Supplementary Material

PETcofold: Predicting conserved interactions and structures of two multiple alignments of RNA sequences

Stefan E. Seemann *, Andreas S. Richter *, Tanja Gesell,
Rolf Backofen † and Jan Gorodkin †

This supplement contains:

Supplementary Information
Supplementary Tables S1 to S3
Supplementary Figures S1 to S3
Supplementary References

*these authors contributed equally to this work
†to whom correspondence should be addressed

Supplementary Information

For this study, the data set of experimentally verified bacterial sRNA-mRNA interactions from Busch *et al.* (2008) was extended by further sRNA-mRNA interactions with experimental support from literature. The extended data set consisted of 15 different sRNAs and 41 interactions from the 3 bacterial species *Escherichia coli* K12 (*E. coli*), *Salmonella typhimurium* LT2 (*Salmonella*) and *Staphylococcus aureus* N315 (*S. aureus*).

To obtain homology information, the RNA family sequence alignment of each sRNA was downloaded from the Rfam database 9.1 (Gardner *et al.*, 2009). However, analysis of the sequence data showed that the Rfam families of some sRNAs miss part of the sequence that is involved in the sRNA-target interaction. This applies to the interactions RyhB-*sdhCDAB* (Massé and Gottesman, 2002), RyhB-*shiA* (Prévost *et al.*, 2007) and Spot42-*galK* (Møller *et al.*, 2002) in *E. coli* and RybB-*ompN* (Bouvier *et al.*, 2008) in *Salmonella*. Therefore, they were excluded in this study. In addition, the interactions of the *E. coli* sRNA OmrB with its targets *cirA*, *ompR* and *ompT* (Guillier and Gottesman, 2008) were neglected, because the final data set already contained interactions of these genes with the sRNA OmrA, which belongs, together with OmrB, to one Rfam family. Finally, two different interactions of the *S. aureus* sRNA RNAIII with its target *rot* (Boisset *et al.*, 2007) were excluded from the data set because they contain crossing interactions, which cannot be predicted by the current version of PETcofold. The final data set of 32 sRNA-mRNA interactions that was used in the study is given in Supplementary Table S1.

Alignments of the target genes and their orthologous were created by the following procedure: First, genome sequences for all 69 species with available complete genome according to the Rfam annotation were downloaded from the EMBL Nucleotide Sequence Database (Cochrane *et al.*, 2009). DNA sequences of all annotated proteins were extracted from each genome. Then the OrthoMCL tool (Li *et al.*, 2003) was used with default parameters to identify groups of orthologous genes separately for the *S. aureus* species (from Rfam family of RNAIII) and all remaining species (from Rfam families of *E. coli* and *Salmonella* sRNAs). For each target gene, we created a set of 250 nt long orthologous target sequences (150 nt upstream and 100 nt downstream of the annotated start codons) according to the OrthoMCL prediction. Target sequences of species that are not contained in the Rfam family of its interaction partner were excluded from the sequence sets. The subsequences of 250 nt length were chosen because all interactions are in range from -132 to +56 relative to the start codon. Flanking regions were included for prediction of mRNA structure and for compensation of misannotated translational start sites. Finally, the sets of targets were locally aligned on sequence level with MAFFT version 6 (using option E-INS-i for generalised affine gap costs) (Katoh and Toh, 2008).

For the final sRNA sequence alignments, all sequences without available complete genome and without a predicted ortholog in the target sequence set were removed from the Rfam sequence alignments followed by removal of gap-only alignment columns.

