Loss of Conserved Noncoding RNAs in Genomes of Bacterial Endosymbionts

Dorota Matelska[1], Malgorzata Kurkowska[1], Elzbieta Purta[1], Janusz M. Bujnicki[1,2,*] and Stanislaw Dunin-Horkawicz[1,*]

[1]Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology, Warsaw, Poland

[2]Laboratory of Structural Bioinformatics, Institute of Molecular Biology and Biotechnology, Adam Mickiewicz University, Poznan, Poland

*Corresponding author: E-mail: sdh@genesilico.pl; iamb@genesilico.pl.

Accepted: January 11, 2016

Abstract

The genomes of intracellular symbiotic or pathogenic bacteria, such as of *Buchnera*, *Mycoplasma*, and *Rickettsia*, are typically smaller compared with their free-living counterparts. Here we showed that noncoding RNA (ncRNA) families, which are conserved in free-living bacteria, frequently could not be detected by computational methods in the small genomes. Statistical tests demonstrated that their absence is not an artifact of low GC content or small deletions in these small genomes, and thus it was indicative of an independent loss of ncRNAs in different endosymbiotic lineages. By analyzing the synteny (conservation of gene order) between the reduced and nonreduced genomes, we revealed instances of protein-coding genes that were preserved in the reduced genomes but lost *cis*-regulatory elements. We found that the loss of *cis*-regulatory ncRNA sequences, which regulate the expression of cognate protein-coding genes, is characterized by the reduction of secondary structure formation propensity, GC content, and length of the corresponding genomic regions.

Key words: endosymbionts, noncoding RNA loss, covariance models, Rfam.

Introduction

The length of fully sequenced bacterial genomes ranges from 110 kb (Bennett and Moran 2013) to 15 Mb (Han et al. 2013). Phylogenetic analysis indicates that the smallest of them are not the closest relatives of ancient bacteria, as was believed earlier (Wallace and Morowitz 1973), but rather originate from larger ancestral genomes subjected to a massive loss of DNA (McCutcheon and Moran 2012). As most of the smallest genomes are of symbiotic and pathogenic bacteria that live within eukaryotic host cells, it is accepted that their reduction was triggered by the change in the living environment (Woese 1987).

The genome reduction process has been characterized in a variety of bacterial genomes (Ochman and Moran 2001). According to the current knowledge, it can be subdivided into stages (McCutcheon and Moran 2012). In the early stages, observable in secondary symbionts, such as *Serratia symbiotica* and *Sodalis glossinidius*, genes are reduced to pseudogenes by a rapid accumulation of mutations (Cole et al. 2001), large DNA fragments spanning functionally unrelated protein-coding genes are removed (Moran and Mira 2001), and chromosomes undergo rearrangements (Belda et al. 2005). In further stages, deletions in nonfunctional regions lead to extremely reduced but stable genomes, represented by long-term obligatory (primary) intracellular endosymbionts such as *Buchnera aphidicola* and *Carsonella ruddii* (Nakabachi et al. 2006). The genome reduction process has been described in detail for protein-coding genes (McCutcheon and Moran 2012). It has been shown that their loss is not a random process and that a similar pattern of loss is observed in distantly related and independently reduced bacteria (Merhej et al. 2009).

The main driving forces of genome reduction are attributed to a deletional bias (Mira et al. 2001), a weakened selection on gene function (Merhej et al. 2009), a reduced opportunity for acquisition of exogenous DNA (Wernegreen et al. 2000), and a reduction of effective population size (Kuo et al. 2009). The reduced genomes are not only characterized by a loss of protein-coding regions but also by lower GC content (Moran 2002) and changes in intergenic regions (IGRs) (Molina and van Nimwegen 2008). IGRs contain a mixture of

nonfunctional neutral sequences (e.g., pseudogenes that are frequently observed in secondary symbionts [Lamelas et al. 2011]) and functional elements, such as transcription-factor binding sites and regions that encode noncoding RNAs (ncRNAs).

The ncRNAs perform various enzymatic, structural, and regulatory functions, and therefore are an important class of functional elements (Eddy 2001; Storz 2002). The regulatory ncRNAs can act either in *cis* (as a part of the mRNA under control) or in *trans* (as standalone RNA molecules that act on other molecules) (Waters and Storz 2009; Storz et al. 2011). The *cis*-regulatory ncRNAs in bacteria are characterized by conserved structure and are typically located within the 5′-untranslated regions (5′-UTRs) of mRNAs. Typically, the signal recognized by a *cis*-regulatory element is directly related to the function of the downstream protein-coding region; for example, riboswitches frequently regulate genes involved in the use or production of a metabolite they recognize, T-boxes found in aminoacyl-tRNA synthetase genes bind the corresponding tRNAs, and ribosomal leaders recognize ribosomal proteins encoded by the downstream genes (Fu et al. 2013). In response to a stimulus, *cis*-regulatory elements alter their structure, and in most cases this change leads to a transcription or translation repression (Serganov and Patel 2008).

