OMICS A Journal of Integrative Biology
Volume 11, Number 1, 2007

DOI: 10.1089/omi.2006.0005

Comparative Genomics of Small RNAs in Bacterial Genomes

STAN LUBAN and DAISUKE KIHARA

ABSTRACT

In recent years, various families of small non-coding RNAs (sRNAs) have been discovered by experimental and computational approaches, both in bacterial and eukaryotic genomes. Although most of them await elucidation of their function, it has been reported that some play important roles in gene regulation. Here we carried out comparative genomics analysis of possible sRNAs that are computationally identified in 30 bacterial genomes from γ- and α-proteobacteria and *Deinococcus radiodurans*. Identified sRNAs are clustered by a complete-linkage clustering method to see conservation among the organisms. On average, sRNAs are found in approximately 30% of intergenic regions of each genome sequence. Of these, 25.7% are conserved among three or more organisms. Approximately 60% of the conserved sRNAs do not locate in orthologous intergenic regions, implying that sRNAs may be shuffled their positions in genomes. The current study implies that sRNAs may be involved in a more extensive range of functions in bacteria.

INTRODUCTION

In addition to transfer RNAs and ribosomal RNAs, a surprising number of small non-coding RNAs (sRNAs) have been found in recent years (Eddy, 2001; Editorial, 2004; Gottesman, 2004a, 2004b; Storz et al., 2004; Wassarman et al., 1999; Wassarman, 2004). sRNAs are nontranslated RNA molecules (*i.e.,* nonmessenger RNAs) that are found to be present and functional in many different organisms, ranging from bacteria to mammals (Wagner and Flardh, 2002). Non-coding RNAs have drawn significant attention in recent years because of their important roles in controlling many regulatory pathways in higher eukaryotes (Kim, 2005) and their practical and clinical importance (Eckstein, 2005; Hammond, 2006). Arguably, sRNAs are one of the most important findings in biology in the past decade; they have opened up a new view of gene regulation.

In eukaryotes, hundreds of sRNAs, termed *microRNAs* (miRNAs), and small interfering RNAs (siRNAs), which are 21–25 nucleotide-long RNAs that negatively regulate gene expression at the post-transcriptional level, have been found in the past couple of years (Ambros et al., 2003; He and Hannon, 2004; Kim, 2005). These examples include *lin-4* regulatory RNA in *C. elegans* (Wightman et al., 1993), small nucleolar RNA (snoRNA), which is involved in rRNA modification (Bachellerie et al., 2002), and *bantam* RNA found in

Department of Computer Science, Department of Biological Sciences, Markey Center for Structural Biology, West Lafayette, Indiana.

Drosophila melanogaster, which suppresses apoptosis and stimulates cell proliferation (Brennecke et al., 2003). Recently non-coding RNAs found in human and mouse have also drawn much attention (Wassarman et al., 2001).

In *Escherichia coli,* so far more than 50 sRNAs are identified (Griffiths-Jones et al., 2005a). The functions of many of the sRNAs remain to be elucidated, although studies of a subset of the sRNAs indicate that they act by three general mechanisms (Storz et al., 2004). A few are integral parts of RNA-protein complexes, such as the 4.5S RNA component of the signal recognition particle and RNase P RNA (Peluso et al., 2000). Another class of sRNAs mimics the structures of other nucleic acids. Examples of this class include the 6S RNA, which binds to a70-RNA polymerase holoenzyme (Wassarman and Storz, 2000). The third class of sRNAs act by base-pairing with other RNAs, thereby regulating expression of genes. sRNAs base-pairing with a target mRNA can have multiple regulatory outcomes in *E. coli*. MicF RNA prevents translation (Andersen and Delihas, 1990; Mizuno et al., 1984), while DsrA RNA facilitates translation (Sledjeski et al., 1996). RyhB RNA base-pairing with its targets is associated with degradation of the mRNAs (Masse and Gottesman, 2002), while base-pairing of some sRNAs with their targets is conceived to block access of a ribonuclease and thus stabilize the mRNA.

sRNAs have been identified by both experimental and computational methods. The experimental large-scale methods used to identify sRNAs include oligonucleotide arrays (Tjaden et al., 2002), EST screening (RNomics) (Huttenhofer et al., 2001), co-immunoprecipitation with the Hfq RNA-binding protein in *E. coli* (Zhang et al., 2003), and co-purification with ribonucleoproteins (Mourelatos et al., 2002). The computational methods used vary greatly, including artificial neural network, a sequence pattern recognition of sets of promoter or terminator motifs in intergenic regions (Argaman et al., 2001; Chen et al., 2002), and identification of conserved RNA secondary structures (Rivas et al., 2001). Several of these computationally identified sRNAs were confirmed by Northern blots. Current computational approaches may not be very accurate, but are still very important in the present situation where our understanding of sRNA in genomes is far from complete; computational analyses can quickly give us a rough idea about the distribution of sRNAs in genome sequences.

Several systematic searches for sRNAs in the *E. coli* genome have identified around 200–400 novel sRNAs: Chen and co-workers' (2002) computational approach produced 227 predictions, while Rivas et al. (2001) identified 275 candidates using probabilistic sequence models compiled as the QRNA program. A neural network–based approach predicted 370 sRNAs (Carter et al., 2001). Recently Zhang et al. (2004) estimated the number of sRNA to be in the range of 118–260 from sequence conservation of intergenic regions. In addition, oligonucleotide arrays by Tjaden et al. (2002) found 334 transcripts of unknown function in intergenic regions. Hershberg et al. (2003) compiled five earlier works (Argaman et al., 2001; Chen et al., 2002; Huttenhofer et al., 2001; Rivas et al., 2001; Zhang et al., 2004), resulting in 1001 nonredundant sRNA candidates. They found that there is not much overlap between the previous studies. Indeed 906 out of 1001 candidates were predicted by one study alone. Hence, it is still not clear how many sRNAs *E. coli* holds.

In this study, we have addressed the following questions: how many sRNAs exist in microbial genome sequences, and how many of them are well conserved among organisms? How well are the locations of sRNAs conserved among genomes? To answer these fundamental questions about microbial sRNAs, we conducted a comparative genome analysis of sRNAs in 30 microorganisms including γ- and α-proteobacteria and Deinococcus-thermus. To the best of our knowledge, this is the first attempt of comparative genomics on sRNAs in bacterial genomes. Comparative genomics is a powerful computational approach to verify predictions made on a single genome, because it is natural to consider that predicted sRNAs showing conservation among multiple organisms are more likely to be correct. A comparative analysis is especially useful for sRNAs, because computational identification of sRNAs is hard due to their less significant statistical signals compared with protein-coding regions (Benson et al., 2004).

We found a large number of sRNAs in each of the 30 organisms, ranging from 265 to 1789 (i.e., on average an sRNA is found in approximately 30% of intergenic regions in each genome). The initially found sRNAs were clustered into "families" (see below section) shared by multiple organisms. There were 869 sRNAs that were included in families shared in five or more organisms. In the *E. coli* genome, a total of 1078 sRNAs was identified. Among them, 380 sRNAs were shared with two or more organisms. The number of 380 roughly agrees with the predicted number of sRNAs in *E. coli* determined in a previous study

(Carter et al., 2001). The current study revealed that there are many more regions than those already known that seem to be able to form RNA secondary structures in the bacterial genomes. Therefore it is expected that more sRNAs will be found to be involved in gene regulatory mechanisms in bacteria by experiments conducted in the near future.

