Prediction of CsrA-regulating small RNAs in bacteria and their experimental verification in *Vibrio fischeri*

Prajna R. Kulkarni, Xiaohui Cui[1], Joshua W. Williams[1], Ann M. Stevens[1] and Rahul V. Kulkarni*

Department of Physics and [1]Department of Biological Sciences, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA

ABSTRACT

The role of small RNAs as critical components of global regulatory networks has been highlighted by several recent studies. An important class of such small RNAs is represented by CsrB and CsrC of *Escherichia coli*, which control the activity of the global regulator CsrA. Given the critical role played by CsrA in several bacterial species, an important problem is the identification of CsrA-regulating small RNAs. In this paper, we develop a computer program (CSRNA_FIND) designed to locate potential CsrA-regulating small RNAs in bacteria. Using CSRNA_FIND to search the genomes of bacteria having homologs of CsrA, we identify all the experimentally known CsrA-regulating small RNAs and also make predictions for several novel small RNAs. We have verified experimentally our predictions for two CsrA-regulating small RNAs in *Vibrio fischeri*. As more genomes are sequenced, CSRNA_FIND can be used to locate the corresponding small RNAs that regulate CsrA homologs. This work thus opens up several avenues of research in understanding the mode of CsrA regulation through small RNAs in bacteria.

INTRODUCTION

Recent studies combining bioinformatic and experimental approaches have led to the discovery of numerous small non-coding RNAs (sRNAs) in bacteria (1–6). Although the functions for a majority of these sRNAs are yet to be determined, an emerging trend is that they play crucial regulatory roles in bacterial adaptation to changing environments (7). In particular, sRNAs have been shown to be critical components of global regulatory networks which coordinate large-scale changes in gene expression (8,9). Further identification and analysis of sRNAs as components of such regulatory networks will aid efforts to elucidate their roles in mediating the global response to changing conditions.

In *Escherichia coli*, the RNA-binding protein CsrA is a key component of one such global regulatory network that is involved in the transition from exponential to stationary growth phase (10,11). The activity of CsrA is modulated by two small RNAs, CsrB and CsrC, which control CsrA levels by binding to multiple copies of the protein (12,13). Recent work has further demonstrated that these sRNAs are activated by the BarA-UvrY two-component system in *E.coli* in a CsrA-dependent manner (14). Homologs of CsrA (e.g. RsmA in *Pseudomonas aeruginosa*) are highly conserved and are found in diverse bacteria where they play key roles in biofilm formation and dispersal (15), and in regulating virulence factors of animal and plant pathogens (16–19). It is interesting to note that, in the proteobacteria, most of the bacterial species having CsrA homologs also contain homologs of BarA and/or UvrY (e.g. the GacA–GacS two-component system in *P.aeruginosa*) and the interaction network between these proteins has been studied in several bacteria (14,19–24).

The presence of both CsrA and BarA–UvrY homologs in several bacterial species naturally leads to the question: Is the method of CsrA regulation via small RNAs also conserved in these species? Indeed, sRNA-encoding genes that regulate CsrA homologs have been identified already in several bacterial species, e.g. *rsmX*, *rsmY* and *rsmZ* in *Pseudomonas fluorescens* (22–25), *rsmB* in *Erwinia carotovora* (26), and *csrB*, *csrC* and *csrD* in *Vibrio cholerae* (19) to name a few. However, there are many bacterial species in which homologs of CsrA and BarA–UvrY are known to be important global regulators [e.g. in *Vibrio fischeri* (27) and in *Legionella pneumophila* (28)] for which the corresponding sRNAs, if they exist, have not been identified to date. The discovery of such sRNAs is complicated by the fact that they cannot all be identified by homology searches alone. Identifying potential CsrA-regulating sRNAs is therefore an important challenge in the field.

In this paper, we develop a procedure to discover potential CsrA-regulating sRNAs in bacteria. Recent experiments have shown that a repeated GGA motif in loop regions is a crucial

element in the small RNAs that regulate CsrA and its homologs (29,30). This suggests that the occurrence of a large number of such sequence motifs in a small genomic region could be a signature of CsrA-binding small RNAs. Building on this basic observation, we have developed a computer program (CSRNA_FIND) to search intergenic regions of the bacteria for potential CsrA-regulating sRNAs. The output of the program, in combination with secondary structure predictions using the program MFOLD (31), identifies all the experimentally known CsrA-regulating sRNAs and also leads to novel predictions for such sRNAs in several bacterial species. The predictions have been confirmed in *V.fischeri* through experiments which demonstrate the transcription of the predicted sRNAs in *V.fischeri* as well their ability to control CsrA levels in *E.coli*. As more genomes are sequenced and further experimental details regarding the binding motifs become available, this approach can be used to locate potential CsrA-regulating sRNAs in these genomes.

Outline of search algorithm

An analysis of the predicted secondary structures of known CsrA-regulating sRNAs indicates that the binding motif for CsrA is the presence of the sequence motif AGGA/ARGGA (where R stands for {T, C, G}) in single-stranded regions, particularly in the loop regions. For example, CsrB in *E.coli* is a 360 bp sRNA which has 16 occurrences of this motif in single-stranded regions (12). This suggests that a high concentration of the above binding motif could be a signature of sequences coding for CsrA-regulating sRNAs. Since the vast majority of bacterial sRNAs discovered to date are located in intergenic regions (6), we developed the program CSRNA_FIND to search for bacterial intergenic regions with high concentrations of the above binding motif to locate potential CsrA-regulating sRNAs. The algorithm steps are outlined below (further details are given in Materials and Methods):

- Obtain the intergenic regions of bacterial species having homologs of CsrA.
- Scan the intergenic regions (using a sliding window) for the number of occurrences of the AGGA/ARGGA-binding motif for a given window size.
- For each intergenic region, note the maximum number of occurrences (N_m) of the binding motif for the given window size.
- Obtain the frequency distribution $f(N_m)$ over the entire genome. Use this to determine the cutoff value N_c: all intergenic regions with $N_m > N_c$ are considered further as potential candidates for regions containing the sRNAs. Sometimes, these intergenic regions contain multiple occurrences of a repeat sequence (each unit being 7 bp or higher). Since these regions are unlikely to code for sRNAs, they are removed from the program output and the remaining intergenic regions are analyzed as follows.
- Scan the intergenic regions for the distribution of binding motifs and the presence of rho-independent terminators to determine putative 5′ and 3′ ends for the sRNA.
- Obtain the secondary structure of the predicted sRNA-encoding region using MFOLD. Compare the number of occurrences of binding motifs in single-stranded regions

with the corresponding number for experimentally known sRNAs of comparable length to determine if the intergenic region encodes a potential CsrA-regulating sRNA.

Since the sRNAs can be of varying lengths, the above procedure is repeated for a range of window sizes to generate a list of predictions for CsrA-regulating sRNAs which are discussed in Results.

