Identification of 17 *Pseudomonas aeruginosa* sRNAs and prediction of sRNA-encoding genes in 10 diverse pathogens using the bioinformatic tool sRNAPredict2

Jonathan Livny*, Anja Brencic[1], Stephen Lory[1] and Matthew K. Waldor

Department of Molecular Biology and Microbiology, Tufts University School of Medicine and Howard Hughes Medical Institute, 136 Harrison Avenue, Boston, MA 02111, USA and [1]Department of Microbiology and Molecular Genetics, Harvard Medical School, 200 Longwood Avenue, Boston, MA 02115, USA

ABSTRACT

sRNAs are small, non-coding RNA species that control numerous cellular processes. Although it is widely accepted that sRNAs are encoded by most if not all bacteria, genome-wide annotations for sRNA-encoding genes have been conducted in only a few of the nearly 300 bacterial species sequenced to date. To facilitate the efficient annotation of bacterial genomes for sRNA-encoding genes, we developed a program, sRNAPredict2, that identifies putative sRNAs by searching for co-localization of genetic features commonly associated with sRNA-encoding genes. Using sRNAPredict2, we conducted genome-wide annotations for putative sRNA-encoding genes in the intergenic regions of 11 diverse pathogens. In total, 2759 previously unannotated candidate sRNA loci were predicted. There was considerable range in the number of sRNAs predicted in the different pathogens analyzed, raising the possibility that there are species-specific differences in the reliance on sRNA-mediated regulation. Of 34 previously unannotated sRNAs predicted in the opportunistic pathogen *Pseudomonas aeruginosa*, 31 were experimentally tested and 17 were found to encode sRNA transcripts. Our findings suggest that numerous genes have been missed in the current annotations of bacterial genomes and that, by using improved bioinformatic approaches and tools, much remains to be discovered in 'intergenic' sequences.

INTRODUCTION

In recent years, it has become evident that small, non-coding RNA species (sRNAs) control diverse cellular processes in numerous prokaryotic species (1,2). Most bacterial sRNAs characterized to date act as post-transcriptional regulators by forming duplexes at the 5′ leader regions of mRNAs, modulating mRNA stability and/or altering the access of mRNAs to the translational machinery (2). Often these interactions are stabilized by Hfq, an RNA chaperone protein conserved in numerous bacterial species (3). Deletion of *hfq* in several bacterial pathogens, including *Vibrio cholerae*, *Salmonella typhimurium*, *Brucella abortus*, *Yersinia enterocolitica* and *Pseudomonas aeruginosa*, has been shown to severely attenuate virulence (4–8). Moreover, deletion of four functionally redundant sRNAs in *V.cholerae* abrogates expression of TCP, a critical *V.cholerae* virulence factor (9). Taken together, these findings suggest that sRNA-mediated regulation is a well-conserved mechanism for modulating bacterial virulence.

The vast majority of the ∼100 bacterial sRNAs known to date have been identified in *Escherichia coli* (2). With the exception of a few highly conserved sRNAs such as tmRNA and RnpB, most *E.coli* sRNAs are well conserved only among closely related species such as *Salmonella* sp. and *Yersinia* sp. (10). Consequently, relatively few putative sRNAs have been identified in other species based solely on primary sequence homology with known *E.coli* sRNAs (11). Many of the recently discovered *E.coli* were initially predicted using integrative bioinformatic approaches that identified putative sRNAs by searching for co-localization of several genetic features commonly associated with sRNA-encoding genes, including promoters, Rho-independent terminators and/or regions of intergenic sequence conservation (12–14). Each of these integrative searches for sRNA-encoding genes involved hundreds or thousands of individual predictive features, whose co-localization was determined either by arduous non-computational methods, severely limiting the rate at which searches could be conducted, or by the *de novo* development and use of novel scripted approaches. Thus, despite the success of these integrative algorithms in identifying novel sRNAs, the lack of computational tools to

*To whom correspondence should be addressed. Tel: +1 617 909 8804; Fax: +1 617 636 2723; Email: jonathan.livny@tufts.edu

efficiently utilize these approaches has severely hindered their implementation. As a consequence, before this study, annotations for sRNA-encoding genes using integrative computational approaches had been conducted in only a few of the nearly 300 sequenced bacterial species (9,10,15–17).

To facilitate the efficient prediction of bacterial sRNAs, we developed sRNAPredict, a C++ program that flexibly integrates different combinations of individual predictive features of sRNAs to rapidly identify putative sRNA-encoding genes in the intergenic regions (IGRs) of any annotated bacterial genome (15). Using sRNAPredict, we predicted dozens of previously unannotated candidate *V.cholerae* sRNAs by searching the *V.cholerae* genome for putative transcriptional terminators encoded downstream of regions of intergenic sequence conservation. Of nine of these predicted sRNAs subjected to experimental verification by northern analysis, five were confirmed.

Although the identification of previously unknown *V.cholerae* sRNAs in our initial search validated sRNAPredict as a bioinformatic tool, it remained unclear whether our general approach was one that could be used to accurately predict novel sRNAs in other bacterial species. Furthermore, the number of *V.cholerae* sRNAs subjected to physical confirmation in our previous study was too small to allow us to identify new features shared by the confirmed sRNAs that could be used to improve the accuracy of our predictive algorithm.

Here we have used sRNAPredict to identify candidate sRNA-encoding genes in the IGRs of the opportunistic Gram-negative pathogen *P.aeruginosa* using putative Rho-independent terminators encoded downstream of regions of sequence conservation as predictors. A total of 34 previously unidentified candidate sRNAs were predicted in this annotation. Of these sRNAs 31 were subjected to physical verification by northern assay and 17 were found to encode sRNA transcripts. Compared with the candidates that do not appear to be transcribed, the confirmed sRNAs tend to have lower BLAST *E*-values, are more often conserved in multiple *Pseudomonas* species, and are more often predicted to encode conserved secondary structure. Using an improved version of sRNAPredict, sRNAPredict2, we identified potential sRNA-encoding genes in 10 additional pathogens. These analyses suggest that all of these diverse pathogens encode numerous sRNAs but that the number of sRNAs per species may vary considerably.

MATERIALS AND METHODS

Summary of the essential features of the sRNAPredict program

sRNAPredict searches for putative sRNAs are limited to regions of the genome that do not encode proteins, tRNAs or rRNAs. sRNAPredict identifies putative sRNAs by searching for the co-localization of genetic features that are associated with many bacterial sRNAs. These genetic features include (i) regions of sequence conservation (ii) Rho-independent terminators and (iii) putative promoters. The various algorithms used by sRNAPredict can utilize any combination of these features. The program only relies on the coordinate locations and, when applicable, strand orientations

of predictive elements; no sequence information is used. The coordinate positions and strand orientations of the predictive elements are automatically extracted from BLAST output files or from the output files of RNAMotif (18) and TransTerm (19), programs that predict putative Rho-independent terminators. The lengths of putative sRNAs reported both by sRNAPredict include the length of the predicted Rho-independent terminator. For a more detailed description of sRNAPredict refer to Livny *et al.* (15).