Trans-acting ncRNAs include small RNAs (sRNAs) and antisense RNAs (asRNAs) and are expressed as standalone RNA molecules. sRNAs, similarly to *cis*-regulatory elements, adopt a well-defined conserved structure. They are involved in various processes, such as tRNA maturation (RNase P ribozyme), translation (transfer-messenger RNA [tmRNA]), trafficking (RNA component of signal recognition particle [SRP]), or transcription regulation (6S RNA) (Waters and Storz 2009; Storz et al. 2011). In contrast to structured and broadly conserved sRNAs and *cis*-regulatory ncRNAs, asRNAs are typically poorly conserved and their regulatory function relies on base-pairing interactions with a target mRNA, which triggers its degradation or inhibits translation (Sesto et al. 2013). The lack of conservation and the fact that some of the asRNAs are encoded on the opposite strand of the protein-coding sequences make them hard to predict using computational methods. However, experimental methods, such as RNA-seq, have shown that massive asRNA expression is common in bacterial genomes (e.g., 20% and 27% genes encode for asRNAs in *Escherichia coli* and *Helicobacter pylori*, respectively) (Güell et al. 2011). Also, in contrast to *cis*-regulatory elements and sRNAs, asRNAs evolve rapidly (Gottesman and Storz 2011) and their transcription patterns differ even between closely related species (Wurtzel et al. 2012).

ncRNAs have been investigated in reduced genomes. For example, by scanning completely sequenced genomes with specific sequence descriptors for ncRNA families, it has been shown that some of the extremely reduced genomes lack the highly conserved sRNA genes encoding SRP RNA (Rosenblad et al. 2009), RNase P (Willkomm et al. 2002), or tmRNA (Hudson et al. 2014). However, other studies have shown that genomes of host-restricted bacteria, despite lacking many genes, still use ncRNAs. For example, a substantial number of IGRs in *Buchnera* species contain functional elements including potentially structured ncRNA (Degnan et al. 2011). This observation was further confirmed by a recent comparative analysis of RNA expression in five *Buchnera* strains, which revealed that a considerable portion of coding genes are expressed with UTRs, indicating their regulatory potential (Hansen and Degnan 2014). Moreover, transcriptomics studies revealed that gene regulation in two host-restricted bacteria *Mycoplasma pneumoniae* and *B. aphidicola* relies on posttranscriptional processes in which asRNAs play a major role (Güell et al. 2009; Hansen and Degnan 2014). As much as 13% and 20% of coding genes in *Mycoplasma* and *Buchnera*, respectively, are covered by antisense transcripts. The majority of the asRNAs is assumed to be rapidly evolving and dynamic, being conserved at most at the level of species (Hansen and Degnan 2014).

Due to the weak sequence conservation, especially in comparison to proteins, the task of identification and classification of ncRNA sequences is difficult. Owing to the rise of computational methods, such as LocaRNA (Will et al. 2012), Infernal (Nawrocki and Eddy 2013), and CMfinder (Yao et al. 2006), which capture not only sequence but also structure conservation, and experimental techniques for transcriptomics (Wang et al. 2009), our knowledge about evolutionarily conserved ncRNAs is systematically expanding and being cataloged in the Rfam database (Burge et al. 2013). The availability of such methods and catalogs of homologous ncRNA instances in different organisms opened up the possibility to conduct large-scale comparative analyses (Hoeppner et al. 2012).

In our study, we applied a comparative genomics approach to study changes in the ncRNA repertoire triggered by the genome reduction process. Our aim was to investigate to which extent ncRNAs conserved among free-living bacteria are maintained in the reduced genomes and to characterize the process of ncRNAs loss.