The resulting data set with sRNAs from Rfam and mRNAs from orthology prediction was processed twofold. mRNA sequences that were very distant from the mRNA sequence of the reference organism were likely to be false positive predicted orthologs and, thus, should be excluded. This was achieved by removing all mRNA sequences from the mRNA alignment that had, in comparison to the reference organism, a pairwise sequence identity at the interaction site and 10 nt of the flanking sequences (PI^{int}_{ref}) below a given threshold. Here PI^{int}_{ref} thresholds of either 40%, 50% or 60% were applied. Redundant sequences were excluded from the data set to avoid a bias by overweighting similar sequence information. To this end, mRNA sequences were clustered with the tool BLASTClust. BLASTClust is part of the standalone BLAST package (Altschul *et al.*, 1997) and is typically used to create non-redundant sequence sets. BLASTClust was called for nucleotide input with a word size of 8 and a percent identity threshold of 100% over an area

covering 90% of each sequence. From each cluster of similar sequences, the sequence with lowest $\mathrm{PI}_{\mathrm{ref}}^{\mathrm{int}}$ was taken. If an mRNA was removed from an mRNA alignment, then the corresponding sRNA, i.e., the sRNA of the same organism that interacts with the mRNA, was removed from the sRNA alignment.

Supplementary Tables

Table S1: Data set of bacterial sRNAs and their target mRNAs used in this study. Rfam acc. denotes the Rfam accession number.

sRNA	Rfam acc.	Target	Organism	Reference
CyaR	RF00112	*luxS*	*E. coli*	De Lay and Gottesman (2009)
CyaR	RF00112	*nadE*	*E. coli*	De Lay and Gottesman (2009)
CyaR	RF00112	*ompX*	*E. coli*	De Lay and Gottesman (2009)
CyaR	RF00112	*yqaE*	*E. coli*	De Lay and Gottesman (2009)
CyaR	RF00112	*ompX*	*Salmonella*	Papenfort *et al.* (2008)
DsrA	RF00014	*hns*	*E. coli*	Lease *et al.* (1998)
DsrA	RF00014	*rpoS*	*E. coli*	Majdalani *et al.* (1998)
GcvB	RF00022	*sstT*	*E. coli*	Pulvermacher *et al.* (2009)
GcvB	RF00022	*argT*	*Salmonella*	Sharma *et al.* (2007)
GcvB	RF00022	*dppA*	*Salmonella*	Sharma *et al.* (2007)
GcvB	RF00022	*gltI*	*Salmonella*	Sharma *et al.* (2007)
GcvB	RF00022	*livJ*	*Salmonella*	Sharma *et al.* (2007)
GcvB	RF00022	*livK*	*Salmonella*	Sharma *et al.* (2007)
GcvB	RF00022	*oppA*	*Salmonella*	Sharma *et al.* (2007)
GcvB	RF00022	STM4351	*Salmonella*	Sharma *et al.* (2007)
GlmZ	RF00083	*glmS*	*E. coli*	Urban and Vogel (2008)
MicA	RF00078	*ompA*	*E. coli*	Udekwu *et al.* (2005)
MicA	RF00078	*lamB*	*Salmonella*	Bossi and Figueroa-Bossi (2007)
MicC	RF00121	*ompC*	*E. coli*	Chen *et al.* (2004)
MicC	RF00121	*ompD*	*Salmonella*	Pfeiffer *et al.* (2009)
MicF	RF00033	*ompF*	*E. coli*	Schmidt *et al.* (1995)
OmrA	RF00079	*cirA*	*E. coli*	Guillier and Gottesman (2008)
OmrA	RF00079	*ompR*	*E. coli*	Guillier and Gottesman (2008)
OmrA	RF00079	*ompT*	*E. coli*	Guillier and Gottesman (2008)
OxyS	RF00035	*fhlA*	*E. coli*	Argaman and Altuvia (2000)
RNAIII	RF00503	SA1000	*S. aureus*	Boisset *et al.* (2007)
RNAIII	RF00503	SA2353	*S. aureus*	Boisset *et al.* (2007)
RNAIII	RF00503	*spa*	*S. aureus*	Huntzinger *et al.* (2005)
RprA	RF00034	*rpoS*	*E. coli*	Majdalani *et al.* (2002)
RyhB	RF00057	*uof-fur*	*E. coli*	Večerek *et al.* (2007)
RyhB	RF00057	*sodB*	*E. coli*	Geissmann and Touati (2004)
SgrS	RF00534	*ptsG*	*E. coli*	Kawamoto *et al.* (2006)