MATERIALS AND METHODS

Genome dataset

We analyzed intergenic regions of 30 complete genome sequences, which were retrieved from Genbank (Benson et al., 2004). The following 25 organisms were selected from the γ-proteobacteria: *Acinetobacter sp. ADP1* (As), *Blochmannia floridanus* (Bf), *Buchnera aphidicola str. Bp.* (Ba), *Coxiella burnetii* (Cb), *Erwinia carotovora* (Ea), *Escherichia coli* K-12 (Eo), *Haemophilus ducreyi* (Hd), *Haemophilus influenzae* (Hi), *Pasteurella multocida* (Pm), *Photorhabdus luminescens* (Pl), *Pseudomonas aeruginosa* (Pa), *Pseudomonas putida* (Pp), *Pseudomonas syringae* (Ps), *Salmonella enterica subsp. enterica serovar Typhi* (Se), *Salmonella typhimurium LT2* (St), *Shewanella oneidensis* (So), *Shigella flexneri* (Sf), *Vibrio cholerae* (Vc), *Vibrio parahaemolyticus* (Vp), *Vibrio vulnificus* (Vv), *Wigglesworthia brevipalpis* (Wb), *Xanthomonas axonopodis* (Xa), *Xanthomonas citri* (Xc), *Xylella fastidiosa* (Xf), *and Yersinia pestis* (Yp). From α-proteobacteria, four organisms—*Agrobacterium tumefaciens* (At), *Brucella melitensis* (Bm), *Caulobacter crescentus* (Cc), and *Mesorhizobium loti* (Ml)—were selected. Finally, *Deinococcus radiodurans* (Dr) was added to the list of organisms from the Deinococcus-thermus, deinococci. (Descriptors of the genomes used in Figure 4 are shown in the parentheses.) Key characteristics of the genomes are summarized in Table 1.

The choice of the genomes was determined by the following: *E. coli* K-12 and its closely related organisms were selected because one of the main foci of this study is to see how many sRNAs in *E. coli* are shared with the other organisms. However, different strains of *E. coli*, such as *E. coli* O153 or *E. coli* CFT073, were not selected because they are too close and the characteristic mutation patterns of RNA regions cannot be well detected by the QRNA program (see below). The *D. radiodurans* genome was added in order to assess the performance of our comparative genomics approach to a distantly related genome. From all of the genome sequences, intergenic regions, which are simply defined as DNA sequences between two adjacent gene-coding regions, were extracted. Definitions of positions of genes were taken from the Genbank annotation. The total number of intergenic regions is 94,464, with an average number of intergenic regions per organism of 3148.8. The total combined length of the intergenic regions is 16,663,732 nt, with an average length of 176.4 nt.

Identification of sRNAs

The entire procedure of identifying and clustering sRNA families is illustrated in Figure 1A. We used the QRNA program to extract potential sRNA regions from the intergenic regions (Rivas and Eddy, 2001). QRNA classifies an input pairwise alignment as either RNA region, coding region, or random sequence, using pair-hidden Markov models and a pair stochastic context-free grammar. Since this algorithm detects the structural characteristics of sRNAs and does not rely on the homology to known sRNA sequences, completely unknown sRNAs are also expected to be captured. The procedure begins with an all-against-all nucleotide-nucleotide BLAST search (Blastall Program, version 2.2.8, http://www.ncbi.nlm.nih.gov/BLAST/) (Altschul et al., 1990) performed on the intergenic regions with the default parameter set. Pairwise alignments from the search with more than the sequence identity of 65% and an E value of <0.01 are kept for the subsequent steps. To reduce false positives, pairwise alignment candidates for sRNAs are eliminated if the alignment does not exist bidirectionally (when query and subject entries are switched and the algorithm is re-run). Then QRNA (version 2.0.2c, http://selab.janelia.org) is run with the default parameters for each of the detected pairwise alignments, and those regions classified as RNA are extracted.

Next we concatenated adjacent sRNAs found in the same intergenic region if they were closer than 25 nt. The reason is because we observed that multiple loop regions of an sRNA tend not to be captured by QRNA as a single, large sRNA but recognized as separate sRNAs. This may be due to the intrinsic limitation of the pair stochastic context-free grammar. The threshold value of 25 nt was determined from known loop regions in the Rfam database. About 70% of the intervals between adjacent sRNAs fall below 25 nt

Table 1. Genome Sequences Used in this Study

Class/order	*Organism*	*Code*[a]	*Genome size (nt)*	*Number of genes*[b]	*Total length of intergenic regions (nt)*	*Genbank accession no.*
γ Proteobacteria/ Alteromonadales	*Shewanella oneidensis*	So	4969803	5016	668224	AE014299
γ/Enterobacteriales	*Erwinia carotovora*	Ea	5064019	4635	712214	BX950851
γ/Enterobacteriales	*Escherichia coli* K-12	Eo	4639221	4558	562552	NC_000913
γ/Enterobacteriales	*Blochmannia floridanus*	Bf	705557	673	106881	BX248583
γ/Enterobacteriales	Photorhabdus *luminescens*	Pl	5688987	5844	1817578	BX470251
γ/Enterobacteriales	*Salmonella enterica Typhi*	Se	4809037	4741	580762	NC_003198
γ/Enterobacteriales	*Salmonella typhimurium LT2*	St	4857432	4725	528319	NC_003197
γ/Enterobacteriales	*Shigella flexneri*	Sf	4607203	4595	1084323	AE005674
γ/Enterobacteriales	*Wigglesworthia brevipalpis*	Wb	697724	689	86185	NC_004344
γ/Enterobacteriales	*Yersinia pestis*	Vp	4653728	4148	766100	NC_003143
γ/Legionellales	*Coxiella burnetii*	Cb	1995275	2182	210457	AE016828
γ/Pasteurellales	*Haemophilus ducreyi*	Hd	1698955	1902	165788	AE017143
γ/Pasteurellales	*Haemophilus influenzae*	Hi	1830138	1881	192954	L42023
γ/Pasteurellales	*Pasteurella multocida*	Pm	2257487	2014	253223	AE004439
γ/Pasteurellales	*Acinetobacter* sp. *ADP1*	As	3598621	3525	406521	CR543861
γ/Pasteurellales	*Pseudomonas aeruginosa*	Pa	6264403	5646	678474	NC_002516
γ/Pasteurellales	*Pseudomonas putida*	Pp	6181863	5462	735926	AE015451
γ/Pasteurellales	Pseudomonas *syringae*	Ps	6397126	5777	813356	AE016853
γ/Vibrionales	*Vibrio cholerae*	Vc	Ch1[c]: 2961149 Ch2: 1072315	Ch1: 3005 Ch2:1123	Ch1: 388302 Ch2: 144611	Ch1: AE003852 Ch2: AE003853
γ/Vibrionales	*Vibrio parahaemolyticus*	Vp	Ch1: 3288558 Ch2: 1877212	Ch1: 3366 Ch2: 1786	Ch1: 391565 Ch2: 244272	Ch1: BA000031 Ch2: A000032
γ/Vibrionales	*Vibrio vulnificus*	Vv	Ch1: 3281945 Ch2: 1844853	Ch1: 3168 Ch2: 1591	Ch1: 511734 Ch2: 258430	Ch1: AE016795 Ch2: AE016796
γ/Xanthomonadales	*Xanthomonas axonopodis*	Xa	5076188	4303	780336	AE008922
γ/Xanthomonadales	*Xanthomonas citri*	Xc	5175554	4436	730162	AE008923
γ/Xanthomonadales	*Xylella fastidiosa*	Xf	2679306	2894	416317	AE003849
α Proteobacteria/ Caulobacterales	*Caulobacter crescentus*	Cc	4016947	3875	364164	AE005673
α/Rhizobiales	*Agrobacterium tumefaciens*	At	2841490	2883	339888	AE008688
α/Rhizobiales	*Brucella melitensis*	Bm	Ch1: 2117144 Ch2: 1177787	Ch1: 2107 Ch2: 1157	Ch1: 312746 Ch2: 149777	Ch1: AE008917 Ch2: AE008918
α/Rhizobiales	*Mesorhizobium loti*	Ml	7036074	6746	966637	NC_002678
Deinococcus-thermus /Deinococcales	*Deinococcus radiodurans*	Dr	Ch1: 2648638 Ch2: 412348	Ch1: 2744 Ch2: 370	Ch1: 266135 Ch2: 30800	Ch1: NC_001263 Ch2: NC_001264

[a]Codes for each genome used in Figure 4 are shown.

[b]Genes include protein and RNA genes described in the Genbank file.

[c]Chromosomes 1 and 2 are shown separately as Ch1 and Ch2.