MATERIALS AND METHODS

Algorithm details and sequence analysis

The program CSRNA_FIND was developed using the programming language PERL and is freely available upon request. Intergenic regions were obtained using the sequence analysis tools at http://rsat.ulb.ac.be/rsat/ (32). The range of window sizes used to scan the intergenic regions was {60, 90, 120, 150, 180, 210, 240}. The distribution of maximal number of occurrences (N_m) of the binding motif for each intergenic region was obtained for the top and bottom strands. This frequency distribution $f(N_m)$ was used to determine the cutoff value N_c for both strands. N_c was chosen to be the first non-zero integer such that $f(N_c + 1)$ is either 0 or 1. Rho-independent terminators were identified by searching for sequence motifs corresponding to GC-rich stem–loop regions followed by a poly(T) tail. The predicted 3′ end of the sRNA was identified with the rho-independent terminator sequence. Sequence information for experimentally known CsrA-regulating sRNAs, in particular the typical distance between the AGGA/ARGGA rich regions and the 5′ end of these sRNAs, was used to estimate the 5′ end of the predicted sRNA. The predicted secondary structures were obtained using the program MFOLD (31). Multiple alignments were carried out using TCoffee (33). The genome context was analyzed using the genome region comparison tool at TIGR. The sequence logos for the upstream binding sites were obtained using the WebLogo program (34) and the corresponding weight matrices were obtained using the program CONSENSUS available at http://rsat.ulb.ac.be/rsat (32,35). The derived weight matrices were used to scan the program output and the corresponding distribution of scores was analyzed to determine the cutoff for potential upstream binding sites.

Bacterial strains and growth conditions

E.coli DH5α or MG1655 were grown at 30 or 37°C in Luria–Bertani (LB) medium with ampicillin (100 μg/ml) when necessary. *V.fischeri* ES114 was grown in LBS medium (36) at 30°C. Kornberg agar plates (1.1% K$_2$HPO$_4$, 0.85% KH$_2$PO$_4$ and 0.6% yeast extract containing 1% glucose) with 1 mM isopropyl-β-D-thiogalactopyranoside (IPTG) and 100 μg/ml ampicillin were used to grow recombinant *E.coli* cultures for the glycogen iodine-staining assay.

DNA manipulation

Standard DNA manipulation procedures (37) were used for all cloning steps. PCR purification, gel extraction and plasmid purification kits were obtained from Qiagen. High-fidelity Deep Vent DNA Polymerase (New England Biolabs) was used to generate PCR products for cloning.

β-Galactosidase assays

The transcriptional fusions containing the promoter and part of the 5′ coding regions of *csrB1* and *csrB2* were separately amplified from *V.fischeri* ES114 chromosomal DNA by PCR with primers 5′-GTGACTTCCTATATTTCAGCTTTGC-3′ and 5′-CGCGGATCCGTGAGCGGTGTCCCTTACAT-3′ for *csrB1* and 5′-TGAGAATTCGTTGATGATTATCAGC-GCTTT-3′ and 5′-CGCGGATCCTTGAGCGGTGTCCTTT-AC-3′ for *csrB2*. EcoRI–BamHI fragments from these PCR products were then subcloned into a *lacZ* expression vector pSP417 (38) and the integrity of their nucleotide sequence was confirmed (Virginia Bioinformatics Institute Core Laboratories). The resulting constructs were used to perform β-galactosidase assays from cells grown to mid-log phase ($OD_{600} = 0.5$) in LB culture. Cell extracts were prepared from cells diluted 1:200 in Z buffer and lysed via chloroform. Assays were performed on 20 μl of cell extract using the Tropix Galacto-Light Plus Kit as per the manufacturer's recommendations. Triplicate assays were performed for each culture and the experiment was repeated three times.

Northern hybridization

V.fischeri cells harvested at four different OD_{600} values were treated with RNAprotect Bacteria Reagent (Qiagen) to stabilize the RNA prior to the RNA isolation. The RNA was isolated using the RNeasy Mini Kit (Qiagen). ^{32}P-labeled *csrB1* and *csrB2* riboprobes were produced by using a random primer DNA labeling kit as described by the manufacturer (Roche). Total cellular RNA (16 μg) was separated on a 1% formaldehyde agarose gel and transferred overnight onto a Nytran supercharge membrane (Turboblotter Gel Transfer Kit; Schleicher & Schuell) in 20× SSC transfer buffer. The RNA was immobilized on the membrane by an UV cross-linker (SpectroLinker; Spectronics Corporation). The membrane was pre-hybridized and hybridized in 10 ml of QuickHyb solution (Stratagene) at 65°C, for 30 min and 2–4 h, respectively, with a probe concentration of 2×10^6 c.p.m./ml and then washed twice for 15 min each in 2× SSC, 0.1% SDS at room temperature and once for 30 min in 0.2× SSC, 0.1% SDS at 60°C. The membrane was air-dried and then exposed to a phosphorimager screen (Molecular Dynamics).

Assays for glycogen production

The gene coding for CsrA was PCR amplified from *V.fischeri* chromosomal DNA with the primers 5′-CCCGGGATGCTA-ATTTTGACTCGCCGTGTAGG-3′ and 5′-AAGCTTTTAG-TGGTGGTGGTGGTGGTGAAAGTTACCTTGCGAAGCC-GCAGGTG-3′. The resulting PCR product encoded CsrA with a C-terminal His$_6$ tag, flanked by SmaI and HindIII restriction sites. The PCR product was ligated into pGEM (Promega, Madison, WI) and sequenced. A SmaI–HindIII fragment from this vector was subsequently ligated into pKK223-3 (39). The primers 5′-CACGGTACCTGGTGTC-GGAAGGATACTGA-3′ and 5′-GTTCTGCAGAAAAACC-CCACCAAGCTCTC-3′ for *csrB1* and 5′-GTAGGTACCT-ATTGGTGTCGGAAGGATGC-3′ and 5′-GTTCTGCAGA-AAAGCCCCACTAGATTTTCA-3′ for *csrB2* were used to amplify these genes from *V.fischeri* chromosomal DNA. KpnI–PstI fragments from these PCR products were ligated

into pUC19 (40) and the integrity of the nucleotide sequences was confirmed. EcoRI–PstI fragments from the *csrB1*- and *csrB2*-pUC19 constructs were then subsequently ligated into the expression vector pKK223-3. *E.coli* MG1655 encoding CsrB1, CsrB2 or CsrA under the control of the IPTG-inducible P_{tac} promoter in pKK223-3, as well as the empty vector, were individually streaked onto Kornberg agar plates. Plates were incubated at 30°C overnight then inverted over iodine crystals until a noticeable change in color could be detected.