Table S2: Prediction performance of PETcofold on data sets with 32 interactions. PI^{int}_{ref} denotes the minimal pairwise interaction site sequence identity to the reference. The predictions were evaluated by their mean interaction MCC. PETcofold was run with values ranging from 0.0 to 1.0 in steps of 0.1 for the intramolecular base pair reliability threshold (δ) and the minimal partial structure probability (γ) either (a) without the option *-noLP* or (b) with *-noLP*. Tables (a) and (b) list the best mean MCC of the 121 parameter combinations, the corresponding median MCC and the values yielding to the best mean MCC. (c) Average number of compensatory interaction base pair exchanges (CBP) of the data sets.

(a)

PETcofold without option *-noLP*			
PI^{int}_{ref}	Best mean MCC	Median MCC	δ/γ
40%	0.507	0.564	0.9/0.9
50%	0.510	0.564	0.9/0.9
60%	0.494	0.546	0.9/[0.0,...,0.5]

(b)

PETcofold with option *-noLP*			
PI^{int}_{ref}	Best mean MCC	Median MCC	δ/γ
40%	0.504	0.561	0.9/0.9
50%	0.511	0.583	0.9/0.9
60%	0.491	0.526	0.9/0.9

(c)

PI^{int}_{ref}	CBP
40%	0.195
50%	0.171
60%	0.114

Table S3: Performance of PETcofold on prediction of four sRNA-mRNA joint secondary structures. PETcofold was run with a value of 0.9 for the intramolecular base pair reliability threshold (δ), varying values for the minimal partial structure probability (γ), with the option *-noLP* and optionally with the option *-extstem*.

	MCC of joint secondary structure										
sRNA-target pair	PETcofold without *-extstem*						PETcofold with *-extstem*				
γ	0.1	0.3	0.5	0.7	0.9		0.1	0.3	0.5	0.7	0.9
MicA-ompA	0.87	0.87	0.87	0.87	0.87		0.83	0.83	0.83	0.83	0.83
OxyS-fhlA	0.80	0.80	0.80	0.71	0.71		0.82	0.82	0.71	0.71	0.71
RyhB-fur	0.13	0.13	0.13	0.13	0.12		0.13	0.13	0.13	0.13	0.12
RyhB-sodB	0.67	0.67	0.67	0.67	0.67		0.71	0.71	0.71	0.70	0.70
Average	0.62	0.62	0.62	0.60	0.59		0.62	0.62	0.60	0.59	0.59