New features of sRNAPredict2

(i) *Reported BLAST E- and score values*: sRNAPredict2 converts coordinates of overlapping regions of sequence conservation identified by BLAST into the coordinates of a single contiguous segment containing all of the overlapping regions. The *E*-value and score assigned to this contiguous segment of conservation correspond to the lowest *E*-value and the highest score of the overlapping regions it contains.

(ii) *Incorporation of QRNA analysis*: A putative sRNA identified by sRNAPredict2 is reported to correspond to a region of conserved secondary structure (denoted by a 'Y' in the 'QRNA?' column of the output file) if that sRNA overlaps any region predicted by QRNA to encode conserved secondary structure (reported as 'winner = RNA' in the QRNA output file).

(iii) *Improved Venn diagram function*: The Venn diagram function can be used to identify putative sRNA-encoding genes that were predicted in multiple independent searches. This is particularly useful when searching for putative sRNA-encoding genes that are conserved in multiple species. When putative terminators are used as one of the predictors of sRNA-encoding genes, two sRNAs predicted in independent searches are reported to correspond to the same sRNA when they are associated with the same terminator. If the respective predicted 5′ ends of these two sRNAs are different, the coordinates, BLAST score and *E*-value corresponding to the shortest of the sRNAs are reported. If either of the sRNAs is predicted to encode conserved secondary structure, the sRNA will be reported to be associated with conserved secondary structure in the final output file.

Search parameters

The maximum and minimum lengths of sRNAs reported in all searches were set to 550 and 60, respectively. The maximum gap allowed between the 3′ end of a region of conservation and the 5′ end of a predicted Rho-independent terminator was 20 bp. Unless otherwise noted, all other sRNAPredict and sRNAPredict2 search parameters (numbers 1–6 in the initial input file) were set to 0. In all sRNAPredict2 searches, the minimum BLAST score allowed was set to 0. BLAST comparisons were conducted using BLASTN 2.0 (20). Search parameters B and V were set to 10 000. Unless otherwise noted, all other search parameters were set to default values. RNAMotif searches were conducted using a motif descriptor provided by D. J. Ecker. TransTerm searches were conducted with the confidence threshold set to 96%. QRNA analyses were conducted using version 2.0.3d with the window size (w) set to 100 and the slide size (x) set to 50.

ORF, tRNA, rRNA and sRNA databases

ORF databases were obtained from the NCBI ftp database (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/) and, when available, from the TIGR ftp database (ftp://ftp.tigr.org/pub/data/Microbial_Genomes/). ORF databases available from NCBI include locus and product names for each ORF but do not include annotated genes with frame-shift mutations. The ORF databases obtained from TIGR do not include locus and product names for each ORF but do include the coordinates of annotated genes with frame-shift mutations. Moreover, they may include annotated ORFs not found in the NCBI database. tRNA and rRNA coordinates were obtained from TIGR. The coordinates of putative and confirmed sRNAs and of predicted riboswitches were obtained from Rfam (http://www.sanger.ac.uk/Software/Rfam/) (11). All *E.coli* sRNA sequences were obtained from the EcoCyc database (http://ecocyc.org/) (21).

RNA isolation and northern analysis

Cultures of *P.aeruginosa* strain PAO1 were grown in Luria–Bertani (LB) at 37°C with shaking at 300 r.p.m. either to exponential (OD_{600} = 0.450) or stationary phase (OD_{600} = 5.700). Total RNA was isolated with Trizol (Invitrogen) according to the manufacturer's protocol. Total RNA per lane (7–10 µg) was fractionated on 6% polyacrylamide urea gels and transferred to BrightStar-Plus nylon membrane (Ambion). RNA was crosslinked to the membrane with UV light. Northern analysis was conducted according to the protocol accompanying the Ultrahyb-oligo buffer (Ambion). All hybridizations were conducted at 40°C. All washes were conducted with 2× SSC 0.5% SDS at room temperature. Northerns were repeated for all confirmed sRNAs.

The sequences of DNA oligonucleotides used as probes for the northern blots are provided in the Supplementary Table S1. Oligonucleotides were designed to be complementary to sequences near the predicted sRNA terminator. T4 polynucleotide kinase (New England Biolabs) was used to end-label ∼0.5–1 pmol of each synthetic DNA oligonucleotide with a 2-fold molar access of [γ-^{32}P]ATP (Perkin Elmer). Radiolabeled oligonucleotides were purified using Sephadex G-25 gel filtration columns (Amersham).

RESULTS

Prediction of *P.aeruginosa* sRNA-encoding genes

We chose to test the utility of sRNAPredict for annotation of candidate sRNA-encoding genes in the IGRs of *P.aeruginosa* PA01 for two reasons. First, the annotated complete genomes of three additional *Pseudomonas* species, *Pseudomonas fluorescens* SBW25, *Pseudomonas putida* KT2440 and *Pseudomonas syringae* DC3000, were available for assessing conservation of *P.aeruginosa* IGR sequences. Second, the *P.aeruginosa* PA01 genome is relatively rich in predicted Rho-independent terminators. We used these putative Rho-independent terminators and regions of conservation to identify putative sRNA-encoding genes in the IGRs of *P.aeruginosa*. Initially, regions of conservation were identified by BLAST searches ($E < 1 \times 10^{-5}$) comparing *P.aeruginosa* intergenic sequences with the complete

genomes of *P.fluorescens*, *P.putida* and *P.syringae*, respectively. Three separate searches for putative sRNAs were conducted, each aimed at identifying putative Rho-independent terminators located within or <20 bp downstream of intergenic sequence conserved between *P.aeruginosa* and one of the other *Pseudomonas* species, respectively. Predicted sRNAs were compared to a database of putative riboswitches, putative sRNA-encoding genes and genes encoding four previously confirmed *P.aeruginosa* sRNAs [Prrf1 and Prrf2 (17,22), RsmY (23,24), and RsmZ (25)]. Subsequently, the Venn diagram function of sRNAPredict was used to identify sRNAs predicted in multiple searches. In total, 38 distinct candidate sRNA-encoding genes were predicted, including all 4 of the previously confirmed *P.aeruginosa* sRNAs.