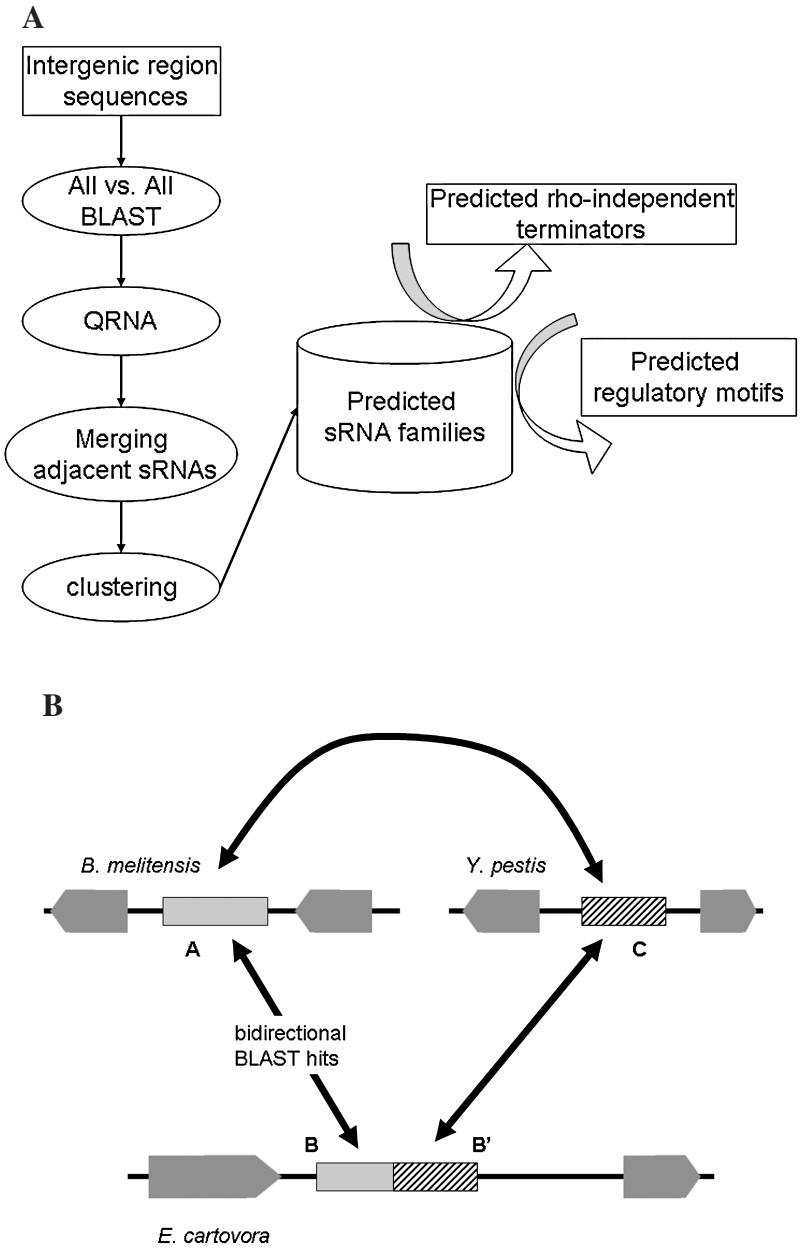

FIG. 1. The procedure of identifying sRNA families. (**A**) The entire procedure. sRNAs predicted by QRNA are clustered into families. Those that overlap with predicted rho-independent transcriptional terminators or regulatory motifs are removed from the families. (**B**) Clustering of sRNAs. In this example, putative sRNAs—A in *B. melitensis*, B(B′) in *E. cartovora*, C in *Y. pestis*—are clustered into a family because all the pairs of A and B, B and C, C and A are initially found pairwise in sRNA alignments. B and B′ are concatenated in the previous step because they are closer than 25 nt.

(Fig. 2). We tried different threshold values, from 5 to 30 with an interval of 5 nt, but the results did not differ greatly.

Clustering of sRNAs

Up to this point, we have pairwise alignments that are classified as sRNA regions. Two sequences in a pairwise alignment come from different organisms (e.g., one from *Y. pestis,* another from *B. melitensis*).

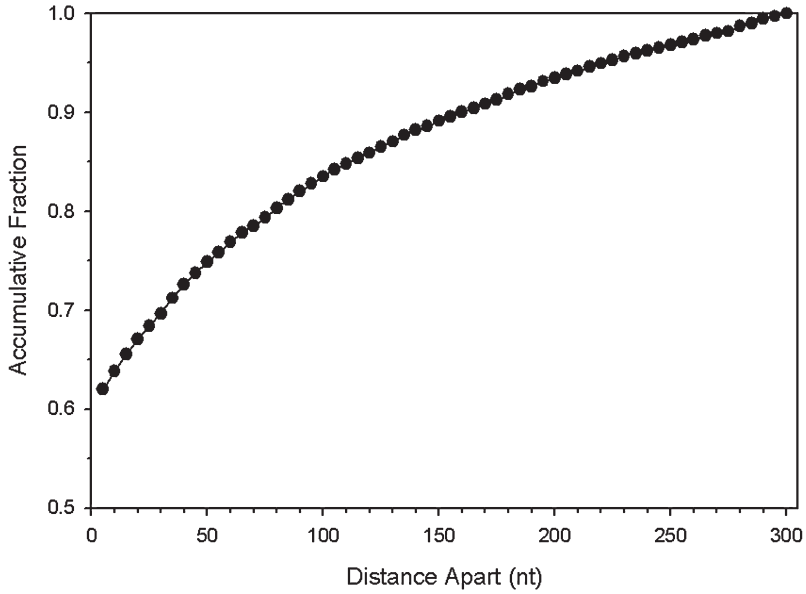

FIG. 2. The accumulative fraction of the distances between initially found sRNA regions by QRNA in the same intergenic region in the 30 genome sequences analyzed.

The next step is to cluster the pairwise alignments of sRNAs to form an "sRNA family" in the following way (Fig. 1B). Suppose there are two pairwise alignments of sRNAs, A with B and B′ with C. Here B and B′ are originally distinct, putative sRNAs located in the same intergenic region of the same organism and closer than 25 nt to each other (termed *the bridging sequences*), thus concatenated in the earlier step. These two alignments of sRNAs are clustered into a family if the pairwise alignment between A and C exists. Thus any pair of members in an sRNA family is guaranteed to have been initially identified as a pairwise alignment of sRNAs (complete-linkage clustering). This procedure is identical to detecting cliques from a graph of sRNAs, where individual sRNAs are represented as nodes and pairs of initially aligned sRNAs are connected by edges. In the same way, the complete-linkage clustering procedure is further carried out to enlarge the size of families. For a merged sRNA family, the highest log-odds score given by QRNA to a component pairwise alignment was assigned. In this study, an sRNA was allowed to participate in multiple families if each of the families satisfies to be a complete cluster.

Eliminating rho-independent transcriptional terminators and regulatory motifs

Next, families overlapping with predicted rho-independent transcriptional terminators (Lesnik et al., 2001) by more than seven nucleotides were removed. In this step, 1.8% of the total sRNAs were removed, resulting in the removal of 1.8% of the families.

At last, families that contained sequences that overlapped with *E. coli* regulatory sequences by more than 50% of the length of either of the sequences were eliminated, since these could have been mistaken as sRNAs due to their similar palindrome pattern. *E. coli* regulatory sequences were taken from RegulonDB (Salgado et al., 2004). In this last step, 2.2% of the total sRNAs were removed, resulting in the removal of 8.0% of the families.

Identifying orthologous protein genes adjacent to sRNAs

We used the precomputed orthologous gene table provided by KEGG (KEGG Orthology Database, KO, http://www.genome.jp) (Kanehisa et al., 2006) to define orthologous protein genes. Small sRNA families with only two members were discarded in the analysis. KO assigns unique KO identities (ID) to orthologous gene groups across species. In the process of corresponding genes to the KO table, KO ID was not

assigned to some portion of the genes because of the inconsistency of gene IDs between Genbank we have used and the KO, orphan genes, which do not have orthologous genes (Siew and Fischer, 2003), and RNA genes, which are not included in KO. Among the 7595 sRNA families shared with three or more organisms (Fig. 5), 6791 (92.0%) of them have two or more sRNA members with at least one of pairs of flanking protein genes with a KO ID.