Supplementary Figures

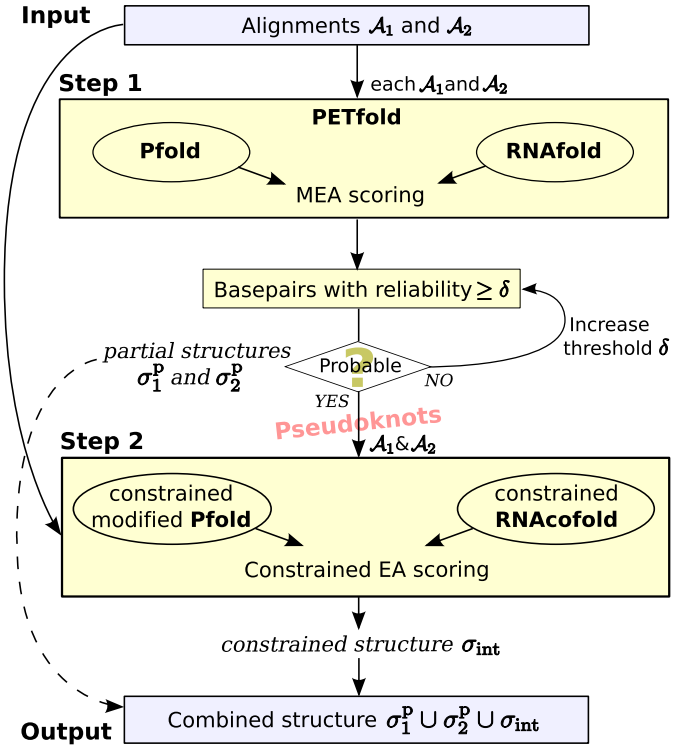

Figure S1: The flow chart diagram is an adapted version of Figure 1 showing the different programs integrated in PETcofold, which are Pfold (Knudsen and Hein, 2003), RNAfold (Hofacker *et al.*, 1994), PETfold (Seemann *et al.*, 2008) and RNAcofold (Bernhart *et al.*, 2006). *MEA* denotes maximum expected accuracy and *EA* denotes expected accuracy.

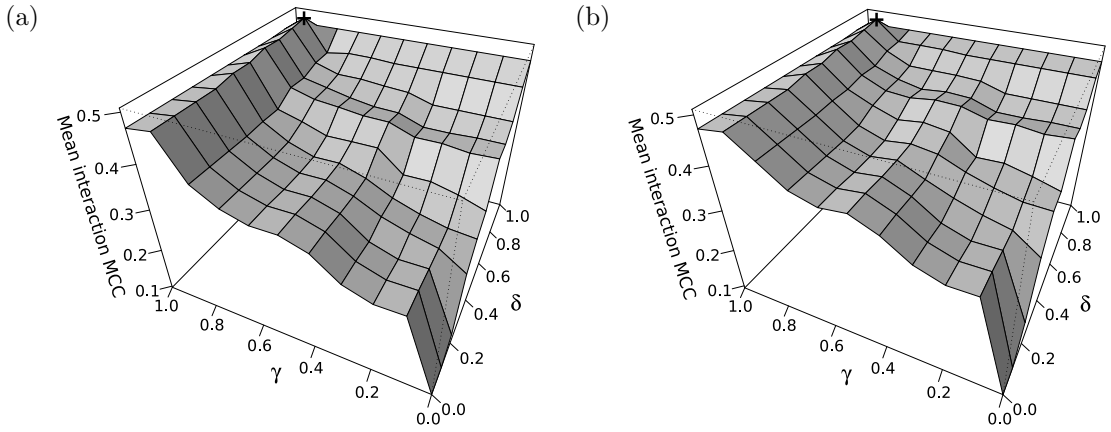

Figure S2: Performance of PETcofold while varying the parameters δ (maximal intramolecular base pair reliability) and γ (minimal partial structure probability). The 3D plot shows the mean MCC of 32 interactions using input sequences with (a) 40% and (b) 50% minimal pairwise interaction site sequence identity to the reference. Predictions were carried out without the option *-noLP*. The maximal MCC is marked with "+".

