrepresented in the top list when compared to the background. The background in this case are all genes for which there is a prediction (not the entire set of genes of an organism). Enrichment scores of 1.3 and higher, suggest statistical significance. However, enrichments also strongly depend on the quality of the annotation of the entered organism of interest. The file is tab delimited. This result is only calculated for the organism of interest.

Regions plots:

These are the same as the ones displayed on the webserver. They can be downloaded in postscript, pdf and png format.

FAQs for the CopraRNA webserver

Other tools for whole genome sRNA target prediction are much faster and do not require previous assembly of homologs. Why should I use CopraRNA?

Truthfully, the runtime of CopraRNA is not excellent and sequence assembly can be tedious. However, the quality of the results outcompetes all other state of the art sRNA target prediction algorithms. Our results show that CopraRNA is even very competitive when compared with the insights gained from micro array analyses. The cost of additional runtime and previous data assembly, is justified by the results being several orders of magnitude better than those computed by other algorithms. Furthermore, CopraRNA is free and fast when compared with microarrays. In some

cases (i.e. GcvB) it allows a complete *in silico* characterization of a certain sRNA's function within the organism.

Why are only organisms supported that are part of the RefSeq database?

In order to guarantee easy usability, CopraRNA requires a certain degree of consistency within the files that it accesses. RefSeq is in most cases a very reliable and consistent database, that meets sensible consistency terms. Find all CopraRNA compatible organisms in this list. Already more than 2000 organisms are CopraRNA compatible.

Why does the target on rank 1 have a p-value = 0 ?

In some cases one of the putative target sequences is encoded on the complementary strand at the same genomic location as the sRNA. In these cases, the complementarity is perfect, which leads to extremely low IntaRNA energy scores and consequently to a p-value of 0. Usually this can be discarded as an artifact. However in some cases it has been shown that sRNAs not only act on trans but also have cis regulatory effects, in which case a putative target with a p-value of 0 should not be disregarded.

What are the q-values and how to interpret them?

The q-values are most easily explained with an example. Assume a q-value cutoff of 0.5. Statistically speaking, 50% of all predicted targets in the list up to this cutoff are assumed to be false positives. The q-value gives you an impression of how many incorrect predictions to expect up to a certain threshold.

When are sRNAs homologous? or Are the sequences I am inputing feasible for CopraRNA?

This is not a trivial question and subject to research in itself. Usually if you find similar sequences of similar lengths with a BLAST search, it is highly likely that the sequences you found are homologous. Yet, if you don't find anything with BLAST this doesn't mean there is nothing to find. In these cases we suggest that you resort to more sophisticated methods to find sRNA homologs, such as GotohScan. Nevertheless, there are cases in which no sRNA homologs exist. In these cases we suggest you resort to an IntaRNA whole genome target prediction.

What are additional homologs*?*

Sometimes the clustering of homologous genes, assigns several genes from one organism to the same cluster. In this case the analysis is only executed on the candidate with the best IntaRNA energy score. In order to prevent losing the other putative targets, they are added at the end as additional homologs.

Are the predictions always good?

Even though we could show that CopraRNA predictions are mostly reliable for Enterobacteria, it is still an *in silico* method and not flawless. You should look at, and think about the output and try to make sense of it, instead of blindly trusting the top list (p-value <=0.01).

Which putative targets should I take a closer look at?

Basically all putative targets with a CopraRNA p-value <=0.01 are statistically speaking interesting. Furthermore putative targets that belong to a certain enriched term are interesting.

Does CopraRNA work for all bacterial and archaeal phyla?

Extensive testing of CopraRNA predictions has so far only been done for enteric bacteria. However, the basic idea is not limited to this branch of microorganisms. It is highly likely that CopraRNA can produce predictions of the same quality for other phyla but it has not yet been experimentally proven.

Is CopraRNA deterministic? It appears your precalculated results are not identical to the results presented in the publication. Why?

Due to the p-value sampling for clusters that do not contain genes from each participating organism, CopraRNA is not a deterministic algorithm. However, usually only slight differences between distinct analyses are to be expected.

Can I download CopraRNA to run batch jobs on my local machine?

The source code for CopraRNA is available from our Software page.

References

1. Busch A, Richter AS, Backofen R (2008) IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics* 24:2849–2856.

2. Eggenhofer F, Tafer H, Stadler PF, Hofacker IL (2011) RNApredator: fast accessibility-based prediction of sRNA targets. *Nucleic Acids Res* 39:W149–W154.

3. Tjaden B (2008) TargetRNA: a tool for predicting targets of small RNA action in bacteria. *Nucleic Acids Res* 36:W109–W113.

4. Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4:44–57.

5. Cline MS et al. (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc* 2:2366–2382.

6. Seemann SE, Richter AS, Gesell T, Backofen R, Gorodkin J (2011) PETcofold: predicting conserved interactions and structures of two multiple alignments of RNA sequences. *Bioinformatics* 27:211–219.