Experimental testing of predicted *P.aeruginosa* sRNAs

To test the accuracy of our predictive searches, 31 of the candidate *P.aeruginosa* sRNAs were subjected to northern analysis (Table 1). For these experiments, ^{32}P-labeled DNA oligonucleotide probes targeting the predicted sRNAs were hybridized to RNA purified from either exponential or stationary phase cultures of *P.aeruginosa* PA01 grown in LB (Figure 1). The previously confirmed *P.aeruginosa* sRNA RsmZ (25) was used as a positive control. Transcripts <400 nt were observed for 21 of the candidate sRNAs. For each of the candidates, the size(s) of its associated transcript(s) was compared with its distance from and the sizes of its flanking ORFs, tRNA or rRNA-encoding genes (Table 1). Based on these analyses, we concluded that the transcript observed for P14 likely corresponds to the mRNA of the upstream ORF PA2853. For P1 and P24, this analysis suggested that while the observed transcripts were relatively large (∼300 nt), they did not correspond to mRNAs, as the genes encoding P1 and P24 are flanked by ORFs that are either on the opposite strand or that encode mRNAs much larger than the transcripts observed (Table 1).

The signals detected for several of the sRNAs were relatively weak, raising concern that they were due to non-specific probe hybridization. We therefore repeated northern analysis of 10 of the candidate sRNAs (boldface in Table 1) using oligonucleotide probes targeting different regions of the transcripts than those targeted by the first set of probes (data not shown). In all but one of these northern assays (P13), transcripts sizes and intensities corresponded to those observed with the first set of probes. Although the respective sizes of the transcripts observed for P7 and P27 were similar with both sets of probes, the intensities of the signals observed for both candidates were particularly weak (Figure 1), making it difficult to rule out the possibility that the transcripts observed were due to non-specific hybridization. We therefore regard the results of our northern analyses of P7 and P27 as inconclusive. Taken together, our findings suggest that, of the 31 candidates subjected to northern analysis, 17 correspond to sRNAs (Figure 1). Thus, of the 38 sRNAs predicted in our annotation, 35 (including the 4 previously confirmed *P.aeruginosa* sRNAs) have been subjected to northern analysis and 21 of these (60%) have been physically confirmed.

As shown in Figure 1, multiple bands were detected for a number of sRNAs. For several of these sRNAs, including

Table 1. Predicted *P.aeruginosa* sRNAs subjected to experimental confirmation

sRNA name[a]	Start	End	Dir.	Length[b]	Approximate observed lengths[c]	Condition[d]	5′ Gene length	Dir.	Distance[e]	3′ Gene length	Dir.	Distance[f]
P1	334 549	334 728	<	179	300,350	E S	1374	<	93	750	>	5
P2	356 478	356 615	>	137	—	—	1227	<	0	1392	>	65
P4	747 255	747 469	<	214	—	—	1212	>	298	1104	>	0
P5	912 791	912 861	>	70	90	E S	1182	<	11	483	>	224
P6	971 645	972 154	<	509	—	—	1953	>	19	777	>	11
P7	971 858	971 972	>	114	140,150	E S	1953	>	232	777	>	193
P8	1 117 532	1 117 609	>	77	130	E S	753	>	141	1359	>	548
P9	1 436 491	1 436 618	>	127	130	E S	510	>	93	405	>	44
P10	1 807 682	1 807 753	>	71	—	—	405	>	33	1578	>	439
P11	1 928 666	1 928 886	<	220	100	E S	2448	<	39	1593	<	7
P13	3 106 919	3 106 994	<	75	70	E S	1920	<	167	861	<	6
P14	3 206 733	3 206 877	>	144	300	—	87	>	235	249	>	36
P15	3 299 022	3 299 269	>	247	180	E S	1014	<	101	1092	>	221
P16	3 318 663	3 318 859	>	196	110	S	1131	>	6	774	<	22
P17	3 677 435	3 677 716	<	281	—	—	312	<	354	609	<	2
P18	3 703 022	3 703 156	<	134	100	E S	855	<	72	1992	<	9
P19	3 705 268	3 705 582	>	314	—	—	1992	<	107	600	<	306
P20	3 705 315	3 705 622	<	307	150	—	1992	<	154	600	<	266
P24	4 444 696	4 444 952	<	256	300	E S	783	>	100	507	>	24
P25	4 444 893	4 444 961	<	68	—	—	783	>	297	507	>	15
P26	4 780 768	4 780 833	<	65	250	E S	4071	<	151	366	<	4
P27	4 781 786	4 781 978	>	192	90	E S	498	<	0	693	<	5
P28	4 956 328	4 956 536	>	208	180	E S	453	<	299	846	<	196
P30	5 308 743	5 308 964	<	221	180	E S	1434	>	318	1401	>	360
P31	5 344 903	5 344 981	>	78	—	—	1134	<	0	804	<	103
P32	5 344 950	5 345 060	<	110	80	E S	1134	<	47	804	<	24
P33	5 775 061	5 775 220	>	159	—	—	1428	>	254	465	<	398
P34	5 835 082	5 835 480	>	398	150	E S	2319	>	12	411	>	0
P35	4 985 782	4 985 843	<	61	55	E S	237	<	52	306	<	2
P36	5 308 433	5 308 493	>	60	75,80	E S	1434	>	8	1401	>	831
P37	5 672 302	5 672 363	<	61	—	—	4443	<	161	1653	<	1

[a]The names of sRNAs tested with two independent probes are in boldface.
[b]The predicted length of the candidate sRNA-encoding genes includes the length of the putative Rho-independent terminator.
[c]The approximate length(s) of the major species detected in each blot.
[d]Growth condition(s) under which the transcript(s) was observed (E, exponential phase; S, stationary phase). Boldface indicate transcript(s) was significantly more abundant under the indicated condition.
[e]The distance between the end of the upstream gene and the predicted start of the candidate sRNA.
[f]The distance between the start of the downstream gene and the predicted end of the candidate sRNA.
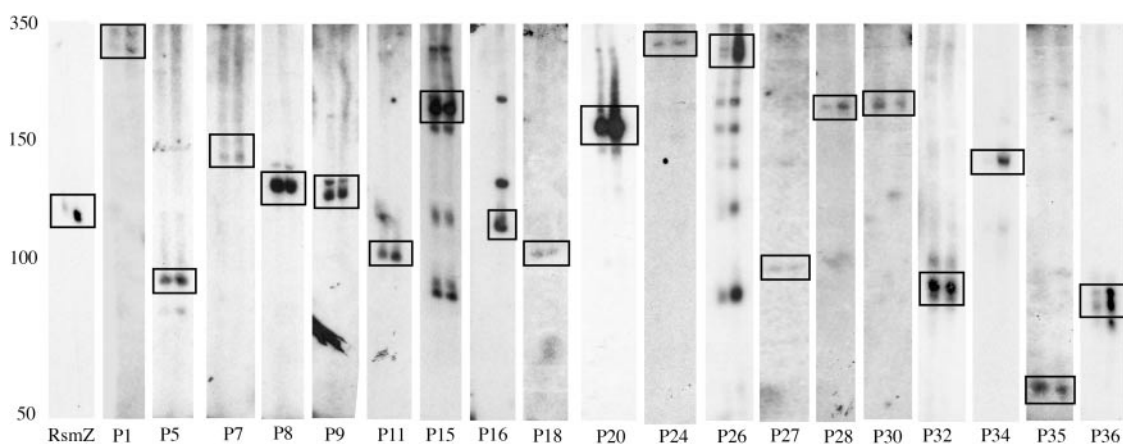