RESULTS

Verification of the procedure

In a computational work, it is crucial to carefully benchmark the accuracy of the methods used. To begin, we tested the procedure using known sRNAs in the *E. coli* genome compiled from three sources: a list from the National Institute of Child Health and Human Development (NICHD) at NIH (<http://dir2.nichd.nih.gov/nichd/cbmb/segr/e.coli.html>); *E. coli* K-12 sequence annotations at University of Wisconsin, Madison (<www.genome.wisc.edu/sequencing/k12.htm#rna>); and supplemental data from Hershberg et al. (2003). The three datasets were merged into a list containing 77 nonredundant experimentally verified sRNAs. Allowing up to 50 nucleotide shifts, 54 sRNAs (70.1%) were successfully identified. By allowing nucleotide shift of 100 and 10, the accuracy was 74.0% and 62.3%, respectively.

In addition, we tested the procedure on the Rfam database (Griffiths-Jones et al., 2005). Out of 71 *E. coli* sRNAs found in the database, only 21 sequences are included in the intergenic regions we used in our analysis. The rest of the RNA sequences include *E. coli* plasmid loci, insertion sequences, and cloning/reporter vector sequences (all of which are not included in the full *E. coli* genome sequence used) and 29 tRNAs, two 4.5S RNAs, and an antisense RNA, which are annotated as coding regions in Genbank, thus eliminating them from our intergenic regions. Among the 21 sequences that we are supposed to find, 15 were found by our procedure. Thus, the accuracy (sensitivity) is calculated to be 71.4%.

Number of sRNAs in each genome

Within 94,464 intergenic region sequences from the 30 microbial genomes, 26,047 local segments were classified as sRNAs. Figure 3 shows the distribution of the log-odds scores from QRNA and the length of

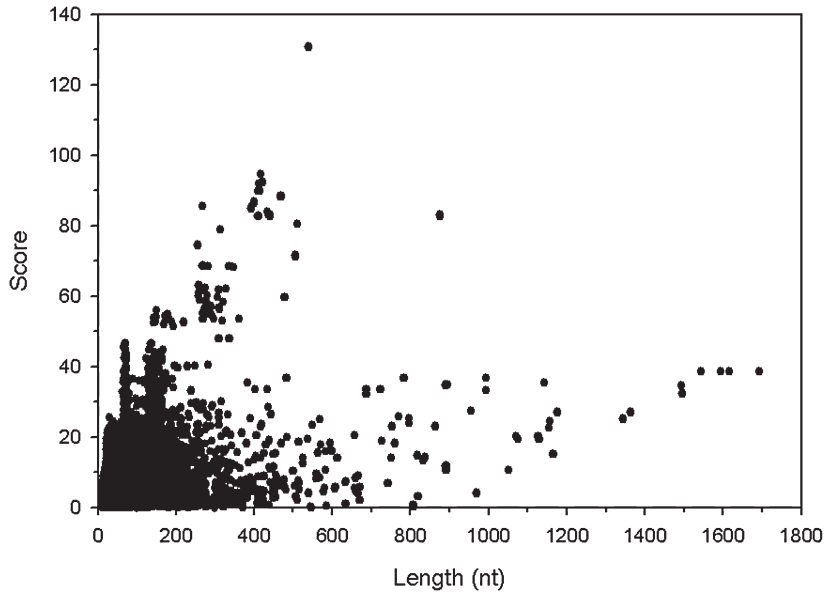

FIG. 3. Distribution of the score relative to the length of the sRNAs.

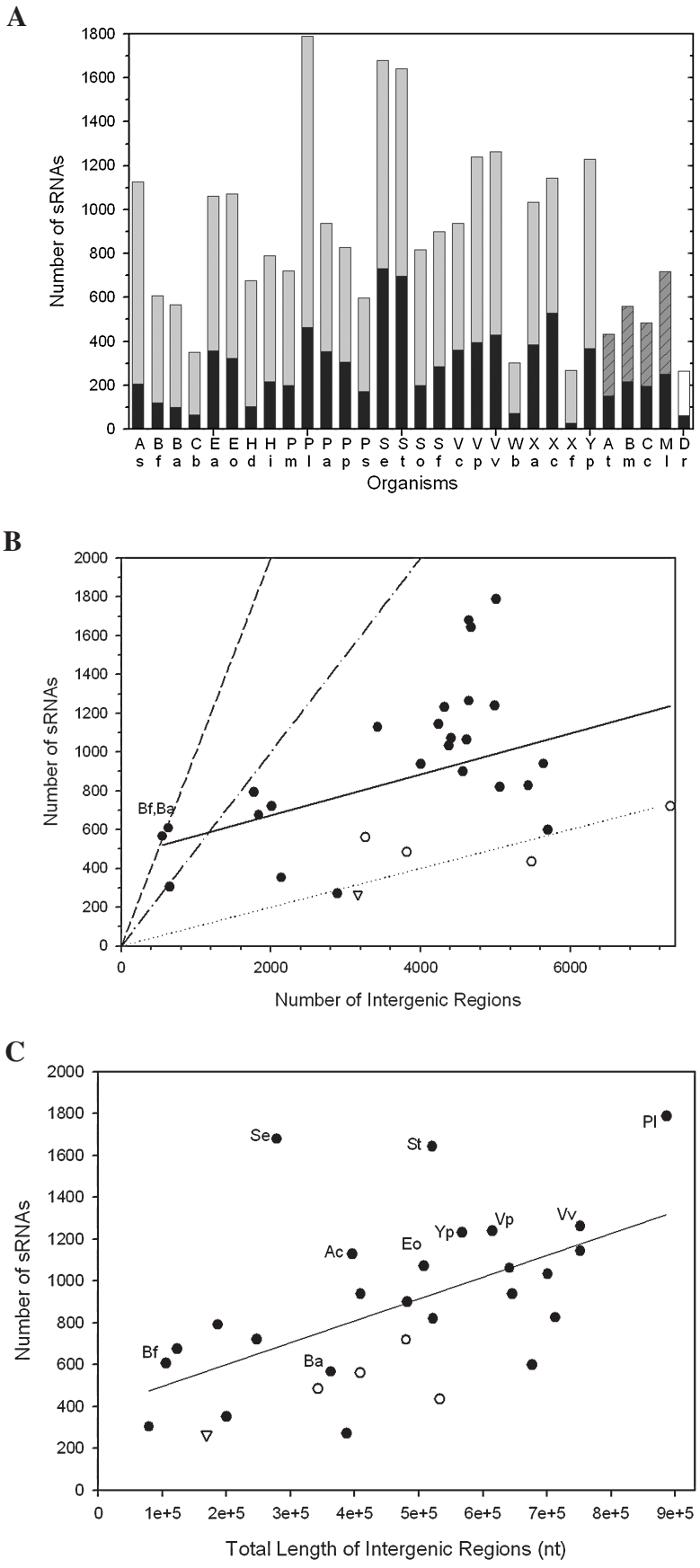

FIG. 4. The number of sRNAs found in each organism. (**A**) The absolute numbers of found sRNAs. Gray bars, γ-proteobacteria; dark gray, striped bars, α-proteobacteria; white bar, deinococci. As, *A. sp. ADP1;* Bf, *B. floridanus;* Ba, *B. aphidicola;* Cb, *C. burnetii;* Ea, *E. carotovora;* Eo, *E. coli* K-12; Hd, *H. ducreyi;* Hi, *H. influenzae;* Pm, *P. multocida;* Pl, *P. luminescens;* Pa, *P. aeruginosa;* Pp, *P. putida;* Ps, *P. syringae;* Se, *S. enterica;* St, *S. typhimurium;* So, *S. oneidensis;* Sf, *S. flexneri;* Vc, *V. cholerae;* Vp, *V. parahaemolyticus;* Vv, *V. vulnificus;* Wb, *W. brevipalpis;* Xa, *X. axonopodis;* Xc, *X. campestris;* Xf, *X. fastidiosa;* Yp, *Y. pestis;* At, *A. tumefaciens;* Bm, *B. melitensis;* Cc, *C. crescentus;* Ml, *M. loti,* Dr, *D. radiodurans*. The black bars show the number of sRNAs with a score of 5 or higher. (**B**) The number of sRNAs found relative to the number of intergenic regions of each genome. Solid circles, γ-proteobacteria; open circles, α-proteobacteria; gray triangle, deinococci. Solid line is a regression line, $y = 463.3 + 0.11x$; dashed line, $y = x$; dash-dot line, $y = 0.3x$; dotted line, $y = 0.1x$. (**C**) Number of sRNAs relative to the total length of the intergenic regions analyzed. The notation of the symbols is the same as the plot B. The solid line is the regression line, $y = 401.0 + 1.06*10^{-3}x$; R^2 coefficient, 0.30.

the detected sRNAs. The number of sRNAs with a score of five or more share 32.0% of the total number of sRNAs. Five is the score used by the authors of QRNA in their paper (McCutcheon and Eddy, 2003). RNAs that are 50 nt or longer comprise 27.1% of the result set.