7. Li AX, Marz M, Qin J, Reidys CM (2011) RNA–RNA interaction prediction based on multiple sequence alignments. *Bioinformatics* 27:456–463.

8. Rehmsmeier M, Steffen P, Höchsmann M, Giegerich R (2004) Fast and effective prediction of microRNA/target duplexes. *RNA* 10:1507 –1517.

9. Sharma CM et al. (2011) Pervasive post-transcriptional control of genes involved in amino acid metabolism by the Hfq-dependent GcvB small RNA. *Mol Microbiol* 81:1144–1165.

10. Hartung J (1999) A Note on Combining Dependent Tests of Significance. *Biometrical Journal* 41:849–855.

11. Demetrescu M, Hassler U, Tarcolea A-I (2006) Combining Significance of Correlated Statistics with Application to Panel Data. *Oxford Bulletin of Economics and Statistics* 68:647–663.

12. Delongchamp R, Lee T, Velasco C (2006) A method for computing the overall statistical significance of a treatment effect among a group of genes. *BMC Bioinformatics* 7:S11.

13. Wallace IM, O'Sullivan O, Higgins DG, Notredame C (2006) M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res* 34:1692–1699.

14. Thompson JD, Higgins DG, Gibson TJ (1994) Improved sensitivity of profile searches through the use of sequence weights and gap excision. *Comput Appl Biosci* 10:19–29.

15. Uchiyama I (2007) MBGD: a platform for microbial comparative genomics based on the automated construction of orthologous groups. *Nucleic Acids Res* 35:D343–D346.

16. Huang DW, Sherman BT, Lempicki RA (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4:44–57.

17. Mandin P, Gottesman S (2010) Integrating anaerobic/aerobic sensing and the general stress response through the ArcZ small RNA. *EMBO J* 29:3094–3107.

18. Papenfort K et al. (2009) Specific and pleiotropic patterns of mRNA regulation by ArcZ, a conserved, Hfq-dependent small RNA. *Mol Microbiol* 74:139–158.

19. Figueroa-Bossi N, Valentini M, Malleret L, Bossi L (2009) Caught at its own game: regulatory small RNA inactivated by an inducible transcript mimicking its target. *Genes Dev* 23:2004 –2015.

20. Overgaard M, Johansen J, Møller-Jensen J, Valentin-Hansen P (2009) Switching off small RNA regulation with trap-mRNA. *Mol Microbiol* 73:790–800.

21. Mandin P, Gottesman S (2009) A genetic approach for finding small RNAs regulators of genes of interest identifies RybC as regulating the DpiA/DpiB two-component system. *Mol Microbiol* 72:551–565.

22. Rasmussen AA et al. (2009) A conserved small RNA promotes silencing of the outer membrane protein YbfM. *Mol Microbiol* 72:566–577.

23. De Lay N, Gottesman S (2009) The Crp-Activated Small Noncoding Regulatory RNA CyaR (RyeE) Links Nutritional Status to Group Behavior. *J Bacteriol* 191:461–476.

24. Papenfort K et al. (2008) Systematic deletion of *Salmonella* small RNA genes identifies CyaR, a conserved CRP-dependent riboregulator of OmpX synthesis. *Mol Microbiol* 68:890–906.

25. Lease RA, Cusick ME, Belfort M (1998) Riboregulation in *Escherichia coli*: DsrA RNA acts by RNA:RNA interactions at multiple loci. *Proc Natl Acad Sci USA* 95:12456–12461.

26. Boysen A, Møller-Jensen J, Kallipolitis B, Valentin-Hansen P, Overgaard M (2010) Translational Regulation of Gene Expression by an Anaerobically Induced Small Non-Coding RNA in *Escherichia Coli*. *J Biol Chem* 285:10690–10702.

27. Durand S, Storz G (2010) Reprogramming of anaerobic metabolism by the FnrS small RNA. *Mol Microbiol* 75:1215–1231.

28. Pulvermacher SC, Stauffer LT, Stauffer GV (2009) The Small RNA GcvB Regulates *sstT* mRNA Expression in *Escherichia coli*. *J Bacteriol* 191:238–248.

29. Pulvermacher SC, Stauffer LT, Stauffer GV (2009) Role of the sRNA GcvB in regulation of *cycA* in *Escherichia coli*. *Microbiology* 155:106–114.

30. Sharma CM, Darfeuille F, Plantinga TH, Vogel J (2007) 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.

31. Jørgensen MG et al. (2012) Small regulatory RNAs control the multi-cellular adhesive lifestyle of *Escherichia coli*. *Mol Microbiol* 84:36–50.

32. Urban JH, Vogel J (2008) Two Seemingly Homologous Noncoding RNAs Act Hierarchically to Activate *glmS* mRNA Translation. *PLoS Biol* 6:e64.