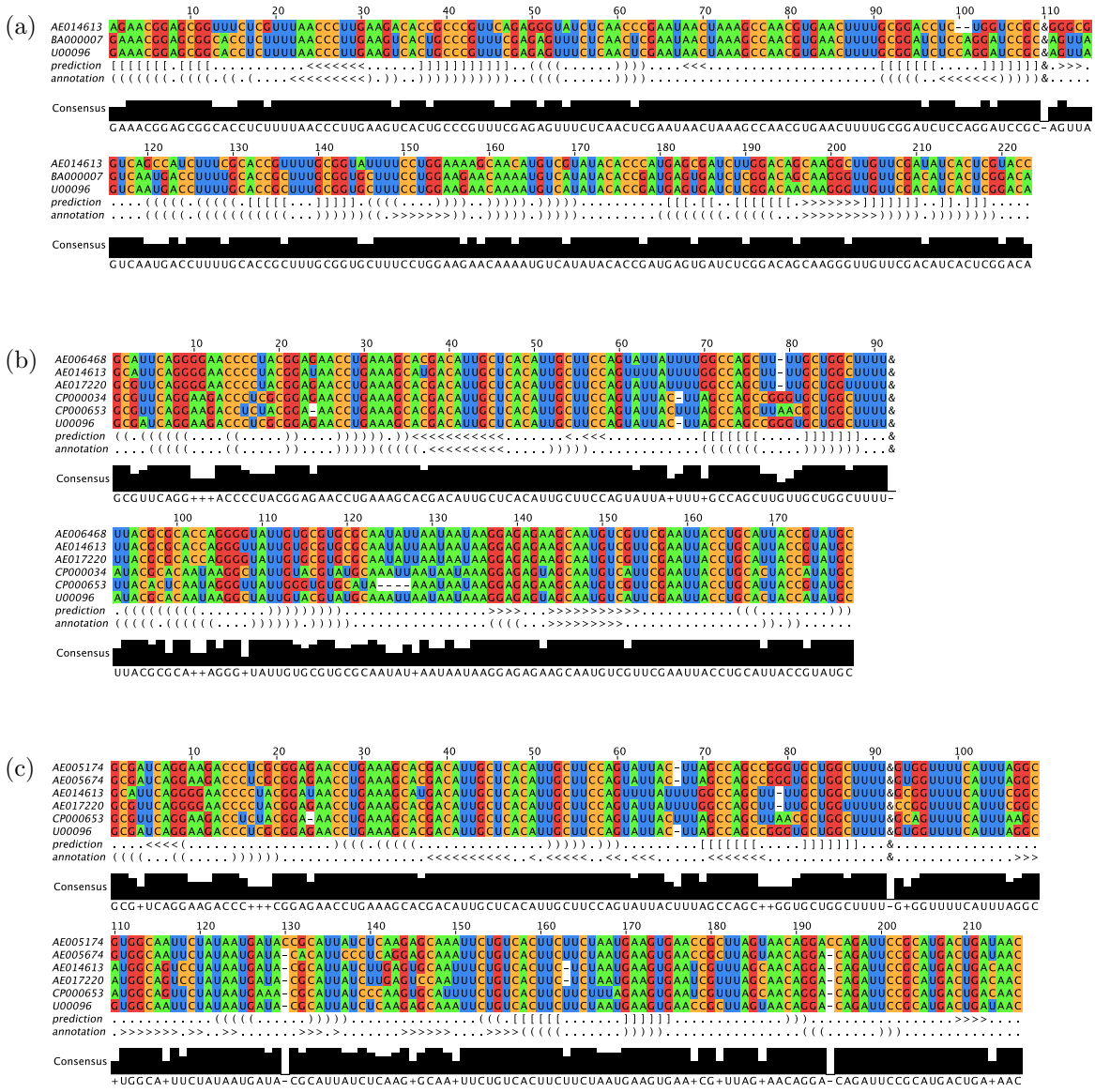

Figure S3: Joint secondary structure of the sRNA-mRNA interaction complex of (a) OxyS-*fhlA*, (b) RyhB-*sodB* and (c) RyhB-*uof-fur*. Each sequence alignment shows the two input alignments concatenated by the linker symbol "&", the joint structure predicted by PETcofold (with parameters $\delta = 0.9$, $\gamma = 0.1$, and options *-noLP* and *-extstem*) and the interaction complex model from literature (Geissmann and Touati, 2004; Večerek *et al.*, 2007). Sequences are labelled with the genome accession numbers of the corresponding organisms. Angle brackets indicate intermolecular base pairs between the two RNAs. Round and square brackets indicate intramolecular base pairs. Square brackets indicate positions that are constrained in step 1 of the PETcofold pipeline. For OxyS-*fhlA*, only columns with $< 50\%$ gaps are shown. The alignments were visualised with Jalview (Waterhouse *et al.*, 2009).