Figure 1. Detection of novel sRNAs by northern analysis. Total RNA was extracted from cultures of *P.aeruginosa* strain PAO1 grown in LB to exponential phase (first lane in each blot) or stationary phase (second lane in each blot). Blots were hybridized to radiolabeled DNA oligonucleotide probes and then exposed for varying times; thus the relative intensities of the signals do not correspond to the relative abundance of each sRNA. The approximate positions of size standards are shown on the left. Boxes are included to highlight the major species observed in each blot.

P15, the same pattern of bands was detected using probes targeting different regions of the sRNA (data not shown), suggesting that these multiple signals were due to sequence-specific hybridizations of the probe to multiple transcripts. In some cases, the observation of multiple bands may reflect an overlap between the gene encoding the sRNA and sequences encoding the untranslated regions (UTRs) of an adjacent ORF. This is likely the explanation for the multiple bands

observed for P26, which is encoded on the same strand as and in close proximity to both its flanking ORFs (Table 1). In other cases, as described for the *E.coli* sRNAs IstR-1 and IstR-2 (26), detection of multiple species may be due to the presence of two overlapping sRNA-encoding genes transcribed from adjacent promoters and sharing a common transcriptional terminator. Alternatively, as recent findings in our laboratory have shown, the existence of several overlapping sRNA species may be the result of post-transcriptional processing of a single sRNA transcript (B. Davis, unpublished data). Several of the sRNAs we detected were significantly more abundant in stationary phase cultures (e.g. P16, P34 and P36), suggesting that transcription of these sRNAs may be regulated by the stationary phase sigma factor RpoS.

As shown in Table 1, the observed sizes of the sRNA transcripts were, in some cases, significantly different from their predicted size. An overestimation of transcript size may result from extension of the conservation of the sRNA-encoding genes beyond the boundaries of the gene into upstream regulatory regions. Underestimation of transcript size may be due to an overlap between the sRNA-encoding gene and an adjacent ORF or conservation of only a portion of the sRNA. For example, the *V.cholerae* sRNA RyhB was initially predicted to be ~80 nt long based on sequence comparison between the *E.coli* RyhB and *V.cholerae* (11); recently, Davis *et al.* (27) showed the actual length of *V.cholerae* RyhB is ~225 nt.

It is possible that some of the 12 candidate sRNAs for which no small transcript was detected may correspond to real sRNAs that are poorly expressed or not expressed at all during *P.aeruginosa* growth in LB. To facilitate detection of those sRNAs transcripts that are only poorly expressed under the conditions tested, all blots were exposed to film for 4 days in the presence of a signal-intensifying screen. Previous studies analyzing dozens of sRNAs in several different species found that while some sRNAs are more abundant during growth in minimal media or under stress conditions, only one (OxyS) was not detectable by northern analysis during growth in LB (14). Based on these findings, the likelihood that even one of the *P.aeruginosa* sRNAs predicted in this study corresponds to a transcript that is detectable only under conditions not tested is very low. Therefore, we assume that the 12 predicted sRNAs for which no transcript <400 nt was detected correspond to false predictions.

Comparison of the *P.aeruginosa* sRNAs to *E.coli* and *V.cholerae* sRNAs and to the genomes of other bacterial species

The 38 predicted *P.aeruginosa* sRNAs were compared by BLAST ($E < 10$) to a database containing 42 confirmed *E.coli* sRNAs and 15 confirmed or putative *V.cholerae* sRNAs. P28 was found to have significant sequence homology with the genes in *E.coli* ($E = 1.5 \times 10^{-24}$) and *V.cholerae* ($E = 7.3 \times 10^{-25}$) encoding the sRNA RnpB. Thus, P28 likely corresponds to the *P.aeruginosa* RnpB, which in *E.coli* is a component of RNase P, the enzyme involved in the processing of 4.5S RNA and tRNA precursor molecules (28). No other *P.aeruginosa* sRNA predicted in this study was found to be homologous to other *E.coli* or *V.cholerae* sRNAs. Interestingly, the previously identified *P.aeruginosa* sRNA RsmZ may be a homolog of the *E.coli*

sRNA RygD (14,29), as the two sequences have some similarity ($E = 0.78$).

The 38 predicted sRNAs were also BLASTed against all sequenced bacteria ($E < 1$). Most did not have homologs outside of the three sequenced Pseudomonas species. However, in addition to P28, which was conserved across a wide variety of species, P9 shared significant similarity ($E = 2 \times 10^{-7}$) with intergenic sequences in both *Bordetella bronchiseptica* and *Bordetella parapertussis*. The degree of sequence similarity between P9 and the two *Bordetella* sp. as measured by *E*-value was nearly 4 logs greater than the sequence similarity shared between P9 and any other species, including other *Pseudomonas* species. Since *P.aeruginosa*, *B.bronchiseptica* and *B.parapertussis* are all respiratory pathogens, it is tempting to speculate that P9 may have a specific and conserved role in mediating *P.aeruginosa* responses to the host respiratory tract.

Numerous bacterial species encode a homolog of the *E.coli ssrA* gene (which encodes tmRNA) (10,30) so it was somewhat surprising that none of the predicted *P.aeruginosa* sRNA-encoding genes corresponded to homologs of the *E.coli ssrA*. The *P.aeruginosa ssrA* is predicted both in Rfam and in the tmRNA website (http://www.indiana.edu/~tmrna/) in a region annotated by TIGR as a tRNA. When this tRNA was removed from our database of putative tRNAs and rRNAs, the gene encoding the *P.aeruginosa* tmRNA homolog was identified in our search. These findings suggest that the *P.aeruginosa* was missed in our search because it was misannotated in the TIGR database as 'tRNA-Pseudo-1' and thus was not included in our database of IGR sequences.