RESULTS

Program output for *E.coli* and *V.fischeri*

Using the search procedure outlined in the previous section, we searched the intergenic regions of 60 bacterial species which have CsrA homologs. The complete list of the bacterial species analyzed in this study is included in the Supplementary Data (List L1). To illustrate the program output, consider first the results obtained from CSRNA_FIND using the intergenic regions of *E.coli* as input. Figure 1A shows the distribution of the maximal number of AGGA/ARGGA-binding motifs in intergenic regions of *E.coli*. As indicated in the figure, two intergenic regions are clearly separated from the genomic background; further analysis reveals that these regions exactly correspond to those encoding CsrB and CsrC in *E.coli*. It should be noted that the experimental identification of CsrC occurred several years after CsrB was first discovered (12,13). The fact that the program was able to identify these two sRNAs in the same iteration highlights the importance of bioinformatic analysis in potentially speeding up the discovery of CsrA-regulating sRNAs. In Figure 1B, we show the results of the program output for *V.fischeri*. Once again, two intergenic regions are clearly separated from the genomic background. Further analysis of these regions for the presence of rho-independent terminators and CsrA-binding sites leads to the prediction of two highly homologous sRNAs (88% sequence identity) which have been named CsrB1 and CsrB2. The predicted sRNAs are 416 and 420 bp long with 21 occurrences of the CsrA-binding motifs, respectively. As expected, the predicted secondary structure for CsrB1 (Figure 2) shows multiple stem–loop structures with most of the AGGA/ARGGA sites located in the loop regions.

Analysis of small RNA upstream sequences

The above procedure was repeated for all the bacterial species studied and the predicted sRNA-encoding sequences (from the program output) were further screened by analyzing their upstream regions. Previous work has shown that the sRNA upstream regions contain a conserved 18 bp sequence which is likely to correspond to the UvrY/GacA-binding site for activation of the sRNAs (19,22,24). The presence of a similar binding site in the upstream region of a putative sRNA can therefore serve as further evidence in support of the prediction. In order to test for the presence of such sites, we derived a weight matrix corresponding to the binding sites using the motif-finding tool CONSENSUS (35). First, the upstream regions of known *csrB* sRNA genes were
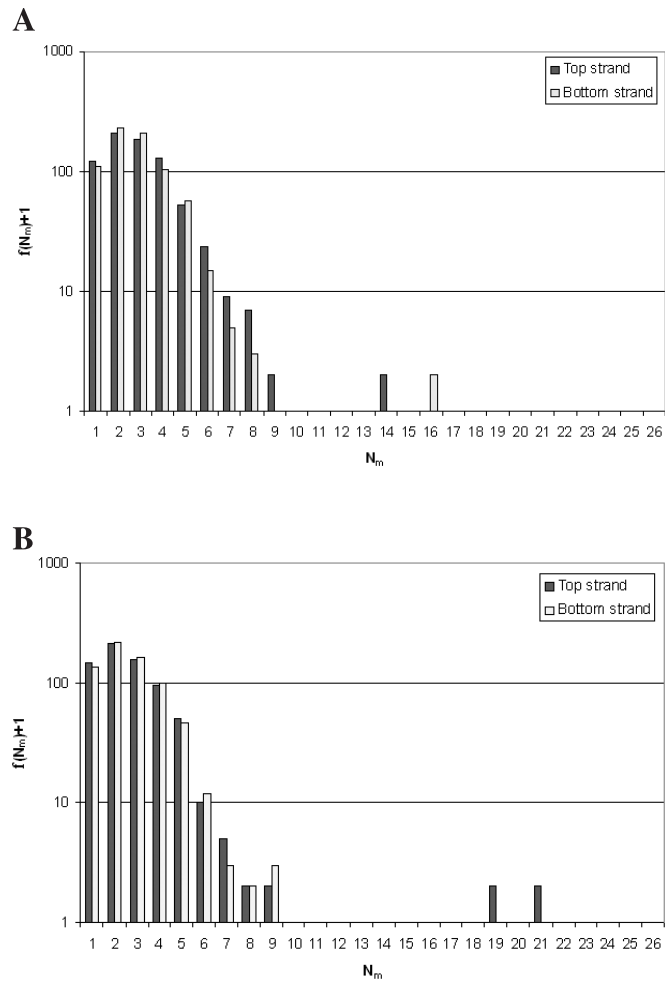
A

B

Figure 1. Distribution of AGGA/ARGGA-binding motifs in intergenic regions. (**A**) Frequency distribution $[f(N_m)]$ of the maximal number (N_m) of AGGA/ARGGA-binding motifs in intergenic regions of *E.coli* using a sliding window covering 240 bp. Two intergenic regions are clearly separated from the genomic background. Closed bars indicate the top strand and open bars indicate the bottom strand. (**B**) The same as (A) but for *V.fischeri*.

Figure 2. Secondary structure of CsrB1 in *V.fischeri*. Predicted secondary structure [obtained using MFOLD (31)] for CsrB1 in *V.fischeri* showing multiple AGGA/ARGGA sequence motifs in the loop regions. The secondary structure for CsrB2 is almost identical to that of CsrB1 since the two sRNAs are highly homologous.

used as the input for CONSENSUS and the derived weight matrix was used to scan the intergenic regions predicted to have CsrA-regulating sRNAs. The predicted sRNAs which showed strong binding sites in their upstream regions (termed '*csrB* upstream site') using the above weight matrix were categorized as *csrB* homologs. Multiple alignment of the upstream regions of these sRNAs (data not shown) also shows strong conservation of the 18 bp upstream sequence further validating their identification as homologous sRNA genes. A similar procedure was carried out using the upstream regions of known *csrC* sRNA genes which led to the identification of the subgroup of predicted sRNAs homologous to *csrC* of *E.coli*. Interestingly, the conserved 18 bp sequences upstream of the *csrC* sRNA genes obtained using CONSENSUS (termed '*csrC* upstream site') are distinct from the *csrB* upstream sites. Finally, a similar procedure was carried out to identify the binding sites in the upstream regions of the RsmA-regulating sRNAs in the Pseudomonads (termed '*rsmY* upstream site'). The *rsmY* upstream site is also
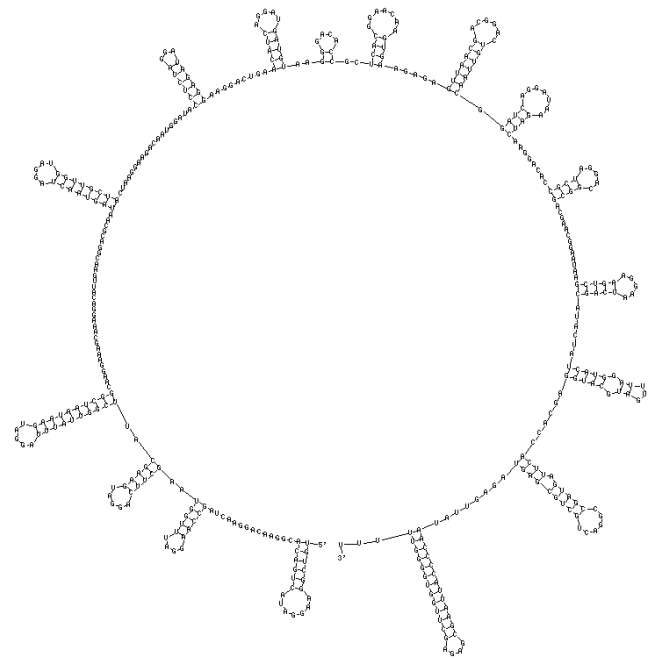
revealed by a multiple alignment of the upstream regions of the corresponding sRNAs; however, this is not the case for the *csrC* upstream site. Since the *csrC* upstream site is revealed by motif-finding tools and not by multiple alignment of the upstream sequences, it is less clear that the proposed binding site for *csrC* corresponds to an upstream activating sequence. The differences (and similarities) between the three sets of binding sites are illustrated by generating the corresponding sequences logos which are shown in Figure 3.