Figure 4 shows the number of sRNAs found in each organism. The number ranges from 265 *(D. radiodurans)* to 1788 (*P. luminescens)* (Fig. 4A). As for *E. coli*, 1078 sRNAs are detected. The ratio of sRNAs to the number of intergenic regions ranges from 0.08 to 0.47, except for two genomes, *B. floridanus* and *B. aphidicola,* whose ratio is 1.01 (both genomes). The average is 0.25/0.30, with the two genomes excluded/included, respectively (Fig. 4B). Therefore, roughly speaking, 30% of the intergenic regions of the microbial genomes contain sRNAs. In Figure 4C, the number of sRNAs relative to the total length of intergenic regions is plotted. Because QRNA works off of pairwise alignments from BLAST output, it tends to find more sRNAs among evolutionarily close organisms than it does among more distant organisms. Since *E. coli, Y. pestis,* and the Salmonella and Vibrio groups are relatively close to each other, they are all observed to have a large number of sRNAs (Fig 4A and C). By the same reasoning, but with the opposite effect, a relatively small number of sRNAs is found in α-proteobacteria and *D. radiodurans* (Fig. 4). In Figure 4B, the two genomes, *B. floridanus* and *B. aphidicola* have a higher ratio of the sRNAs relative to the number of intergenic regions, but the ratio relative to the total length of intergenic regions is similar to other genomes (Fig. 4C). This is because annotation of these two genomes gives a smaller number, but a larger size of intergenic regions. The regression line in Figure 4C is $y = 392.6 + 1.04*10^{-3}x$ with an R^2 coefficient value of 0.30. It will be interesting to see if the number of predicted sRNAs in eukaryotes fit these regression lines from microbial genomes.

A total of 67.4% (79.9%) of relatively longer intergenic regions with $\geq$500 nt have sRNAs residing within them. For shorter intergenic regions of $<$500 nt, sRNAs are found among 24.5% (34.1%). (The data for *E. coli* are given in the parentheses.)

Conservation of sRNAs among organisms

Initially found sRNAs were clustered into 38,050 families (Fig. 5). Note here that the number of families is larger than the number of unique sRNAs reported in Figure 4, because an sRNA is allowed to be a member of multiple families. Within each family, it is guaranteed that every pair of sRNAs is fully connected, that is, they initially formed a pairwise alignment that is recognized as RNA by the QRNA program. An alternative way of clustering is to carry out single-linkage clustering, where a member in a family is connected to at least another member in the family. We have not used the single-linkage clustering because it formed a fewer number of too-large sRNA families. We allowed an sRNA to belong to multiple families rather than assigning it to a single family by considering the highest score (the E value from the initial BLAST search or the log-odds score of QRNA) because, unlike the sequence similarity for classifying a protein family, the significance of these scores is not well established in the case of sRNAs. The distribution of the number of organisms sharing the same sRNA family is shown in Figure 5. Note that there is no family with a single sRNA, because the detecting procedure starts from pairwise alignments. There are 572 families formed that comprise sRNAs shared by six or more organisms. The number of unique sRNAs participating in families of a certain size is shown in Figure 6. 6681 sRNAs (25.4% among the total) are included in families of at least three. In *E. coli*, 1589 sRNA families shared by at least three organisms were identified (Fig. 5), which include 380 sRNA members (Fig. 6).

Table 2 shows the most common combinations of organisms sharing the same sRNA family. For the frequent combinations of three organisms, permutations of the closely related organisms, *E. coli, S. enterica, S. flexneri,* and *S. typhimurium LT2*, are evident (Table 2A). *B. melitensis* from α-proteobacteria frequently forms combinations with *M. loti*, which is another organism from α-proteobacteria, as part of combinations with Pseudomonas and others from γ-proteobacteria (Tab. 2B). Pairs of organisms sharing sRNA families are illustrated in Figure 7. Because the current procedure starts from pairwise alignments with a certain significance score, more sRNAs are found between closely related organisms. However, sRNAs are still found between distantly related organisms. For example, *D. radiodurans*, which is most evolutionary distant organism among the ones used in this study, finds common sRNAs with γ-proteobacteria.

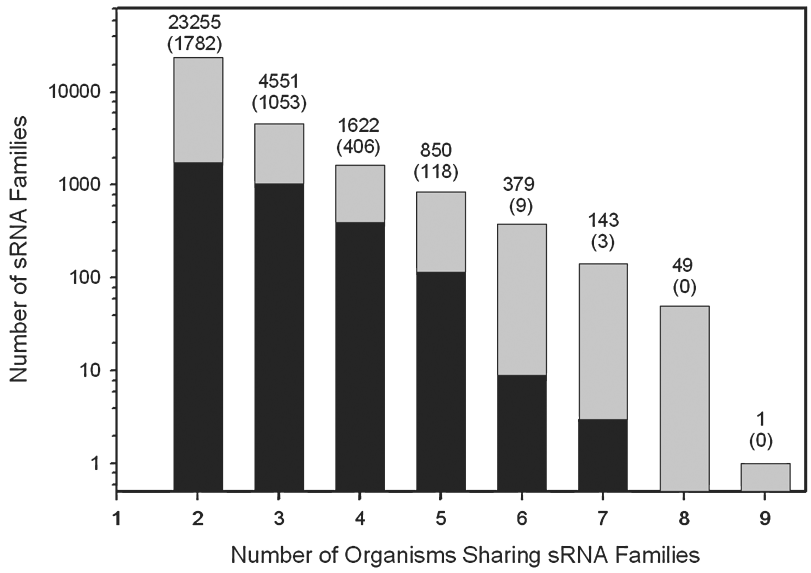

FIG. 5. Conservation of sRNAs. The gray bars show the number of sRNAs shared by some of the 30 organisms, and the black bars and the numbers in the parentheses show sRNA families shared by *E. coli* and the others. The y-axis is plotted in the log scale.

Orthology of adjacent protein genes to sRNA families

Next, we investigated if flanking genes to sRNA members of a family are orthologous or not. When flanking genes to sRNAs of the same family are orthologous to each other, it indicates that most probably the entire region, including the flanking genes and the intergenic region with the sRNAs in it, are conserved among the organisms (orthologous intergenic regions). Among the sRNA families shared with three organisms or more, 37.7% of them have at least a pair of sRNAs with at least one of the pair of flanking protein genes orthologous. Furthermore, in 15.7% of the families, sRNA members share orthologous protein

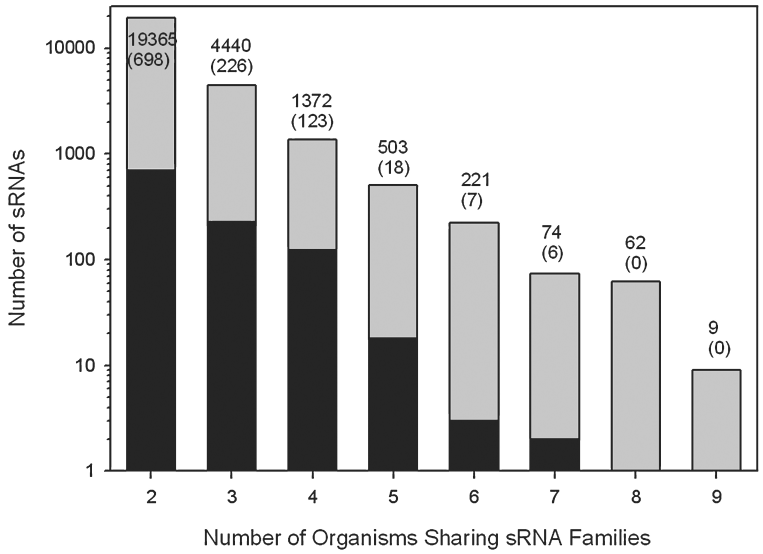

FIG. 6. The number of unique sRNAs in families of each size. The black bars show unique sRNAs in families in *E. coli* shared by the other organisms.