Features that distinguish confirmed from unconfirmed sRNAs

We sought to determine if the experimentally confirmed *P.aeruginosa* sRNAs shared any common attributes that distinguished them from sRNAs that were not detected in the northern analysis, the presumed false predictions. The features we examined included (i) the degree to which the sRNAs are conserved, (ii) the number of BLAST partners in which the sRNAs are conserved, (iii) the predictive program(s) used to identify the terminators associated with the sRNAs, (iv) the distance of the sRNAs from their flanking ORFs, tRNAs or rRNAs, and (v) whether the sRNAs were predicted by QRNA to encode conserved secondary structure. QRNA is a program that utilizes BLAST-generated sequence alignments to identify patterns of sequence homology that likely represent conservation of RNA secondary structure (29,31). Both P7 and P27 were excluded from these analyses since the experimental testing of these candidates was inconclusive.

To facilitate the analysis of the experimentally tested *P.aeruginosa* sRNAs, we developed sRNAPredict2, a program that has five new features compared to sRNAPredict. First, when conservation is used as a predictor, each sRNA predicted by sRNAPredict2 is annotated with its associated BLAST score and *E*-value. The maximum *E*- and minimum score allowed for conserved regions in each search can be set in the initial input file. Second, the Venn diagram function of sRNAPredict, which was designed to compare candidate sRNAs predicted in a maximum of two independent searches,
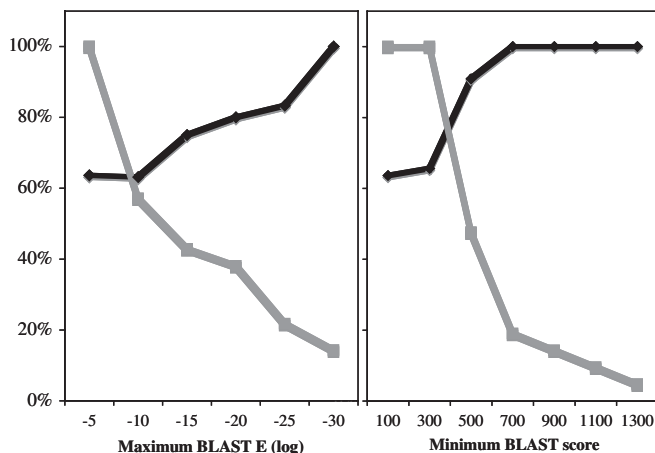
Figure 2. The accuracy (closed diamond) and sensitivity (closed square) of the predictive search increases and decreases, respectively, as the BLAST stringency is increased. The accuracy corresponds to the percentage of sRNAs predicted at the indicated BLAST stringencies that were confirmed. The sensitivity corresponds to the percentage of all 23 experimentally confirmed *P.aeruginosa* sRNAs that were predicted at the indicated BLAST stringencies.

Figure 3. Venn diagram showing the number of novel sRNAs confirmed per the number of novel sRNAs predicted based on conservation between *P.aeruginosa* and the three BLAST partner species used for comparison.

was replaced with a function that compares sets of sRNAs predicted in an unlimited number of searches and reports the total number of independent searches in which the sRNA was identified as well as the name of the BLAST partner(s) in which it was found to be conserved. Third, when terminators are used as predictors, sRNAPredict2 reports which program(s) was (were) used to identify the terminator associated with each predicted sRNA. Fourth, sRNAPredict2 can use QRNA output files to identify candidate sRNAs that are predicted to encode conserved secondary structures. Finally, the format of the sRNAPredict2 output file can be opened as a tab-delimited spreadsheet in Excel, facilitating the efficient sorting of predicted sRNAs by any of their associated features.

Using sRNAPredict2, we repeated the searches for candidate *P.aeruginosa* sRNAs ($E < 1 \times 10^{-5}$) and analyzed the predictions based on the following features:

(i) The degree to which the sRNA is conserved. The list of predicted *P.aeruginosa* sRNAs was sorted by E-value and accuracy of predictions versus the E-value was plotted. As shown in Figure 2, the percent of predicted sRNAs that were physically confirmed increased as the E-value decreased. In contrast, the sensitivity of the searches, as measured by the proportion of the 21 confirmed sRNAs predicted, decreased steadily as E-value decreased. Similar analyses revealed that increased BLAST score leads to increased accuracy and decreased sensitivity. Thus, predicted sRNAs that possess a higher degree of conservation are more likely to correspond to real sRNAs than those that posses a lower degree of conservation.

(ii) The number of BLAST partners in which the sRNA is conserved. We used the new Venn diagram function of sRNAPredict2 to identify sRNAs predicted based on conservation between *P.aeruginosa* and multiple BLAST partners (Figure 3). Excluding P7 and P27, 64% of all the sRNAs experimentally tested were confirmed while
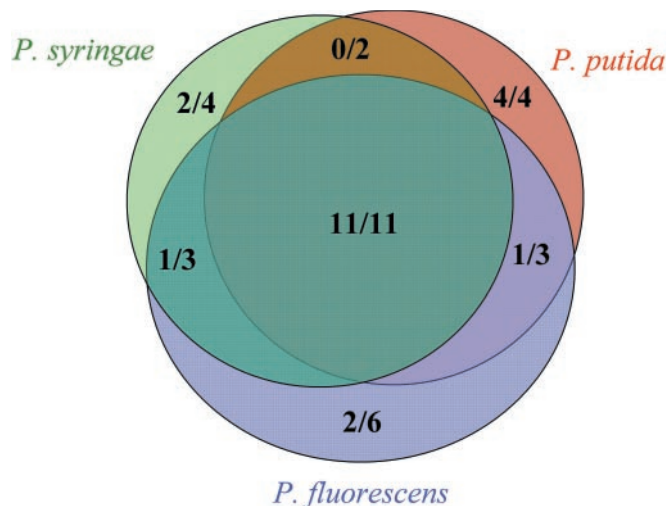
69 and 100% of the sRNAs predicted using more than one BLAST partner and more than two BLAST partners, respectively, were confirmed. These findings suggest that an sRNA predicted in searches based on conservation in multiple BLAST partners is less likely to be a false positive prediction than one that is predicted based on conservation in only one BLAST partner.

(iii) The predictive program used to identify the sRNA's terminator. We examined whether an sRNA associated with a terminator predicted by RNAMotif was more or less likely to correspond to a real sRNA than one associated with a terminator predicted by TransTerm. Of the 28 sRNA-encoding genes predicted based on a putative terminator identified by RNAMotif, 64% corresponded to physically confirmed sRNAs. Of the 11 sRNA-encoding genes predicted based on a putative terminator identified by TransTerm, the proportion that corresponded to physically confirmed sRNAs was also 64%. These findings suggest that a candidate sRNA associated with a terminator predicted by TransTerm is not significantly more or less likely to correspond to a real sRNA than one that is associated with a terminator predicted by TransTerm.