33. Bossi L, Figueroa-Bossi N (2007) A small RNA downregulates LamB maltoporin in *Salmonella*. *Mol Microbiol* 65:799–810.

34. Coornaert A et al. (2010) MicA sRNA links the PhoP regulon to cell envelope stress. *Mol Microbiol* 76:467–479.

35. Gogol EB, Rhodius VA, Papenfort K, Vogel J, Gross CA (2011) Small RNAs endow a transcriptional activator with essential repressor functions for single-tier control of a global stress regulon. *Proc Natl Acad Sci USA* 108:12875 –12880.

36. Udekwu KI et al. (2005) Hfq-Dependent Regulation of OmpA Synthesis Is Mediated by an Antisense RNA. *Genes Dev* 19:2355–2366.

37. Chen S, Zhang A, Blyn LB, Storz G (2004) MicC, a Second Small-RNA Regulator of Omp Protein Expression in *Escherichia coli*. *J Bacteriol* 186:6689–6697.

38. Pfeiffer V, Papenfort K, Lucchini S, Hinton JCD, Vogel J (2009) Coding sequence targeting by MicC RNA reveals bacterial mRNA silencing downstream of translational initiation. *Nat Struct Mol Biol* 16:840–846.

39. Corcoran CP et al. (2012) Superfolder GFP reporters validate diverse new mRNA targets of the classic porin regulator, MicF RNA. *Mol Microbiol* 84:428–45.

40. Holmqvist E, Unoson C, Reimegård J, Wagner EGH (2012) A mixed double negative feedback loop between the sRNA MicF and the global regulator Lrp. *Mol Microbiol* 84:414–27.

41. Suzuki T, Ueguchi C, Mizuno T (1996) H-NS regulates OmpF expression through *micF* antisense RNA in *Escherichia coli*. *J Bacteriol* 178:3650–3653.

42. Holmqvist E et al. (2010) Two antisense RNAs target the transcriptional regulator CsgD to inhibit curli synthesis. *EMBO J* 29:1840–1850.

43. Guillier M, Gottesman S (2008) The 5’ end of two redundant sRNAs is involved in the regulation of multiple targets, including their own regulator. *Nucl Acids Res* 36:6781–6794.

44. Altuvia S, Zhang A, Argaman L, Tiwari A, Storz G (1998) The *Escherichia coli* OxyS regulatory RNA represses fhlA translation by blocking ribosome binding. *EMBO J* 17:6069–6075.

45. Majdalani N, Hernandez D, Gottesman S (2002) Regulation and mode of action of the second small RNA activator of RpoS translation, RprA. *Mol Microbiol* 46:813–826.

46. Mika F et al. (2012) Targeting of *csgD* by the small regulatory RNA RprA links stationary phase, biofilm formation and cell envelope stress in *Escherichia coli*. *Mol Microbiol* 84:51–65.

47. Papenfort K, Bouvier M, Mika F, Sharma CM, Vogel J (2010) Evidence for an Autonomous 5' Target Recognition Domain in an Hfq-Associated Small RNA. *Proc Natl Acad Sci USA* 107:20435–20440.

48. Bouvier M, Sharma CM, Mika F, Nierhaus KH, Vogel J (2008) Small RNA Binding to 5' mRNA Coding Region Inhibits Translational Initiation. *Mol Cell* 32:827–837.

49. Balbontín R, Fiorini F, Figueroa-Bossi N, Casadesús J, Bossi L (2010) Recognition of heptameric seed sequence underlies multi-target regulation by RybB small RNA in Salmonella enterica. *Mol Microbiol* 78:380–394.

50. Desnoyers G, Massé E (2012) Noncanonical repression of translation initiation through small RNA recruitment of the RNA chaperone Hfq. *Genes Dev* 26:726–739.

51. Johansen J, Rasmussen AA, Overgaard M, Valentin-Hansen P (2006) Conserved Small Non-coding RNAs that belong to the σ^E Regulon: Role in Down-regulation of Outer Membrane Proteins. *J Mol Biol* 364:1–8.

52. Salvail H et al. (2010) A small RNA promotes siderophore production through transcriptional and metabolic remodeling. *Proc Natl Acad Sci USA* 107:15223–8.

53. Desnoyers G, Morissette A, Prévost K, Massé E (2009) Small RNA-induced differential degradation of the polycistronic mRNA *iscRSUA*. *EMBO J* 28:1551–1561.

54. Večerek B, Moll I, Bläsi U (2007) Control of Fur synthesis by the non-coding RNA RyhB and iron-responsive decoding. *EMBO J* 26:965–975.

55. Prévost K et al. (2007) The small RNA RyhB activates the translation of *shiA* mRNA encoding a permease of shikimate, a compound involved in siderophore synthesis. *Mol Microbiol* 64:1260–1273.