Table 2A. Five Most Frequent Combinations of Three Organisms Sharing sRNA Family

Organisms	*Frequency*
E. coli, S. enterica, S. typhimurium LT2	696
S. enterica, S. flexneri, S. typhimurium LT2	676
E. coli, S. typhimurium LT2, S. flexneri	580
E. coli, S. flexneri, S. enterica	511
E. carotovora, S. enterica, Y. pestis	492

Table 2B. Five Most Frequent Combinations of Five Organisms Sharing sRNA Family

Organisms	*Frequency*
E. carotovora, E. coli, S. enterica, S. flexneri, S. typhimurium LT2	87
B. melitensis, E. carotovora, M. loti, P. aeruginosa, P. multocida	81
B. melitensis, E. carotovora, M. loti, P. aeruginosa, Y. pestis	77
B. melitensis, E. carotovora, M. loti, P. multocida, Y. pestis	77
B. melitensis, E. carotovora, S. enterica, S. typhimurium LT2, Y. pestis	73

genes as at least one of the pair of flanking genes. In many cases, the organism pairs that share a sRNA family with flanking orthologous protein genes are closely related. Typical cases include pairs among *S.enterica*, *S. typhimurium LT2*, *S. flexneri*, *E. coli*, and *Y. pestis*; *P. ptuida* and *P. aeruginosa*; or *V. parahaemolyticus, V. vulnificus*, and *V. cholerae*.

On the other hand, it is also interesting that approximately the remaining 60% of the sRNA families are found in "nonorthologous" intergenic regions. This may have two causes, which are not necessarily mutually exclusive. Firstly, orthologous/homologous sRNAs shared among the organisms, which are most probably responsible for the same function, locate at different genome positions. It would not be surprising if orthologous sRNAs are located at different positions in different genomes, because the location of protein-coding genes are known to be shuffled when organisms start to diverge (Suyama and Bork, 2001; Watanabe et al., 1997). Our procedure detects clusters of sRNAs that have a reasonably good sequence similarity and thus can have the same secondary structure if they are actually transcribed as RNA molecules. The second possibility is that sRNAs clustered into a family share the same secondary structure but not necessarily the same function. This is not unreasonable because function of sRNAs may be determined by several key nucleotides (most likely in the loop regions) but not necessarily by the whole sequences. All the results in this study, including detected sRNAs and families in *E. coli*, are available online (<http://dragon.bio.purdue.edu/sRNA>).

Comparison with the other results

To verify our prediction more rigorously, we compared our results further with previously published results of computationally predicted sRNAs in the *E. coli* genome reported by Chen *et al*. (2002). Their approach was to find pairs of sigma-70 promoter motif and a terminator motif in intergenic regions. An sRNA is predicted if a motif pair is found on a same strand and if the pair is greater than 45 but less than 350 nt apart. Open reading frames and possible ribosome binding sites were searched for downstream of each promoter, and, if found, candidate regions were removed because they were more likely to be protein-coding regions. As a result, Chen and co-workers predicted 227 sRNAs between 80 and 400 nt in length in the *E. coli* genome, 32 of which were already known sRNAs. They also performed Northern hybridization to verify transcription for some of the candidate sRNAs.

We have compared our results with theirs, as follows: For a given predicted sRNA in their paper, if both left and right boundaries of an sRNA of our prediction are found within 100 nt of those of an sRNA from

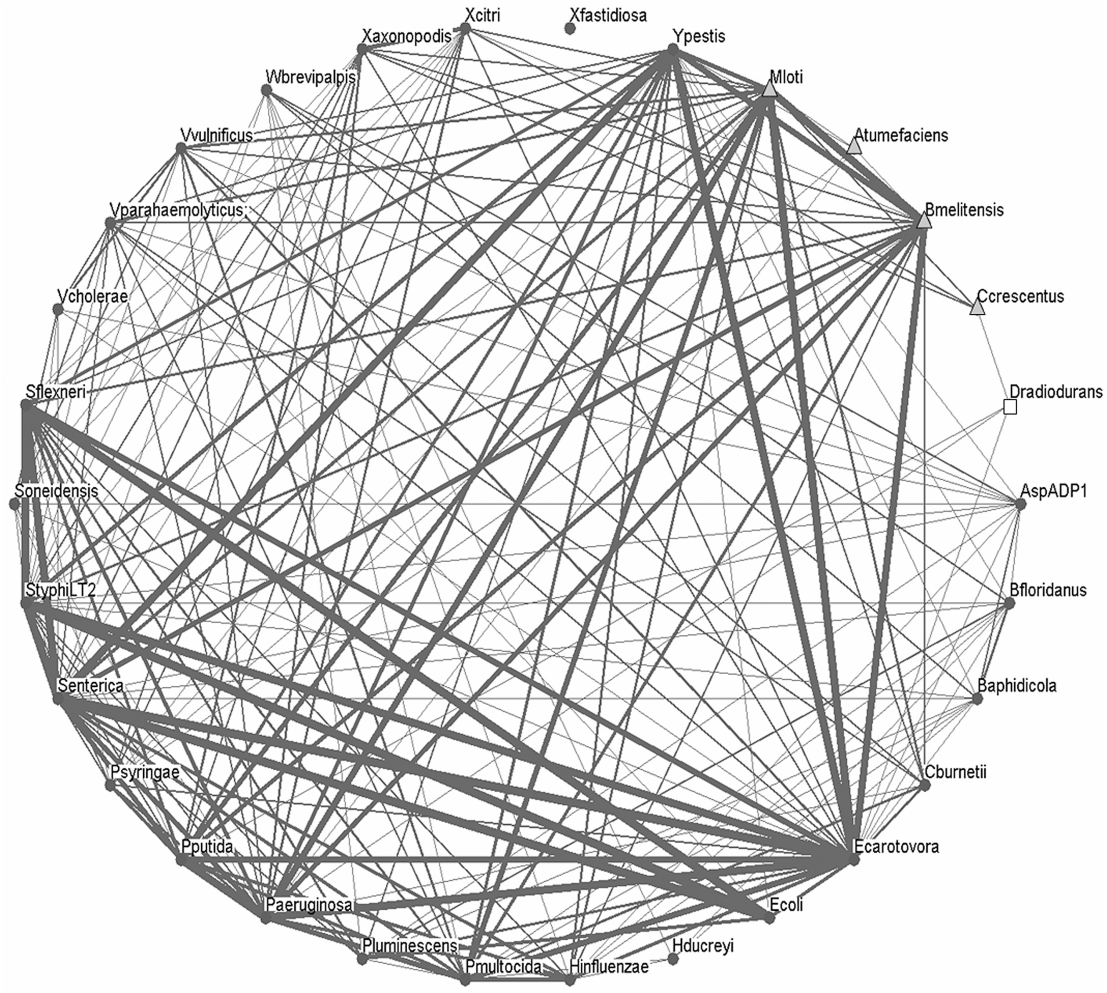

FIG. 7. Organism pairs sharing sRNAs. The number of sRNAs shared by organism pairs is represented by edges with a thickness of ten different levels. sRNA families shared by three or more organisms are considered. Pairs with less than 50 occurrences are discarded. Circular nodes, γ-proteobacteria; gray triangles, α-proteobacteria; white square, deinococcus. The Ospray package (Breitkreutz et al., 2003) was used to generate this diagram.

their prediction, the two sRNAs are considered to correspond to each other. Among the 227 sRNAs predicted by them, 114 (50.2%) overlapped with our predicted sRNAs. When the threshold value is changed to 50 nt and 1000 nt, 101 (44.4 %) and 157 (69.2 %) of them corresponded to our prediction, respectively. The overlap between our prediction and Chen and his colleagues' prediction is considerably high compared with the very few overlaps found between the other existing predictions. (It has been reported that out of 1001 nonredundant sRNA candidates identified in five works, only 9.5% of them are predicted in multiple works [Hershberg et al., 2003]).