Supplementary References

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**(17), 3389–402.

Argaman, L. and Altuvia, S. (2000). *fhlA* repression by OxyS RNA: kissing complex formation at two sites results in a stable antisense-target RNA complex. *Journal of Molecular Biology*, **300**(5), 1101–12.

Bernhart, S. H., Tafer, H., Mückstein, U., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2006). Partition function and base pairing probabilities of RNA heterodimers. *Algorithms Mol Biol*, **1**(1), 3.

Boisset, S., Geissmann, T., Huntzinger, E., Fechter, P., Bendridi, N., Possedko, M., Chevalier, C., Helfer, A. C., Benito, Y., Jacquier, A., Gaspin, C., Vandenesch, F., and Romby, P. (2007). *Staphylococcus aureus* RNAIII coordinately represses the synthesis of virulence factors and the transcription regulator Rot by an antisense mechanism. *Genes Dev*, **21**(11), 1353–66.

Bossi, L. and Figueroa-Bossi, N. (2007). A small RNA downregulates LamB maltoporin in *Salmonella*. *Mol Microbiol*, **65**(3), 799–810.

Bouvier, M., Sharma, C. M., Mika, F., Nierhaus, K. H., and Vogel, J. (2008). Small RNA binding to 5' mRNA coding region inhibits translational initiation. *Mol Cell*, **32**(6), 827–37.

Busch, A., Richter, A. S., and Backofen, R. (2008). IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*, **24**(24), 2849–56.

Chen, S., Zhang, A., Blyn, L. B., and Storz, G. (2004). MicC, a second small-RNA regulator of Omp protein expression in *Escherichia coli*. *J Bacteriol*, **186**(20), 6689–97.

Cochrane, G., Akhtar, R., Bonfield, J., Bower, L., Demiralp, F., Faruque, N., Gibson, R., Hoad, G., Hubbard, T., Hunter, C., Jang, M., Juhos, S., Leinonen, R., Leonard, S., Lin, Q., Lopez, R., Lorenc, D., McWilliam, H., Mukherjee, G., Plaister, S., Radhakrishnan, R., Robinson, S., Sobhany, S., Hoopen, P. T., Vaughan, R., Zalunin, V., and Birney, E. (2009). Petabyte-scale innovations at the European Nucleotide Archive. *Nucleic Acids Research*, **37**(Database issue), D19–25.

De Lay, N. and Gottesman, S. (2009). The Crp-activated small noncoding regulatory RNA CyaR (RyeE) links nutritional status to group behavior. *J Bacteriol*, **191**(2), 461–76.

Gardner, P. P., Daub, J., Tate, J. G., Nawrocki, E. P., Kolbe, D. L., Lindgreen, S., Wilkinson, A. C., Finn, R. D., Griffiths-Jones, S., Eddy, S. R., and Bateman, A. (2009). Rfam: updates to the RNA families database. *Nucleic Acids Research*, **37**(Database issue), D136–40.

Geissmann, T. A. and Touati, D. (2004). Hfq, a new chaperoning role: binding to messenger RNA determines access for small RNA regulator. *EMBO J*, **23**(2), 396–405.

Guillier, M. and Gottesman, S. (2008). The 5' end of two redundant sRNAs is involved in the regulation of multiple targets, including their own regulator. *Nucleic Acids Research*.

Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, S., Tacker, M., and Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Monatshefte Chemie*, **125**, 167–188.

Huntzinger, E., Boisset, S., Saveanu, C., Benito, Y., Geissmann, T., Namane, A., Lina, G., Etienne, J., Ehresmann, B., Ehresmann, C., Jacquier, A., Vandenesch, F., and Romby, P. (2005). *Staphylococcus aureus* RNAIII and the endoribonuclease III coordinately regulate *spa* gene expression. *EMBO J*, **24**(4), 824–35.

Katoh, K. and Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform*, **9**(4), 286–98.