(iv) The distance between the sRNA and its flanking ORFs. In our annotations of both *V.cholerae* and *P.aeruginosa* sRNAs, as well as in previous predictions of sRNAs in *E.coli*, a number of predicted sRNA-encoding genes were found to be conserved UTRs of mRNAs (14,15). Thus we examined if sRNAs predicted near ORFs were more likely to be false predictions than those predicted far from annotated genes. The average distances between confirmed sRNAs and their flanking genes was not found to be greater than the average distance between unconfirmed sRNAs and their flanking genes (data not shown), suggesting that an sRNA-encoding gene predicted far from an ORF is no more likely to correspond to a real sRNA than one predicted near an ORF.

(v) Whether the sRNA is predicted by QRNA to encode conserved secondary structure. QRNA is a program

designed to identify secondary structure conservation that was used by Rivas *et al.* (29) to predict 275 sRNAs in the IGRs of *E.coli*, including 80% of the previously known *E.coli* sRNAs. Of the 49 candidates for novel sRNAs subjected to experimental verification in this study, transcripts were detected for only 11 (22%), suggesting that using predicted secondary structure conservation as the sole predictor of sRNA-encoding genes produces a relatively high proportion of false positive predictions. We sought to determine if integrating QRNA analysis into our predictive approach would enable us to distinguish between physically confirmed and unconfirmed *P.aeruginosa* sRNAs. BLAST alignments ($E < 1$) between the *P.aeruginosa* IGRs containing the 38 predicted sRNAs and the genomes of *P.syringae*, *P.putida* or *P.fluorescens*, respectively, were analyzed using QRNA to identify regions that appear to encode conserved RNA secondary structure. Of the candidate sRNAs 23 overlapped sequences predicted by QRNA to encode conserved secondary structure. Of these 20 corresponded to experimentally tested sRNAs and 17 (85%) corresponded to sRNAs experimentally confirmed either in this study or in previous studies. These observations suggest that candidate sRNAs predicted by QRNA to encode conserved secondary structure are more likely to correspond to real sRNAs than ones that are not. However, QRNA analysis only identified 17 of the 21 (81%) confirmed sRNAs, suggesting either that some confirmed sRNAs lack conserved secondary structures or that QRNA is unable to detect some conserved structures.

Prediction of sRNAs in other pathogens

Using sRNAPredict2, we annotated the IGRs of 10 diverse Gram-negative and Gram-positive pathogens for candidate sRNA-encoding genes. Since our analysis of the experimentally tested *P.aeruginosa* sRNAs suggested that predictions of novel sRNAs based on homology with more than one BLAST partner are more accurate than those based on homology with just one BLAST partner, we limited our analyses to those pathogens for which the genome sequences of at least 3 other family members were available. In the annotation of each genome, Rho-independent terminators adjacent to regions of sequence conservation ($E < 1 \times 10^{-5}$) between the species of interest and three to seven related species were used as predictors of sRNA-encoding genes. Remarkably, a total of 2912 candidate sRNA-encoding genes were predicted in these searches (Table 2). Only 153 of the predicted sRNAs correspond to previously annotated sRNAs or riboswitches. Of the 2759 novel sRNAs predicted, 1758 (64%) and 561 (21%) were predicted based on conservation in more than one and in more than two BLAST partners, respectively, and 2308 (84%) were predicted by QRNA to encode conserved RNA structure. Although these candidate sRNA-encoding genes represent the strongest candidates, our analysis of the *P.aeruginosa* sRNAs suggests that a smaller but significant proportion of the genes predicted based on conservation with only one BLAST partner and not predicted to encode conserved structure are also likely to encode sRNA

transcripts. These annotations are available in Supplementary Tables S3–S12.

DISCUSSION

Prior to this study, genome-wide annotation for sRNA-encoding genes using integrative bioinformatic approaches had been performed in only a few of the nearly 300 sequenced bacterial species. Here, we conducted genome-wide searches for sRNA-encoding genes in the IGRs of 11 diverse pathogens, leading to the prediction of 2793 previously unannotated genes. The fact that so many of these predicted sRNAs, including all 17 of the physically confirmed *P.aeruginosa* sRNA, did not correspond to sRNAs annotated in the Rfam database highlights the inherent limitations of predicting sRNAs based solely on homology with previously confirmed sRNAs. Especially when considering the limitations of our predictive approach (see below), the large number of candidate sRNA-encoding genes predicted in our searches reveals critical gaps in the current annotations of bacterial genomes.

Clearly not all of our predictions correspond to real sRNAs; however, since 66% of the *P.aeruginosa* sRNAs predicted in this study and 56% of the *V.cholerae* sRNAs predicted in our previous study were detected by northern analysis, it seems reasonable to speculate that a significant proportion of our predictions in other species are also correct. Our *post hoc* analysis of the experimentally tested *P.aeruginosa* sRNAs suggests that within a set of sRNA-encoding genes identified by sRNAPredict those that are highly conserved, conserved in multiple species and/or predicted by QRNA to encode conserved secondary structure represent the strongest candidates for bona fide sRNA transcripts. The databases of candidate sRNAs provided in Supplementary Data can be sorted by any of the characteristics listed above, allowing researchers studying any of the pathogens annotated in this study to quickly and easily identify the strongest candidate sRNAs in their species of interest.

Although nearly 3000 sRNAs were predicted in our annotations, some sRNAs were undoubtedly missed. As shown in Table 2, only about half of sRNAs annotated in the Rfam database were identified in our search. For example, 22 of the 39 previously annotated sRNAs in *Salmonella enterica* were identified. Interestingly, all 39 of these sRNAs were identified when IGR sequence conservation was used as the only predictor, suggesting that most of the previously annotated sRNAs missed in our search were not identified because they are not associated with putative Rho-independent terminators. Some of the sRNAs missed in our searches may be associated with a Rho-dependent terminator. Others may be associated with Rho-independent terminators that conform poorly to the consensus motifs used by RNAMotif and TransTerm to identify putative terminators. These motifs are based primarily on the structure of Rho-independent terminators in *E.coli*. Since the structure of Rho-independent terminators may be different in other species, these two programs may be significantly less effective in identifying terminators in certain species. Our findings suggest that the sensitivity of our predictive approach is limited by its reliance on putative Rho-independent terminators as predictors of sRNA-encoding genes.