DISCUSSION

We have identified possible sRNAs in intergenic regions of 25 γ-proteobacterium, including *E. coli* and its relatively evolutionary close organisms, 4 α-proteobacterium, and *D. radiodurans*, by a computational method. Here we employed a comparative genomics approach on a large number of organisms. Comparative genomics approaches have been proven to be very powerful in gene finding (Dewey et al., 2004; Taher

et al., 2003) or protein function prediction (Snel et al., 2000). To the best of our knowledge, this is the first attempt of a large-scale comparative genomics study of sRNAs.

The primary contribution of this paper is that we propose a new procedure to cluster sRNAs into families and by applying the procedure we found 25.4% of the identified sRNAs are shared by more than three organisms. An interesting finding is that flanking protein-coding genes to sRNAs in the same family are not always orthologous, which implies that sRNAs are shuffled in a genome during evolution in the same way that protein-coding genes do. The degree of the conservation of sRNAs among organisms adds another measure for assessing the plausibility of computational predictions. In *E. coli*, among 1078 predicted sRNAs, 380 are shared by three organisms or more.

Although our prediction has not been verified by experiment, we emphasize that our computational work has been done in a very careful manner. First, only pairwise alignments of intergenic regions, which satisfy the sequence similarity threshold from both directions, are used in order to reject alignments of spurious similarity. Second, regions that overlap with known regulatory motifs are removed from sRNA hits, because a quasi-palindrome pattern of regulatory motifs can be misrecognized as a stem region of RNA secondary structure. Third, regions that overlap with predicted rho-independent transcriptional terminator regions were removed. Fourth, we have also merged initially predicted sRNAs in the same intergenic region if they locate close to each other, because QRNA tends to predict very short RNA regions, and as a result the length of initially identified sRNAs is largely biased to a shorter side. Fifth, we have benchmarked our procedure by applying it to experimentally verified sRNAs in order to estimate the accuracy of the prediction made in the current study. Finally, we compared our prediction in the *E. coli* genome with previous work reported by Chen *et al.* (2002) to gauge the consistency. Their work employs a completely different method, thus yielding appropriate data to cross-validate our results. To the best of our knowledge, this report represents the most carefully executed computational work for identifying sRNAs, although some of the previous works employed wet experiments, such as Northern blot.

At this juncture, it is appropriate to discuss limitation of our computational method. Because the prediction procedure starts from alignments constructed by BLAST, the number of predicted sRNAs in a genome depends on how many closely related genomes are available for the target genome, as illustrated in Table 2A and Figure 7, which show the most frequent combinations of three organisms are permutations of *E. coli* and its five closely related organisms. We added *D. radiodurans*, a relatively distant organism in the list of genomes examined, to investigate how many sRNAs can still be detected in it. Apparently, the ratio of the number of detected sRNAs to the intergenic region size in *D. radiodurans* is lower than that in the other organisms (Fig. 4B and C). However, we still identified 265 sRNAs in *D. radiodurans* from pairwise alignments between organisms in γ-proteobacteria. The key in optimizing the sRNA finding procedure is to prepare many genome sequences that are closely related to allow for pairwise alignments, yet evolutionarily distinct to allow for compensatory mutations consistent with base-paired secondary structure of sRNAs.

We verified the sensitivity of our prediction by running the procedure on datasets of known sRNAs, which was above 70%. It is difficult to estimate the specificity (*i.e.*, the ratio of correct predictions among all the predictions) of our procedure, because our current knowledge of sRNAs in genomes is limited and there is not a reliable way to judge if a newly predicted sRNA is real or a false positive. Although we did careful work on removing predicted rho-independent transcriptional terminators from our sRNA prediction, the current procedure may still not perfectly eliminate the possibility of contamination of cis-regulatory RNA structures (Rivas and Eddy, 2000). Although it is not avoidable that some amount of false positives is included in the present prediction, modestly speaking, our study has revealed for the first time that there are hundreds of regions that can possibly form secondary structures if transcribed as RNAs in the genome sequences; more than 60% of the sRNA families may be shuffled their location in genomes during evolution.

In this study, detected sRNAs are clustered into families by a complete-linkage clustering. The number of families changes if a different clustering method is used. We also tried to cluster sRNAs solely by their secondary structure predicted by MFOLD (Mathews et al., 1999), because the secondary structure is essential for the function of some sRNAs (*e.g.*, the kissing complex formation of OxyS [Argaman and Altuvia, 2000]). However, it has become apparent that this was not trivial because the best-

scoring secondary structures can vary a great amount, even for a pair of highly homologous RNA sequences. Certainly a set of suboptimal secondary structures should be considered when the structure similarity of two sRNAs is examined. Structure analysis of sRNAs along this line is left for the future work. Another interesting direction is to predict target regions for identified sRNAs (Lewis et al., 2003; Rehmsmeier et al., 2004).

Taking the abundance and the prevalence of the sRNAs among organisms into account, sRNAs are surely playing a variety of important roles in gene regulation. sRNAs may lead, if properly identified and catalogued, to opening the door to the systematic understanding of the overall picture of the regulatory network of organisms (Madan Babu and Teichmann, 2003).

CONCLUSION

A comparative genomics study of sRNAs in 30 microbial genomes was carried out to investigate conservation of sRNAs among genomes and to assign an additional reliability measure to the predictions. Possible sRNAs are found in approximately 30% of intergenic regions of the genomes, and approximately 25% of them are conserved among three organisms or more. In the *E. coli* K-12 genome, 1078 sRNAs are found, among which 380 are conserved in three organisms or more. The number of 380 sRNAs roughly agrees with previous studies by other groups. More than 60% of sRNAs are shuffled their location in genomes during evolution.

ACKNOWLEDGMENTS

The authors thank Bin Li for technical help on the clustering sRNA families. We thank Troy Hawkins for proofreading the manuscript and Carol Greski for preparing the manuscript. The authors are also grateful to Barry L. Wanner for valuable discussions. D.K. acknowledges supports from NIH (R01GM075004 and U24 GM077905) and NSF (DMS 0604776). S.L. was supported in part by the Howard Hughes Summer Internship Program in 2004 and 2005.

REFERENCES

ALTSCHUL, S.F., GISH, W., MILLER, W., MYERS, E.W., and LIPMAN, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. **215,** 403–410.

AMBROS, V., BARTEL, B., BARTEL, D.P., BURGE, C.B., CARRINGTON, J.C., CHEN, X., et al. (2003). A uniform system for microRNA annotation. RNA **9,** 277–279.

ANDERSEN, J., and DELIHAS, N. (1990). micF RNA binds to the 5′ end of ompF mRNA and to a protein from *Escherichia coli*. Biochemistry **29,** 9249–9256.

ARGAMAN, L., and ALTUVIA, S. (2000). fhlA repression by OxyS RNA: kissing complex formation at two sites results in a stable antisense-target RNA complex. J. Mol. Biol. **300,** 1101–1112.

ARGAMAN, L., HERSHBERG, R., VOGEL, J., BEJERANO, G., WAGNER, E.G., MARGALIT, H., et al. (2001). Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. Curr. Biol. **11,** 941–950.

BACHELLERIE, J.P., CAVAILLE, J., and HUTTENHOFER, A. (2002). The expanding snoRNA world. Biochimie **84,** 775–790.

BENSON, D.A., KARSCH-MIZRACHI, I., LIPMAN, D.J., OSTELL, J., and WHEELER, D.L. (2004). GenBank: update. Nucl. Acids Res. **32,** D23–26.

BREITKREUTZ, B.J., STARK, C., and TYERS, M. (2003). Osprey: a network visualization system. Genome Biol. **4,** R22.