Predictions for CsrA-regulating small RNAs

The sRNA genes predicted by the program output, which also showed the presence of upstream binding sites (using the weight matrix search), have been categorized into three classes: *csrB* homologs, *csrC* homologs and *rsmX/Y/Z* homologs. The resulting output is summarized in Table 1 and more detailed information about the corresponding small RNAs (including their predicted lengths and genomic location) is provided in Supplementary Table S1. For the species considered, the above list includes all the experimentally confirmed sRNAs as well as predictions for several new sRNAs which have not yet been confirmed experimentally. Additionally, the program output contains several predicted sRNAs which satisfy all the search criteria but do not show a conserved binding site in their upstream regions. The sequence information for these predicted sRNAs (See Discussion) is provided in Table 2 and the detailed information about these sRNAs (including their predicted lengths and genomic location) is provided in Supplementary Table S2. The information regarding the predicted *csrB*, *csrC* and
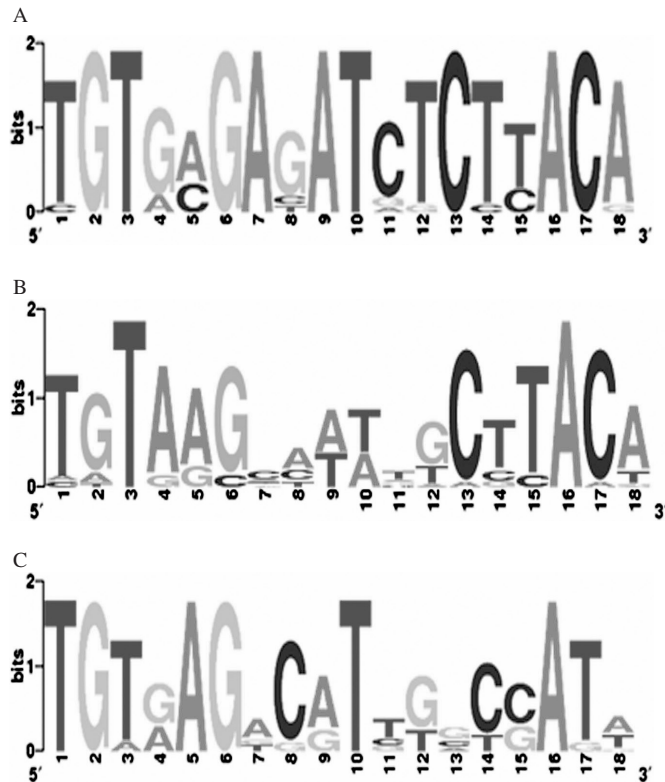
Figure 3. Sequence logos for upstream binding sites of predicted sRNAs. The sequence logos [generated using Web Logo (40)] for conserved upstream sites for all the known and predicted (**A**) *csrB*, (**B**) *rsmX/Y/Z* and (**C**) *csrC* sRNA genes.

rsmY upstream sites of the sRNAs is provided in Supplementary Table S3.

An interesting feature of the above predictions is that while many species appear to have multiple copies of *csrB* homologs, *csrC* is present only in single copy in the species that have it. A striking example is *Photobacterium profundum*, where our analysis predicts as many as four sRNAs homologous to *csrB* (in addition to a *csrC* sRNA). Multiple sRNAs have also been predicted in species such as *Vibrio parahaemolyticus*, *Vibrio vulnificus* and *Shewanella oneidensis*. Previous analysis had already identified the three *csrB* homologs in *V.parahaemolyticus* and *V.vulnificus* (19); however, the current work also revealed the presence of *csrC* in these species. Interestingly, we find no evidence of a *csrC* homolog in the closely related species *V.fischeri* and *V.cholerae*. In the Pseudomonads, the output from CSRNA_FIND led to the identification of three RsmA-regulating sRNAs in *P.fluorescens* in perfect agreement with experiments [the presence of the third sRNA was experimentally confirmed only recently (24)]. The above analysis also predicts the existence of three such sRNAs in *Pseudomonas syringae* whereas in *P.aeruginosa*, only two RsmA-regulating sRNAs are predicted. In the human pathogen *L.pneumophila*, for which CsrA functions as the key regulator for differentiation from the transmissive to the replicative phase (28), two CsrA-regulating sRNAs are predicted. The predicted RNAs are similar to those regulating RsmA in the Pseudomonads and accordingly have been named *rsmY* and *rsmZ*.

As more completed genome sequences become available, CSRNA_FIND can be used to locate the corresponding CsrA-regulating sRNAs. This is illustrated by the predictions for the corresponding sRNAs in *Pseudoalteromonas haloplanktis*, *Colwellia psychrerythraea* and *Psychrobacter arcticum* for which the completed genomes were made available only recently. It should be noted, however, that all the predictions presented in Table 1 correspond to bacterial species in the gammaproteobacteria. Thus the probability of the program predicting novel sRNAs in newly sequenced bacterial genomes is likely to correlate with the phylogeny of the species. Accordingly, the phylogenetic context of the predicted sRNAs from Table 1 is highlighted in the Supplementary Data (List L1).

In summary, our analysis leads to predictions for several new CsrA-regulating sRNAs in bacteria and also suggests a way of categorizing them based on conserved upstream sequences. In order to test the validity of these predictions, the corresponding experiments were carried out in *V.fischeri* as discussed below.

EXPERIMENTAL RESULTS

Transcription of *csrB1* and *csrB2* in *V.fischeri*

The presence of two *V.fischeri* sRNAs, CsrB1 and CsrB2, was confirmed. First, the existence of functional promoters for these two genes was measured via transcriptional fusions to *lacZ* in recombinant *E.coli* (Figure 4A). Second, the expression rates of CsrB1 and CsrB2 in *V.fischeri* were analyzed over time via northern blots. The total amount of the pool of CsrB1 and CsrB2 appears to remain steady between an OD_{600} of 0.25 and 2.0 as identical results were obtained using probes against either sRNA. Given that CsrB1 and CsrB2 are only 4 bp different in size and 88% identical, a single band of the appropriate size and thought to be representative of both sRNAs was observed (Figure 4B and data not shown).

Activity of CsrA, CsrB1 and CsrB2 in recombinant *E.coli*

A qualitative iodine-staining assay (13) was used to visualize glycogen production in recombinant *E.coli* strains overexpressing *V.fischeri* CsrA, CsrB1 and CsrB2 (Figure 5). Cells overexpressing *V.fischeri* CsrA had a noticeably lighter yellow–brown appearance than cells containing only the pKK223-3 vector. Over-expression of CsrA leads to decreased glycogen accumulation, which causes the lighter staining to be seen. Cells overexpressing CsrB1 or CsrB2 showed a much darker brown color than the other strains, which indicates that they overproduce glycogen as a result of the inactivation of CsrA. Hence, the genes predicted to encode CsrA, CsrB1 and CsrB2 from *V.fischeri* are able to function in *E.coli* and interact with the glycogen regulatory network in a manner consistent with that of their *E.coli* protein counterparts.