56. Večerek B, Moll I, Afonyushkin T, Kaberdin V, Bläsi U (2003) Interaction of the RNA chaperone Hfq with mRNAs: direct and indirect roles of Hfq in iron metabolism of *Escherichia coli*. *Mol Microbiol* 50:897–909.

57. Prévost K, Desnoyers G, Jacques J-F, Lavoie F, Massé E (2011) Small RNA-induced mRNA degradation achieved through both translation block and activated cleavage. *Genes Dev* 25:385–96.

58. Kawamoto H, Koide Y, Morita T, Aiba H (2006) Base-pairing requirement for RNA silencing by a bacterial small RNA and acceleration of duplex formation by Hfq. *Mol Microbiol* 61:1013–1022.

59. Rice JB, Vanderpool CK (2011) The small RNA SgrS controls sugar–phosphate accumulation by regulating multiple PTS genes. *Nucleic Acids Res* 39:3806 –3819.

60. Papenfort K, Podkaminski D, Hinton JCD, Vogel J (2012) The ancestral SgrS RNA discriminates horizontally acquired *Salmonella* mRNAs through a single G-U wobble pair. *Proc Natl Acad Sci U S A* 109:E757–E764.

61. Papenfort K, Sun Y, Miyakoshi M, Vanderpool CK, Vogel J (2013) Small RNA-Mediated Activation of Sugar Phosphatase mRNA Regulates Glucose Homeostasis. *Cell* 153:426–437.

62. Møller T, Franch T, Udesen C, Gerdes K, Valentin-Hansen P (2002) Spot 42 RNA mediates discoordinate expression of the *E. coli* galactose operon. *Genes Dev* 16:1696–1706.

63. Beisel CL, Storz G (2011) The Base-Pairing RNA Spot 42 Participates in a Multioutput Feedforward Loop to Help Enact Catabolite Repression in *Escherichia coli*. *Mol Cell* 41:286–297.

64. Urban JH, Vogel J (2007) Translational control and target recognition by *Escherichia coli* small RNAs in vivo. *Nucleic Acids Res* 35:1018–1037.

65. Tattersall J et al. (2012) Translation Inhibition of the Developmental Cycle Protein HctA by the Small RNA IhtA Is Conserved across *Chlamydia*. *PLoS ONE* 7:e47439.

66. Wilderman PJ et al. (2004) Identification of tandem duplicate regulatory small RNAs in *Pseudomonas aeruginosa* involved in iron homeostasis. *Proc Natl Acad Sci USA* 101:9792–9797.

67. Oglesby AG et al. (2008) The Influence of Iron on *Pseudomonas aeruginosa* Physiology *A REGULATORY LINK BETWEEN IRON AND QUORUM SENSING*. *J Biol Chem* 283:15558–15567.

68. Nielsen JS et al. (2011) A Small RNA Controls Expression of the Chitinase ChiA in *Listeria monocytogenes*. *PLoS ONE* 6:e19019.

69. Nielsen JS et al. (2010) Defining a role for Hfq in Gram-positive bacteria: evidence for Hfq-dependent antisense regulation in *Listeria monocytogenes*. *Nucl Acids Res* 38:907–919.

70. Smaldone GT, Antelmann H, Gaballa A, Helmann JD (2012) The FsrA sRNA and FbpB Protein Mediate the Iron-Dependent Induction of the *Bacillus subtilis* LutABC Iron-Sulfur-Containing Oxidases. *J Bacteriol* 194:2586–2593.

71. Gaballa A et al. (2008) The *Bacillus subtilis* iron-sparing response is mediated by a Fur-regulated small RNA and three small, basic proteins. *Proc Natl Acad Sci USA* 105:11927–11932.

72. Heidrich N, Chinali A, Gerth U, Brantl S (2006) The small untranslated RNA SR1 from the *Bacillus subtilis* genome is involved in the regulation of arginine catabolism. *Mol Microbiol* 62:520–536.

73. 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.

74. Richter AS, Backofen R (2012) Accessibility and conservation: General features of bacterial small RNA-mRNA interactions? *RNA Biol* 9:954–965.

75. Smith C, Heyne S, Richter AS, Will S, Backofen R (2010) Freiburg RNA Tools: a web server integrating INTARNA, EXPARNA and LOCARNA. *Nucleic Acids Res* 38:W373–W377.

76. Bernhart SH, Hofacker IL, Will S, Gruber AR, Stadler PF (2008) RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics* 9:474.

77. Boysen A, Møller-Jensen J, Kallipolitis B, Valentin-Hansen P, Overgaard M (2010) Translational Regulation of Gene Expression by an Anaerobically Induced Small Non-Coding RNA in *Escherichia Coli*. *J Biol Chem* 285:10690–10702.