Kawamoto, H., Koide, Y., Morita, T., and Aiba, H. (2006). Base-pairing requirement for RNA silencing by a bacterial small RNA and acceleration of duplex formation by Hfq. *Mol Microbiol*, **61**(4), 1013–22.

Knudsen, B. and Hein, J. (2003). Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Research*, **31**(13), 3423–8.

Lease, R. A., Cusick, M. E., and Belfort, M. (1998). Riboregulation in *Escherichia coli*: DsrA RNA acts by RNA:RNA interactions at multiple loci. *Proc. Natl. Acad. Sci. USA*, **95**(21), 12456–61.

Li, L., Stoeckert, C. J. J., and Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*, **13**(9), 2178–89.

Majdalani, N., Cunning, C., Sledjeski, D., Elliott, T., and Gottesman, S. (1998). DsrA RNA regulates translation of RpoS message by an anti-antisense mechanism, independent of its action as an antisilencer of transcription. *Proc. Natl. Acad. Sci. USA*, **95**(21), 12462–7.

Majdalani, N., Hernandez, D., and Gottesman, S. (2002). Regulation and mode of action of the second small RNA activator of RpoS translation, RprA. *Mol Microbiol*, **46**(3), 813–26.

Massé, E. and Gottesman, S. (2002). A small RNA regulates the expression of genes involved in iron metabolism in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA*, **99**(7), 4620–5.

Møller, T., Franch, T., Udesen, C., Gerdes, K., and Valentin-Hansen, P. (2002). Spot 42 RNA mediates discoordinate expression of the *E. coli* galactose operon. *Genes Dev*, **16**(13), 1696–706.

Papenfort, K., Pfeiffer, V., Lucchini, S., Sonawane, A., Hinton, J. C. D., and Vogel, J. (2008). Systematic deletion of *Salmonella* small RNA genes identifies CyaR, a conserved CRP-dependent riboregulator of OmpX synthesis. *Mol Microbiol*, **68**(4), 890–906.

Pfeiffer, V., Papenfort, K., Lucchini, S., Hinton, J. C. D., and Vogel, J. (2009). Coding sequence targeting by MicC RNA reveals bacterial mRNA silencing downstream of translational initiation. *Nat Struct Mol Biol*, **16**(8), 840–6.

Prévost, K., Salvail, H., Desnoyers, G., Jacques, J.-F., Phaneuf, É., and Massé, E. (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**(5), 1260–73.

Pulvermacher, S. C., Stauffer, L. T., and Stauffer, G. V. (2009). The small RNA GcvB regulates *sstT* mRNA expression in *Escherichia coli*. *J Bacteriol*, **191**(1), 238–48.

Schmidt, M., Zheng, P., and Delihas, N. (1995). Secondary structures of *Escherichia coli* antisense *micF* RNA, the 5'-end of the target *ompF* mRNA, and the RNA/RNA duplex. *Biochemistry*, **34**(11), 3621–31.

Seemann, S. E., Gorodkin, J., and Backofen, R. (2008). Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments. *Nucleic Acids Research*, **36**(20), 6355–62.

Sharma, C. M., Darfeuille, F., Plantinga, T. H., and 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**(21), 2804–17.

Udekwu, K. I., Darfeuille, F., Vogel, J., Reimegård, J., Holmqvist, E., and Wagner, E. G. H. (2005). Hfq-dependent regulation of OmpA synthesis is mediated by an antisense RNA. *Genes Dev*, **19**(19), 2355–66.

Urban, J. H. and Vogel, J. (2008). Two seemingly homologous noncoding RNAs act hierarchically to activate *glmS* mRNA translation. *PLoS Biol*, **6**(3), e64.

Večerek, B., Moll, I., and Bläsi, U. (2007). Control of Fur synthesis by the non-coding RNA RyhB and iron-responsive decoding. *EMBO J*, **26**(4), 965–75.

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., and Barton, G. J. (2009). Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**(9), 1189–91.