DISCUSSION

Sequence criteria for CsrA-regulating small RNAs

Several hitherto undiscovered CsrA-regulating small RNAs have now been predicted using the program CSRNA_FIND.

Table 1. CsrA-regulating sRNA genes from the program output

Bacterial species	sRNA gene	Flanking genes	Orientation[a]	References[b]
Acinetobacter sp. ADP1	*rsmX*	*ACIAD0938/ACIAD0939*	→ ← →	Predicted[c]
	rsmY	*ACIAD0480/gltP*	→ ← →	Predicted[c]
	rsmZ	*ACIAD3594/ACIAD3596*	← → ←	Predicted[c]
C.psychrerythraea	*csrB1*	*CPS_3528/CPS_3529*	← ← ←	Predicted[c]
	csrB2	*CPS_3528/CPS_3529*	← ← ←	Predicted[c]
E.carotovora SCRI1043	*rsmB*	*syd/aepA*	→ → →	(26)
E.coli K12	*csrB*	*yqcC/syd*	← ← ←	(12)
	csrC	*yihA/yihI*	← → →	(13)
L.pneumophila (P)	*rsmY*	*gyrB/lpp0005*	→ ← →	Predicted[c]
	rsmZ	*lpp1662/lpp1663*	→ ← →	Predicted[c]
P.profundum	*csrB1*	*PBPRA2976/PBPRA2977*	→ ← →	Predicted[c]
	csrB2	*PBPRA2976/PBPRA2977*	→ ← →	Predicted[c]
	csrB3	*PBPRA2976/PBPRA2977*	→ ← →	Predicted[c]
	csrB4	*PBPRB1150/PBPRB1151*	→ ← ←	Predicted[c]
	csrC	*PBPRA3500/PBPRA3501*	← ← →	Predicted[c]
Photorhabdus luminescens	*csrB*	*syd/plu0664*	→ → →	Predicted[c]
P.haloplanktis	*csrB1*	*PSHAa1973/PSHAa1975*	→ ← ←	Predicted[c]
	csrB2	*PSHAa2664/gabD*	← → ←	Predicted[c]
	csrC	*PSHAa2751/PSHAa2752*	← ← ←	Predicted[c]
P.aeruginosa	*rsmY*	*dnr/PA0528*	← → ←	(22)[d]
	rsmB/rsmZ	*fdxA/rpoS*	→ ← ←	(16,45)
P.fluorescens Pf-5	*rsmX*	*PFL4112/PFL4113*	→ ← ←	(24)
	rsmY	*PFL5683/PFL5684*	→ ← ←	(22)
	rsmZ	*rpoS/fdxA*	→ → ←	(23)
P.putida	*rsmY*	*PP0370/PP0371*	→ → ←	(22)[d]
	rsmZ	*PP1624/PP1625*	→ → ←	(22)[d]
P.syringae	*rsmX*	*PSPTO3698/PSPTO3699*	→ ← →	Predicted[c]
	rsmY	*PSPTO0506/PSPTO0507*	→ → →	(22)[d]
	rsmZ	*PSPTO1566/PSPTO1567*	← ← →	(22)[d]
P.arcticum	*rsmY*	*Psyc_1521/Psyc_1522*	→ ← ←	Predicted[c]
Salmonella enterica Typhi	*csrB*	*yqcC/syd*	← ← ←	(6)[d]
	csrC	*yihI/STY3880*	← ← →	(6)[d]
Salmonella typhimurium	*csrB*	*yqcC/syd*	← ← ←	(17)
	csrC	*yihA/yihI*	← → →	(46)
S.oneidensis	*csrB1*	*SO1615/SO1616*	→ → →	Predicted[c]
	csrB2	*SO1616/SO1617*	→ → →	Predicted[c]
Shigella flexneri 2a str. 301	*csrB*	*SF2805/syd*	← ← ←	(6)[d]
	csrC	*yihA/yihI*	← → →	(6)[d]
V.cholerae	*csrB1*	*VC0882/VC0883*	← → ←	(19)
	csrB2	*VC0190/VC0191*	→ ← ←	(19)
	csrB3	*VCA0839/VCA0840*	→ ← →	(19)
V.fischeri ES114	*csrB1*	*VF0602/VF0603*	← → →	Predicted[c]
	csrB2	*VF0051/VF0052*	→ → ←	Predicted[c]
V.parahaemolyticus	*csrB1*	*VP2326/VP2327*	← ← →	(19)[d]
	csrB2	*VP3011/VP3012*	→ → ←	(19)[d]
	csrB3	*VPA0175/VPA0176*	← ← ←	(19)[d]
	csrC	*VP0110/VP0111*	→ → ←	Predicted[c]
V.vulnificus CMCP6	*csrB1*	*VV11848/VV11852*	← → →	(19)[d]
	csrB2	*VV10946/VV10949*	→ → ←	(19)[d]
	csrB3	*VV20844/VV20845*	← → ←	(19)[d]
	csrC	*VV10897/VV10899*	← ← ←	Predicted[c]
Yersinia pestis CO92	*csrB*	*syd/tnp*	→ → →	(6)[d]
	csrC	*YPO0019/YPO0020*	← → →	(6)[d]
Yersinia pseudotuberculosis	*csrB*	*YPTB3010/syd*	← ← ←	Predicted[c]
	csrC	*YPTB0019/YPTB0020*	← → →	Predicted[c]

[a]The genes on the top strand are indicated by '→' whereas the genes on the bottom strand are indicated by '←'. The middle arrow indicates the orientation of the sRNA and the flanking arrows indicate the orientation of the adjacent genes.
[b]Previous work in which the sRNA has been discussed and/or experimentally demonstrated.
[c]sRNA predicted by CSRNA_FIND in the present study.
[d]Previous work in which the sRNA has been discussed but not experimentally verified.

The predicted sRNA-encoding sequences (Table 1) all satisfy the following requirements:

(i) located in intergenic regions;
(ii) high concentration of the putative CsrA-binding motif AGGA/ARGGA;
(iii) presence of a rho-independent terminator;
(iv) predicted secondary structure showing repeated occurrences of the sequence element GGA in loop and free regions; and
(v) presence of a conserved upstream sequence categorized as either a *csrB, csrC* or *rsmY* upstream site.