Table 2. Summary of sRNAPredict2 annotations for sRNA-encoding genes in 11 species of pathogens

Species[a]	No. of BLAST partners[a]	Total no. of sRNAs predicted[b]	No. of unique sRNAs predicted[c]	No. of novel sRNAs predicted…	…conserved in >1 partner	…conserved in >2 partners	…encoding predicted conserved RNA structure	Total kb IGR conservation[e]	Proportion of previously annotated sRNAs predicted	No. of IG terminators	No. of unique sRNA/Mb genome	No. of unique sRNA/10 kb IGR	No. of unique sRNAs/IG terminator	No. of total sRNAs/100 kb IGR conservation
B.anthracis	6	1828	947	889	688	23	858	1235.4	2/3	2895	181.1	12.73	0.327	14.80
Y.pestis	7	1478	772	755	443	221	550	1627.5	13/24	1201	167.8	9.45	0.643	10.62
S.enterica	7	1284	599	572	321	145	460	1486.3	22/39	1289	116.8	7.73	0.465	8.64
Staphylococcus aureus	3	344	168	144	84	48	106	366.2	3/4	819	59.8	3.36	0.205	9.39
Clostridium tetani	4	426	161	138	108	77	127	534.3	0/3	852	56.1	4.22	0.189	7.97
Streptococcus pneumoniae	5	111	63	62	29	13	48	78.8	0/3	675	29.2	2.52	0.093	14.08
Streptococcus pyogenes	5	126	62	56	26	19	49	247.1	2/3	613	33.5	2.28	0.101	5.10
Mycobacterium tuberculosis	4	78	56	56	13	6	47	199.6	0/3	114	12.7	1.75	0.491	3.91
Helicobacter pylori	6	130	50	50	42	29	8	114.6	0/3	126	29.9	3.60	0.397	11.35
C.trachomatis	5	51	43	43	4	3	30	81.2	0/3	167	41.3	4.58	0.257	6.28
P.aeruginosa	3	72	38	34	19	13	25	152.3	5/6	1018	6.1	0.59	0.037	4.73
Total (10 genomes[f])		6073	2912	2759	1758	584	2308							
Max. fold difference[g]		35.8	20.3	22.2							29.7	21.6	17.4	3.8

[a]Strain names of the species annotated and of their BLAST partners as well as the sources of their genomic sequence files are listed in Supplementary Table S2.
[b]The total number of sRNA-encoding genes predicted in all searches between the species of interest and each of its BLAST partners.
[c]The number of distinct sRNA-encoding genes predicted in the species of interest as determined by using the sRNAPredict2 Venn diagram function.
[d]The number of predicted sRNAs that do not correspond to riboswitches or sRNAs annotated in the Rfam database.
[e]The amount of genome sequence conserved between the species of interest and all partner species.
[f]Not including *P.aeruginosa.*
[g]The fold difference between the largest and smallest values in each column.

Another significant limitation of our predictive approach stems from its reliance on sequence conservation. First, this effectively limits our annotations to regions of the genome that do not encode proteins or structural RNAs on either strand. Second, sensitive and accurate prediction of sRNAs based on sequence conservation requires that the evolutionary distance between the species of interest and its BLAST partners be appropriate. More specifically, the evolutionary distance between the two species must be close enough so that sRNA-encoding genes are conserved but far enough so that intergenic sequences which do not encode sRNAs no longer share significant homology. For some species, the genome sequences of appropriately diverged BLAST partners are not currently available. For example, we were unable to annotate sRNA-encoding genes in the Gram-negative pathogen *Francisella tularensis*, as no other *Francisella* sp. genome sequences were available. Other species, such as *S.enterica* and *Bacillus anthracis*, are members of families in which numerous species have been sequenced. For these species, it is difficult to determine which of the many potential BLAST partners will yield reliable predictions of sRNA-encoding genes. As more of the candidate sRNAs are experimentally tested, we may be able to define parameters of evolutionary divergence that will allow us to identify BLAST partners that will produce the most accurate and sensitive annotations for sRNA-encoding genes.

As shown in Table 2, the number of candidate sRNAs predicted in each of the 11 genomes analyzed varied significantly, from 947 in *B.anthracis* to 38 in *P.aeruginosa*. The number of sRNAs predicted in each species did not correlate with either the overall size of the genome, the amount of intergenic sequence in the genome or the number of BLAST partners used (Table 2). Although some of the variations in the number of sRNAs predicted among these diverse pathogens can be explained by the variation in the number of predicted intergenic terminators and by the total amount of conserved intergenic sequence, these two features do not fully account for the wide range in the number of predicted sRNAs in the 11 genomes analyzed (Table 2). A possible explanation for these discrepancies is that some species, such as *B.anthracis* and *Yersinia pestis*, are inherently richer in sRNA-encoding genes than others, such as *P.aeruginosa* and *Chlamydia trachomatis*, much like certain species, such as *Streptomyces coelicolor*, encode many more alternative sigma factors per base pair of genomic sequence than others, such as *Mycoplasma* sp. (32).

Another explanation for the discrepancies in the numbers of sRNAs predicted may be that the accuracy and sensitivity of annotations among the species analyzed varies significantly. For some species, the BLAST partners used may be more effective in the prediction of sRNAs than for others. For example, the evolutionary relationships between *B.anthracis* and one or more of its BLAST partners may be too close, leading to a high rate of false positive predictions. Similarly, the relatively low number of sRNAs predicted in *P.aeruginosa* may be due to the fact that *P.aeruginosa* is too evolutionarily distant from its BLAST partners. Alternatively, as discussed above, the differences in the number of sRNAs predicted may be due to the fact that the proportion of sRNAs associated with Rho-independent terminators varies significantly among the species analyzed. Determining whether the wide range in the numbers of sRNAs predicted in the 11 species analyzed reveals real differences in the density of sRNA-encoding genes among bacterial species or simply reflects the limitations of our approach in predicting sRNAs in certain species versus others will require subjecting more of the sRNA candidates predicted in our annotations to experimental verification.

As shown in Table 2, over 700 sRNA-encoding genes were predicted in both *B.anthracis* and *Y.pestis*. If the accuracy of our predictions in these organisms is similar to the accuracy of our predictions in *V.cholerae* and *P.aeruginosa*, our annotations would include several hundred bona fide sRNAs for each of these species. Although this may seem unlikely considering that <100 sRNAs have been confirmed in *E.coli*, it is important to note that the number of *E.coli* sRNAs that has been subjected to experimental verification represents only a small proportion of all *E.coli* sRNAs predicted to date. Indeed, Hershberg *et al*. (10) have reported that a total of 1001 non-redundant predicted *E.coli* sRNAs remain untested, suggesting that even in *E.coli* the total number of sRNAs remains unclear. Thus, without subjecting a significant portion of the sRNAs predicted in *B.anthracis* and *Y.pestis* to experimental verification, it is difficult to determine whether or not the relatively high number of sRNAs predicted in these species reflects a relatively high rate of false positive predictions.