BRENNECKE, J., HIPFNER, D.R., STARK, A., RUSSELL, R.B., and COHEN, S.M. (2003). Bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila. Cell **113,** 25–36.

CARTER, R.J., DUBCHAK, I., and HOLBROOK, S.R. (2001). A computational approach to identify genes for functional RNAs in genomic sequences. Nucleic Acids Res. **29,** 3928–3938.

CHEN, S., LESNIK, E.A., HALL, T.A., SAMPATH, R., GRIFFEY, R.H., ECKER, D.J., et al. (2002). A bioinformatics based approach to discover small RNA genes in the *Escherichia coli* genome. Biosystems **65,** 157–177.

DEWEY, C., WU, J.Q., CAWLEY, S., ALEXANDERSSON, M., GIBBS, R., and PACHTER, L. (2004) Accurate identification of novel human genes through simultaneous gene prediction in human, mouse, and rat. Genome Res. **14,** 661–664.

ECKSTEIN, F. (2005). Small non-coding RNAs as magic bullets. Trends Biochem. Sci. **30,** 445–452.

EDITORIAL. (2004). Desperately seeking RNAs. Nat. Struct. Mol. Biol. **11,** 799.

EDDY, S.R. (2001). Non-coding RNA genes and the modern RNA world. Nat. Rev. Genet. **2,** 919–929.

GOTTESMAN, S. (2004a). Small RNAs shed some light. Cell **118,** 1–2.

GOTTESMAN, S. (2004b). The small RNA regulators of *Escherichia coli*: roles and mechanisms. Ann. Rev. Microbiol. **58,** 303–328.

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,** D121–124.

HAMMOND, S.M. (2006). MicroRNAs as oncogenes. Curr. Opin. Genet. Dev. **16,** 4–9.

HE, L., and HANNON, G.J. (2004). MicroRNAs: small RNAs with a big role in gene regulation. Nat. Rev. Genet. **5,** 522–531.

HERSHBERG, R., ALTUVIA, S., and MARGALIT, H. (2003). A survey of small RNA-encoding genes in *Escherichia coli*. Nucl. Acids Res. **31,** 1813–1820.

HUTTENHOFER, A., KIEFMANN, M., MEIER-EWERT, S., O'BRIEN, J., LEHRACH, H., BACHELLERIE, J.P., et al. (2001). RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. EMBO J. **20,** 2943–2953.

KANEHISA, M., GOTO, S., HATTORI, M., OKI-KINOSHITA, K.F., ITOH, M., KAWASHIMA, S., et al. (2006). From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. **34,** D354–357.

KIM, V.N. (2005). MicroRNA biogenesis: coordinated cropping and dicing. Nat. Rev. Mol. Cell Biol. **6,** 376–385.

LESNIK, E.A., SAMPATH, R., LEVENE, H.B., HENDERSON, T.J., MCNEIL, J.A., and ECKER, D.J. (2001). Prediction of rho-independent transcriptional terminators in *Escherichia coli*. Nucleic Acids Res. **29,** 3583–3594.

LEWIS, B.P., SHIH, I.H., JONES-RHOADES, M.W., BARTEL, D.P., and BURGE, C.B. (2003). Prediction of mammalian microRNA targets. Cell **115,** 787–798.

MADAN BABU, M., and TEICHMANN, S.A. (2003). Evolution of transcription factors and the gene regulatory network in *Escherichia coli*. Nucleic Acids Res. **31,** 1234–1244.

MASSE, E., and GOTTESMAN, S. (2002). A small RNA regulates the expression of genes involved in iron metabolism in *Escherichia coli*. Proc. Natl. Acad. Sci. USA **99,** 4620–4625.

MATHEWS, D.H., SABINA, J., ZUKER, M., and TURNER, D.H. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J. Mol. Biol. **288,** 911–940.

MCCUTCHEON, J.P., and EDDY, S.R. (2003). Computational identification of non-coding RNAs in Saccharomyces cerevisiae by comparative genomics. Nucleic Acids Res. **31,** 4119–4128.

MIZUNO, T., CHOU, M.Y., and INOUYE, M. (1984). A unique mechanism regulating gene expression: translational inhibition by a complementary RNA transcript (micRNA). Proc. Natl. Acad. Sci. USA **81,** 1966–1970.

MOURELATOS, Z., DOSTIE, J., PAUSHKIN, S., SHARMA, A., CHARROUX, B., ABEL, L., et al. (2002). miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs. Genes Dev. **16,** 720–728.

PELUSO, P., HERSCHLAG, D., NOCK, S., FREYMANN, D.M., JOHNSON, A.E., and WALTER, P. (2000). Role of 4.5S RNA in assembly of the bacterial signal recognition particle with its receptor. Science **288,** 1640–1643.

REHMSMEIER, M., STEFFEN, P., HOCHSMANN, M., and GIEGERICH, R. (2004). Fast and effective prediction of microRNA/target duplexes. RNA **10,** 1507–1517.

RIVAS, E., and EDDY, S.R. (2001). Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics. **2,** 8.

RIVAS, E., and EDDY, S.R. (2000). Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. Bioinformatics **16,** 583–605.

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.

SALGADO, H., GAMA-CASTRO, S., MARTINEZ-ANTONIO, A., DIAZ-PEREDO, E., SANCHEZ-SOLANO, F., PERALTA-GIL, M., et al. (2004). RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. Nucleic Acids Res **32** (database issue), 303–306.

SIEW, N., and FISCHER, D. (2003). Analysis of singleton ORFans in fully sequenced microbial genomes. Proteins **53,** 241–251.

SLEDJESKI, D.D., GUPTA, A., and GOTTESMAN, S. (1996). The small RNA, DsrA, is essential for the low temperature expression of RpoS during exponential growth in *Escherichia coli*. EMBO J. **15,** 3993–4000.

SNEL, B., LEHMANN, G., BORK, P., and HUYNEN, M.A. (2000). STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Res. **28,** 3442–3444.

STORZ, G., OPDYKE, J.A., and ZHANG, A. (2004). Controlling mRNA stability and translation with small, noncoding RNAs. Curr. Opin. Microbiol. **7,** 140–144.

SUYAMA, M., and BORK, P. (2001). Evolution of prokaryotic gene order: genome rearrangements in closely related species. Trends Genet. **17,** 10–13.

TAHER, L., RINNER, O., GARG, S., SCZYRBA, A., BRUDNO, M., BATZOGLOU, S., et al. (2003). AGenDA: homology-based gene prediction. Bioinformatics **19,** 1575–1577.

TJADEN, B., SAXENA, R.M., STOLYAR, S., HAYNOR, D.R., KOLKER, E., and ROSENOW, C. (2002). Transcriptome analysis of *Escherichia coli* using high-density oligonucleotide probe arrays. Nucl. Acids Res. **30,** 3732–3738.

WAGNER, E.G., and FLARDH, K. (2002). Antisense RNAs everywhere? Trends Genet. **18,** 223–226.

WASSARMAN, K.M. (2004). RNA regulators of transcription. Nat. Struct. Mol. Biol. **11,** 803–804.

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.

WASSARMAN, K.M., and STORZ, G. (2000). 6S RNA regulates E. coli RNA polymerase activity. Cell **101,** 613–623.

WASSARMAN, K.M., ZHANG, A., and STORZ, G. (1999). Small RNAs in *Escherichia coli.* Trends Microbiol. **7,** 37–45.

WATANABE, H., MORI, H., ITOH, T., and GOJOBORI, T. (1997). Genome plasticity as a paradigm of eubacteria evolution. J. Mol. Evol. **44**(suppl 1), 57–64.

WIGHTMAN, B., HA, I., and RUVKUN, G. (1993). Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in *C. elegans.* Cell **75,** 855–862.

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.

ZHANG, Y., ZHANG, Z., LING, L., SHI, B., and CHEN, R. (2004). Conservation analysis of small RNA genes in *Escherichia coli*. Bioinformatics **20,** 599–603.

Address reprint requests to:
Daisuke Kihara
Department of Biological Sciences
Department of Computer Science
Markey Center for Structural Biology
Lilly Hall, B207
915 West State Street
West Lafayette, IN 47907

E-mail: dkihara@purdue.edu