Table 2. Additional predictions for CsrA-regulating sRNA genes

Bacterial species	sRNA genes	Flanking genes	Orientation[a]
Acinetobacter sp ADP1	*rsmB*	*crc/ACIAD3528*	→ → →
B.subtilis	*rsmY*	*yybO/yybN*	→ ← →
H.pylori J99	*csrB1*	*jhp0951/jhp0952*	← ← →
	csrB2	*rbn/jhp1300*	→ ← →
P.putida KT2440	*rsmX*	*PP4094/PP4095*	→ → →
P.arcticum	*rsmX*	*prc/slyD*	→ → ←
	rsmZ	*Psyc_0155*[b]	
S.oneidensis	*csrC*	*mutM/SO4727*	→ → →

[a]The genes on the top strand are indicated by '→' whereas the genes on the bottom strand are indicated by '←'. The middle arrow indicates the orientation of the sRNA and the flanking arrows indicate the orientation of the adjacent genes.
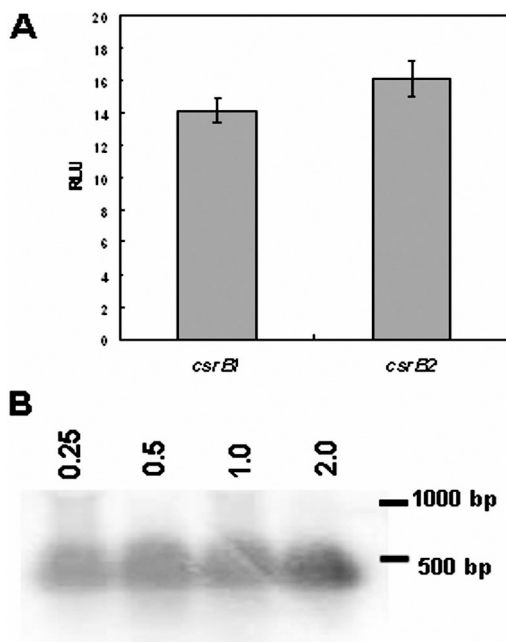[b]sRNA located entirely within coding sequence of given gene.

Figure 4. Transcription of *csrB1* and *csrB2*. (**A**) β-Galactosidase activity levels of recombinant DH5α strains encoding *csrB1*- or *csrB2-lacZ* transcriptional fusions in pSP417. Background levels of β-galactosidase produced from the negative control pSP417 were 0.063 ± 0.004 RLU. Error bars represent the standard deviation of assays performed in triplicate from three independent samples. (**B**) Northern blot analysis of the rate of transcription of *csrB1* and *csrB2* in *V.fischeri* ES114 grown to different OD values as indicated using *csrB2* sequences as a probe. Identical results were obtained when *csrB1* sequences were used as a probe (data not shown). The blot shown is representative of two independent experiments. The migration of RNA size standards is indicated on the right.

The criteria given above are met by all experimentally known CsrA-regulating sRNA homologs and can be considered to be the defining features of such sRNAs. Since the predicted novel sRNAs in Table 1 also satisfy all the above requirements, this suggests a high degree of confidence in the validity of these predictions.

Additional predictions

In addition to the sRNAs listed in Table 1, our analysis revealed several sRNAs satisfying most but not all the criteria
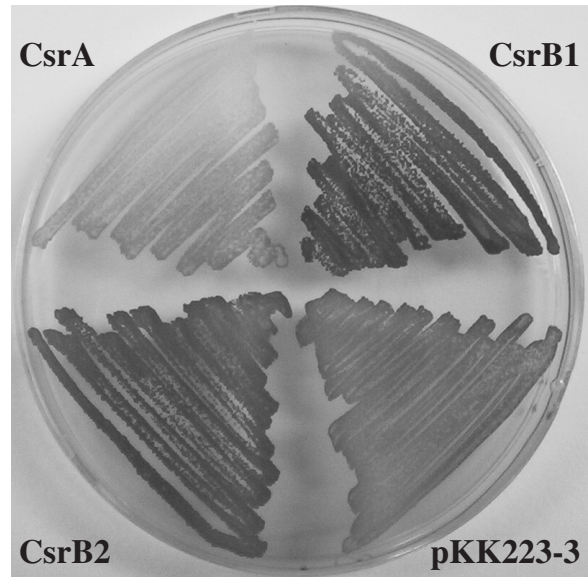

Figure 5. Effects of *V.fischeri* proteins on glycogen regulation. Recombinant *E.coli* MG1655 overexpressing *V.fischeri* CsrA, CsrB1, CsrB2 or no protein from *V.fischeri* were grown on Kornberg agar plates supplemented with 1 mM IPTG and 100 μg/ml ampicillin and qualitatively assayed for levels of glycogen production.

listed above. The sequence information relating to these sRNAs is provided in Table 2 and the predicted sRNAs are discussed further below.

- In both *Pseudomonas putida* and *Acinetobacter* sp., the program predicts additional sRNAs satisfying conditions (i)–(iv) above, both of which, however, lack a conserved upstream binding site.
- In *P.arcticum*, on the other hand, there are two additional predicted sRNAs both of which show a strong *rsmY* upstream site. One of the sRNAs does not have a high concentration of the AGGA/ARGGA motif; however, the predicted secondary structure shows multiple occurrences of GGA in the loop regions. The other sRNA is not in the intergenic regions but is located entirely in the coding sequence of a predicted hypothetical protein. Since the predicted sRNA-encoding sequence satisfies conditions (ii)–(v) above, it is very likely that the sequence codes for a CsrA-regulating sRNA rather than being part of a hypothetical protein as suggested by the annotation.
- In *Helicobacter pylori*, the program predicts two highly homologous sRNAs satisfying conditions (i)–(iv) above but lacking a conserved binding site in the upstream regions, which is not surprising since *H.pylori* does not have a UvrY ortholog. Regardless, the lack of a predicted upstream site reduces the degree of confidence in the prediction. However, it would be interesting to experimentally test these predictions since a previous study, carrying out a detailed analysis of the role of CsrA in *H.pylori* infections (41), attempted to locate CsrA-regulating sRNAs in this organism without success.
- In *Bacillus subtilis*, the program predicts a sRNA-encoding sequence satisfying conditions (i)–(iv) above but lacking a conserved upstream site consistent with the absence of a UvrY homolog. If the prediction is experimentally

confirmed, this would be an exciting development, since it would be, to our knowledge, the first instance of a CsrA-regulating sRNA in the Gram-positive bacteria.

- In *S.oneidensis*, our analysis predicts an additional sRNA which does not have a high concentration of the AGGA/ARGGA-binding motif. However, the presence of the *csrC* upstream site, in conjunction with conservation of genome context (see below) strongly suggests that the region codes for a sRNA homologous to *csrC*.