Reliable annotation of bacterial genomes for sRNA-encoding genes has several important implications. Knowing where putative sRNAs are encoded will aid in the design of more precise genetic manipulations. Moreover, the identification of a putative sRNA-encoding gene in a sequence of interest may help explain a phenotype whose dependence on a protein-encoding gene cannot be demonstrated. Most importantly, annotation for sRNA-encoding genes will greatly facilitate the identification and characterization of previously unknown sRNAs. As the number of confirmed sRNAs and the variety of species in which sRNAs have been studied continue to grow, larger issues in sRNA biology, such as sRNA evolution, can be addressed. By continuing to develop more effective bioinformatic algorithms and more efficient and accessible computational tools for predicting sRNA-encoding genes, we hope to ensure that sensitive and accurate annotation for sRNAs will one day become a standard feature of every sequenced bacterial genome.

PROGRAM AVAILIBILITY

sRNAPredict2 is written in C++. The sRNAPredict2 source code, user instructions and Mac OS X-compatible executable are available for download at http://www.tufts.edu/sackler/waldorlab/sRNAPredict/.

SUPPLEMENTARY DATA

Supplementary Data available at NAR online.

ACKNOWLEDGEMENTS

REFERENCES

1. Dennis,P.P. and Omer,A. (2005) Small non-coding RNAs in archaea. *Curr. Opin. Microbiol.*, **8**, 685–694.
2. Gottesman,S. (2005) Micros for microbes: non-coding regulatory RNAs in bacteria. *Trends Genet.*, **21**, 399–404.
3. 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.
4. Brown,L. and Elliott,T. (1996) Efficient translation of the RpoS sigma factor in *Salmonella typhimurium* requires host factor I, an RNA-binding protein encoded by the hfq gene. *J. Bacteriol.*, **178**, 3763–3770.
5. Ding,Y., Davis,B.M. and Waldor,M.K. (2004) Hfq is essential for *Vibrio cholerae* virulence and downregulates sigma expression. *Mol. Microbiol.*, **53**, 345–354.
6. Nakao,H., Watanabe,H., Nakayama,S. and Takeda,T. (1995) yst gene expression in *Yersinia enterocolitica* is positively regulated by a chromosomal region that is highly homologous to *Escherichia coli* host factor 1 gene (hfq). *Mol. Microbiol.*, **18**, 859–865.
7. Robertson,G.T. and Roop,R.M.J. (1999) The *Brucella abortus* host factor I (HF-I) protein contributes to stress resistance during stationary phase and is a major determinant of virulence in mice. *Mol. Microbiol.*, **34**, 690–700.
8. Sonnleitner,E., Hagens,S., Rosenau,F., Wilhelm,S., Habel,A., Jager,K.E. and Blasi,U. (2003) Reduced virulence of a hfq mutant of *Pseudomonas aeruginosa* O1. *Microb. Pathog.*, **35**, 217–228.
9. 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.
10. Hershberg,R., Altuvia,S. and Margalit,H. (2003) A survey of small RNA-encoding genes in *Escherichia coli*. *Nucleic Acids Res.*, **31**, 1813–1820.
11. Griffiths-Jones,S., Moxon,S., Marshall,M., Khanna,A., Eddy,S.R. and Bateman,A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, **33**, D121–D124.
12. 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.
13. Chen,S., Lesnik,E.A., Hall,T.A., Sampath,R., Griffey,R.H., Ecker,D.J. and Blyn,L.B. (2002) A bioinformatics based approach to discover small RNA genes in the *Escherichia coli* genome. *Biosystems*, **65**, 157–177.
14. 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.
15. 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.
16. Pichon,C. and Felden,B. (2005) Small RNA genes expressed from *Staphylococcus aureus* genomic and pathogenicity islands with specific expression among pathogenic strains. *Proc. Natl Acad. Sci. USA*, **102**, 14249–14254.
17. Wilderman,P.J., Sowa,N.A., FitzGerald,D.J., FitzGerald,P.C., Gottesman,S., Ochsner,U.A. and Vasil,M.L. (2004) Identification of tandem duplicate regulatory small RNAs in *Pseudomonas aeruginosa* involved in iron homeostasis. *Proc. Natl Acad. Sci. USA*, **101**, 9792–9797.
18. Macke,T.J., Ecker,D.J., Gutell,R.R., Gautheret,D., Case,D.A. and Sampath,R. (2001) RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res.*, **29**, 4724–4735.
19. Ermolaeva,M.D., Khalak,H.G., White,O., Smith,H.O. and Salzberg,S.L. (2000) Prediction of transcription terminators in bacterial genomes. *J. Mol. Biol.*, **301**, 27–33.
20. 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 Res.*, **25**, 3389–3402.
21. Keseler,I.M., Collado-Vides,J., Gama-Castro,S., Ingraham,J., Paley,S., Paulsen,I.T., Peralta-Gil,M. and Karp,P.D. (2005) EcoCyc: a comprehensive database resource for *Escherichia coli*. *Nucleic Acids Res.*, **33**, D334–D337.
22. Alifano,P., Rivellini,F., Limauro,D., Bruni,C.B. and Carlomagno,M.S. (1991) A consensus motif common to all Rho-dependent prokaryotic transcription terminators. *Cell*, **64**, 553–563.
23. Heeb,S., Heurlier,K., Valverde,C., Cámara,M., Haas,D. and Williams,P. (2004) Post-transcriptional regulation in *Pseudomonas* spp. via the Gac/Rsm regulatory network. In Ramos,J.L. (ed.), *The Pseudomonads*.. Kluwer Academic/Plenum Publishers, NY,Vol. 2, pp. 239–255.
24. 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.
25. 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.
26. Vogel,J., Argaman,L., Wagner,E.G. and Altuvia,S. (2004) The small RNA IstR inhibits synthesis of an SOS-induced toxic peptide. *Curr. Biol.*, **14**, 2271–2276.
27. Davis,B.M., Quinones,M., Pratt,J., Ding,Y. and Waldor,M.K. (2005) Characterization of the small untranslated RNA RyhB and its regulon in *Vibrio cholerae*. *J. Bacteriol.*, **187**, 4005–4014.
28. Gopalan,V., Vioque,A. and Altman,S. (2002) RNase P: variations and uses. *J. Biol. Chem.*, **277**, 6759–6762.
29. Rivas,E., Klein,R.J., Jones,T.A. and Eddy,S.R. (2001) Computational identification of noncoding RNAs in *E.coli* by comparative genomics. *Curr. Biol.*, **11**, 1369–1373.
30. Withey,J.H. and Friedman,D.I. (2003) A salvage pathway for protein structures: tmRNA and *trans*-translation. *Annu. Rev. Microbiol.*, **57**, 101–123.
31. Rivas,E. and Eddy,S.R. (2001) Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, **2**, 8.
32. Rhodius,V.A., Suh,W.C., Nonaka,G., West,J. and Gross,C.A. (2005) Conserved and variable functions of the sigma(E) stress response in related genomes. *PLoS Biol.*, **4**, 757–774.