Classification of CsrA-regulating small RNAs

In addition to predicting novel sRNAs, our study has enabled a classification of two types of CsrA-regulating sRNA genes in the gamma proteobacteria: those that are homologous to *csrB* and those that are homologous to *csrC*. The classification of the predicted small RNAs as either a *csrB* homolog or a *csrC* homolog is based on multiple lines of evidence. First, analysis of the upstream regions gives rise to distinct activator binding site motifs for *csrB* and *csrC* (Figure 3) which is used to classify the sRNAs. This classification is further validated by homology searches: for all the bacterial species having two or more predicted *csrB* sRNAs, one of the *csrB* homologs can be used to identify all the others in that organism using BLAST searches. On the other hand, the sequences of *csrB* and *csrC* within each bacterial species are sufficiently different such that neither can be identified from the other using homology searches. Finally, analysis of the genome context of *csrC* homologs reveals that the sRNA is always located in the neighborhood of the genes *yihI* and *yihA* (which are the flanking genes for *csrC* in *E.coli*). A similar analysis for *csrB* sRNAs reveals that at least one of the *csrB* homologs in all the bacterial species (with the exception of *V.parahaemolytcius* and *V.vulnificus*) is in the genome neighborhood (i.e. separated by <20 genes) of the *syd* gene (which is one of the flanking genes for *csrB* in *E.coli*). This conservation of genome context further strengthens the validity of the predicted novel sRNAs and supports the classification based on conserved upstream binding sites.

Connections to other global regulatory networks

Recent work has shown that there is a close connection between the quorum-sensing regulatory network and the CsrA regulon in *V.cholerae* (19). Studying the genome context of the predicted sRNAs also suggests further connections between the CsrA regulon and global regulatory networks such as the quorum-sensing regulon. For example, one of the flanking genes for *csrB4* in *P.profundum* is *PBPRB1151*. The ortholog of this gene in *V.fischeri* (VFA1016) was shown recently to be part of a regulatory locus that is differentially regulated by quorum sensing (42). Furthermore, as noted earlier, *csrC* is always found in the genome neighborhood of the gene *yihA* which has been shown to be essential for normal cell division (43). This suggests a hypothesis linking the CsrA regulon with the regulation of cell division. The suggested connection is further strengthened by the observation that in *E.coli*, the protein SdiA (which is a homolog of the quorum-sensing regulator LuxR of *V.fischeri*) has been shown to regulate

both transcription of *csrB* and *csrC* (14) as well as the transcription of *ftsZ* (a gene that is essential for cell division) (44). It would be of interest to explore these connections further in *V.fischeri* to study the integration of these global regulatory networks.

CONCLUSIONS

In conclusion, we have developed an algorithm for the discovery of CsrA-regulating sRNAs in bacteria. Our analysis recovers all experimentally known sRNAs and makes novel predictions for such sRNAs in important species such as *L.pneumophila*, *V.parahaemolyticus*, *S.oneidensis* and *P.haloplanktis* to name a few. Our experimental results have verified the predictions in *V.fischeri* and also provide the groundwork for future studies exploring the connections between the CsrA regulon and other global regulatory networks. It should be noted that while predictions have been made for some species, there are many more bacterial species with CsrA homologs for which our program could not find a definitive signature of CsrA-regulating sRNAs. This may be because the mode of regulation of CsrA (via sRNAs) is not conserved in the other species. Alternatively, in the species with distant CsrA homologs, the mode of regulation (via sRNAs) is retained but the binding motifs for CsrA have changed to the extent that these sRNAs cannot be identified using our present scheme. It is hoped that future experimental studies in combination with similar bioinformatic approaches will be instrumental in unraveling the mode of CsrA regulation in additional bacterial species.

REFERENCES

1. Wassarman,K.M., Repoila,F., Rosenow,C., Storz,G. and Gottesman,S. (2001) Identification of novel small RNAs using comparative genomics and microarrays. *Genes Dev.*, **15**, 1637–1651.
2. Argaman,L., Hershberg,R., Vogel,J., Bejerano,G., Wagner,E.G., Margalit,H. and Altuvia,S. (2001) Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. *Curr. Biol.*, **11**, 941–950.
3. Vogel,J., Bartels,V., Tang,T.H., Churakov,G., Slagter-Jager,J.G., Huttenhofer,A. and Wagner,E.G. (2003) RNomics in *Escherichia coli* detects new sRNA species and indicates parallel transcriptional output in bacteria. *Nucleic Acids Res.*, **31**, 6435–6443.
4. Zhang,A., Wassarman,K.M., Rosenow,C., Tjaden,B.C., Storz,G. and Gottesman,S. (2003) Global analysis of small RNA and mRNA targets of Hfq. *Mol. Microbiol.*, **50**, 1111–1124.

5. Livny,J., Fogel,M.A., Davis,B.M. and Waldor,M.K. (2005) sRNAPredict: an integrative computational approach to identify sRNAs in bacterial genomes. *Nucleic Acids Res.*, **33**, 4096–4105.

6. Hershberg,R., Altuvia,S. and Margalit,H. (2003) A survey of small RNA-encoding genes in *Escherichia coli*. *Nucleic Acids Res.*, **31**, 1813–1820.

7. Repoila,F., Majdalani,N. and Gottesman,S. (2003) Small non-coding RNAs, co-ordinators of adaptation processes in *Escherichia coli*: the RpoS paradigm. *Mol. Microbiol.*, **48**, 855–861.

8. Lenz,D.H., Mok,K.C., Lilley,B.N., Kulkarni,R.V., Wingreen,N.S. and Bassler,B.L. (2004) The small RNA chaperone Hfq and multiple small RNAs control quorum sensing in *Vibrio harveyi* and *Vibrio cholerae*. *Cell*, **118**, 69–82.

9. Masse,E., Vanderpool,C.K. and Gottesman,S. (2005) Effect of RyhB small RNA on global iron use in *Escherichia coli*. *J. Bacteriol.*, **187**, 6962–6971.

10. Romeo,T. (1998) Global regulation by the small RNA-binding protein CsrA and the non-coding RNA molecule CsrB. *Mol. Microbiol.*, **29**, 1321–1330.

11. Majdalani,N., Vanderpool,C.K. and Gottesman,S. (2005) Bacterial small RNA regulators. *Crit. Rev. Biochem. Mol. Biol.*, **40**, 93–113.

12. Liu,M.Y., Gui,G., Wei,B., Preston,J.F.,III, Oakford,L., Yüksel,U., Giedroc,D.P. and Romeo,T. (1997) The RNA molecule CsrB binds to the global regulatory protein CsrA and antagonizes its activity in *Escherichia coli*. *J. Biol. Chem.*, **272**, 17502–17510.

13. Weilbacher,T., Suzuki,K., Dubey,A.K., Wang,X., Gudapaty,S., Morozov,I., Baker,C.S., Georgellis,D., Babitzke,P. and Romeo,T. (2003) A novel sRNA component of the carbon storage regulatory system of *Escherichia coli*. *Mol. Microbiol.*, **48**, 657–670.

14. Suzuki,K., Wang,X., Weilbacher,T., Pernestig,A.K., Melefors,O., Georgellis,D., Babitzke,P. and Romeo,T. (2002) Regulatory circuitry of the CsrA/CsrB and BarA/UvrY systems of *Escherichia coli*. *J. Bacteriol.*, **184**, 5130–5140.

15. Jackson,D.W., Suzuki,K., Oakford,L., Simecka,J.W., Hart,M.E. and Romeo,T. (2002) Biofilm formation and dispersal under the influence of the global regulator CsrA of *Escherichia coli*. *J. Bacteriol.*, **184**, 290–301.

16. Heurlier,K., Williams,F., Heeb,S., Dormond,C., Pessi,G., Singer,D., Camara,M., Williams,P. and Haas,D. (2004) Positive control of swarming, rhamnolipid synthesis, and lipase production by the posttranscriptional RsmA/RsmZ system in *Pseudomonas aeruginosa* PAO1. *J. Bacteriol.*, **186**, 2936–2945.

17. Altier,C., Suyemoto,M., Ruiz,A.I., Burnham,K.D. and Maurer,R. (2000) Characterization of two novel regulatory genes affecting *Salmonella* invasion gene expression. *Mol. Microbiol.*, **35**, 635–646.

18. Ma,W., Cui,Y., Liu,Y., Dumenyo,C.K., Mukherjee,A. and Chatterjee,A.K. (2001) Molecular characterization of global regulatory RNA species that control pathogenicity factors in *Erwinia amylovora* and *Erwinia herbicola* pv. gypsophilae. *J. Bacteriol.*, **183**, 1870–1880.

19. Lenz,D.H., Miller,M.B., Zhu,J., Kulkarni,R.V. and Bassler,B.L. (2005) CsrA and three redundant small RNAs regulate quorum sensing in *Vibrio cholerae*. *Mol. Microbiol.*, **58**, 1186–1202.

20. Cui,Y., Chatterjee,A. and Chatterjee,A.K. (2001) Effects of the two-component system comprising GacA and GacS of *Erwinia carotovora* subsp. *carotovora* on the production of global regulatory rsmB RNA, extracellular enzymes, and Harpin_Ecc. *Mol. Plant Microbe Interact.*, **14**, 516–526.

21. Teplitski,M., Goodier,R.I. and Ahmer,B.M. (2003) Pathways leading from BarA/SirA to motility and virulence gene expression in *Salmonella*. *J. Bacteriol.*, **185**, 7257–7265.

22. Valverde,C., Heeb,S., Keel,C. and Haas,D. (2003) RsmY, a small regulatory RNA, is required in concert with RsmZ for GacA-dependent expression of biocontrol traits in *Pseudomonas fluorescens* CHA0. *Mol. Microbiol.*, **50**, 1361–1379.

23. Heeb,S., Blumer,C. and Haas,D. (2002) Regulatory RNA as mediator in GacA/RsmA-dependent global control of exoproduct formation in *Pseudomonas fluorescens* CHA0. *J. Bacteriol.*, **184**, 1046–1056.

24. Kay,E., Dubuis,C. and Haas,D. (2005) Three small RNAs jointly ensure secondary metabolism and biocontrol in *Pseudomonas fluorescens* CHA0. *Proc. Natl Acad. Sci. USA*, **102**, 17136–17141.

25. Aarons,S., Abbas,A., Adams,C., Fenton,A. and O'Gara,F. (2000) A regulatory RNA (PrrB RNA) modulates expression of secondary metabolite genes in *Pseudomonas fluorescens* F113. *J. Bacteriol.*, **182**, 3913–3919.

26. Liu,Y., Cui,Y., Mukherjee,A. and Chatterjee,A.K. (1998) Characterization of a novel RNA regulator of *Erwinia carotovora* ssp. *carotovora* that controls production of extracellular enzymes and secondary metabolites. *Mol. Microbiol.*, **29**, 219–234.

27. Whistler,C.A. and Ruby,E.G. (2003) GacA regulates symbiotic colonization traits of *Vibrio fischeri* and facilitates a beneficial association with an animal host. *J. Bacteriol.*, **185**, 7202–7212.

28. Molofsky,A.B. and Swanson,M.S. (2003) *Legionella pneumophila* CsrA is a pivotal repressor of transmission traits and activator of replication. *Mol. Microbiol.*, **50**, 445–461.

29. Valverde,C., Lindell,M., Wagner,E.G. and Haas,D. (2004) A repeated GGA motif is critical for the activity and stability of the riboregulator RsmY of *Pseudomonas fluorescens*. *J. Biol. Chem.*, **279**, 25066–25074.

30. Dubey,A.K., Baker,C.S., Romeo,T. and Babitzke,P. (2005) RNA sequence and secondary structure participate in high-affinity CsrA–RNA interaction. *RNA*, **11**, 1579–1587.

31. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.

32. van Helden,J. (2003) Regulatory sequence analysis tools. *Nucleic Acids Res.*, **31**, 3593–3596.

33. Notredame,C., Higgins,D. and Heringa,J. (2000) T-Coffee: a novel method for multiple sequence alignments. *J. Mol. Biol.*, **302**, 205–217.

34. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.

35. Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.

36. Dunlap,P.V. (1989) Regulation of luminescence by cyclic AMP in *cya*-like and *crp*-like mutants of *Vibrio fischeri*. *J. Bacteriol.*, **171**, 1199–1202.

37. Sambrook,J., Fritsch,E.F. and Maniatis,T. (1989) *Molecular Cloning: A Laboratory Manual, 2nd edn*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

38. Podkovyrov,S.M. and Larson,T.J. (1995) A new vector-host system for construction of *lacZ* transcriptional fusions where only low-level gene expression is desirable. *Gene*, **156**, 151–152.

39. Amann,E., Brosius,J. and Ptashne,M. (1983) Vectors bearing a hybrid *trp-lac* promoter useful for regulated expression of cloned genes in *Escherichia coli*. *Gene*, **25**, 167–178.

40. Yanisch-Perron,C., Vieira,J. and Messing,J. (1985) Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors. *Gene*, **33**, 103–119.

41. Barnard,F.M., Loughlin,M.F., Fainberg,H.P., Messenger,M.P., Ussery,D.W., Williams,P. and Jenks,P.J. (2004) Global regulation of virulence and the stress response by CsrA in the highly adapted human gastric pathogen *Helicobacter pylori*. *Mol. Microbiol.*, **51**, 15–32.

42. Lupp,C. and Ruby,E.G. (2005) *Vibrio fischeri* uses two quorum-sensing systems for the regulation of early and late colonization factors. *J. Bacteriol.*, **187**, 3620–3629.

43. Dassain,M., Leroy,A., Colosetti,L., Carole,S. and Bouche,J.P. (1999) A new essential gene of the 'minimal genome' affecting cell division. *Biochimie*, **81**, 889–895.

44. Garcia-Lara,J., Shang,L.H. and Rothfield,L.I. (1996) An extracellular factor regulates expression of *sdiA*, a transcriptional activator of cell division genes in *Escherichia coli*. *J. Bacteriol.*, **178**, 2742–2748.

45. Burrowes,E., Abbas,A., O'Neill,A., Adams,C. and O'Gara,F. (2005) Characterisation of the regulatory RNA RsmB from *Pseudomonas aeruginosa* PAO1. *Res. Microbiol.*, **156**, 7–16.

46. Fortune,D.R., Suyemoto,M. and Altier,C. (2006) Identification of CsrC and characterization of its role in epithelial cell invasion in *Salmonella enterica serovar Typhimurium*. *Infect. Immun.*, **74**, 331–339.