Materials and Methods

Genome Characteristics, Identification of Rfam Families, and Phylogeny

Sequences of 1,156 bacterial genomes with annotations were downloaded from National Center for Biotechnology Information (NCBI) Genome (Sayers et al. 2012). Descriptions of the organisms' lifestyle were extracted from NCBI Genome and corrected manually based on the information obtained from the literature (supplementary table S1, Supplementary Material online). Genome length, the number of Clusters of Orthologous Groups (COGs) (Tatusov et al. 2000) types, and the number of protein-coding genes were calculated based on concatenated chromosomes and

plasmids sequences of the individual organisms. Covariance models representing families from Rfam 11 (Burge et al. 2013) were used to scan genomes with cmscan (with default parameters; *E*-value threshold of 10^{-3}) (Nawrocki and Eddy 2013). The obtained matches were grouped to classes (tRNA, rRNA, RNase P, SRP, 6S, tmRNA, asRNA, standalone sRNA, ribosomal proteins regulator, riboswitch, thermoregulator, other *cis*-regulatory element) following Rfam annotations after manual correction (supplementary tables S2 and S3, Supplementary Material online). In addition, we divided the *cis*-regulatory families (i.e., ribosomal proteins regulators, riboswitches, and thermoregulators) according to their localization relative to the regulated genes: 5′-UTR, 3′-UTR, IGR (between Open Reading Frames [ORFs] in an mRNA), and ORF (if more than 70% of an ncRNA sequence overlaps with an ORF). The presence of the individual families and classes was mapped on a maximum-likelihood bacterial phylogeny tree calculated based on 31 protein markers obtained from the work of Wu and Eisen (2008). The γ-proteobacterial phylogeny was taken from the study of Husník et al. (2011) (inferred by maximum-likelihood approach from the set of 69 genes, under the CAT+GTR [general time reversible] model, after removal positions containing both A/T and G/C states). *Serratia symbiotica* and *Blochmannia vafer* were placed manually as sister species to *Serratia proteamaculans* and *Blochmannia floridanus*, respectively, based on the phylogenetic trees obtained from previous studies (Williams and Wernegreen 2010; Burke and Moran 2011).

Estimation of GC Content Detection Limits of Individual Rfam Covariance Models

To check whether the existing covariance models of Rfam families are able to detect ncRNAs with GC contents observed in the reduced bacterial genomes, "GC content detection index" was estimated for each model. This index corresponds to the minimal GC content of a sequence that enables its detection by a corresponding Rfam model. For each potentially lost Rfam family, a reference ncRNA sequence with the lowest observed GC content was selected from the most closely related free-living bacterium. Based on this sequence, an ensemble of 100 artificial test sequences with GC content lower than a given GC content threshold was generated. The artificial test sequences were generated by randomly mutating GC base pairs and unpaired G and C nucleotides according to the nucleotide occurrence frequencies observed in an Rfam seed alignment; that is, the probabilities of consecutive mutations were proportional to the relative occurrences of A and U residues in the respective columns of the seed alignment. The starting GC content threshold was set to 13%, which corresponds to the genomic GC content of *Candidatus Zinderia insecticola* CARI. If less than 90% of the generated sequences were not matching the corresponding Rfam model (as assessed with the model-specific bit-score gathering thresholds) (Nawrocki and Eddy 2013), the procedure was repeated with an increased GC content threshold (e.g., 15%, 18%, 20%, 23%, 25%) until at least 90% of the artificial sequences were matching the model (see fig. 1*A*).

We assumed that a given Rfam family can be confidently predicted to be absent if it passed the GC content test; that is, at least 90% of artificial sequences were detectable (at the bit-score gathering thresholds) at the GC content level corresponding to the average GC content of all IGRs of a given "reduced" genome. The results confirmed 68% of ncRNA absences (1,427 out of total 2,090 from the three groups). If the threshold is set based on the average genomic GC content, then the number of confirmed losses increases to 77%.

Estimation of Effectiveness of Rfam Covariance Models in the Detection of Truncated Sequences

To assess whether covariance models are sensitive to small deletions, which are frequently observed in the reduced genomes, the "sequence length limit index" was estimated for each model. Similarly to the GC content detection index, this limit corresponds to the minimal length of a sequence that enables its detection by a corresponding Rfam model. For each potentially lost Rfam family, the shortest sequence from its seed alignment was selected as a starting point. By introducing single nucleotide deletions at random positions (100 independent runs per sequence), an initial ensemble of test sequences was generated. Next, small deletions were introduced iteratively to each test sequence. To mimic deletions encompassing continuous fragments of a sequence, the deletions were introduced randomly, but only in positions adjacent to already deleted regions. Furthermore, to constrain the predicted secondary structure, the base-paired nucleotides were removed together. The final set of test sequences was scanned with cmsearch. The sequence length limit was calculated as the minimal length for which at least 90% of the generated sequences could be matched to the corresponding covariance model under the model-specific bit-score gathering threshold (see fig. 1*B*).

De Novo Prediction of Structured ncRNAs

To test whether our observations do not result from a bias specific to Rfam, we confronted them with the statistics based on de novo predicted *cis*-regulatory motifs. For each group (γ-proteobacteria, α-proteobacteria, and Tenericutes), we considered genomes shown in figure 5 and several additional closely related nonreduced genomes. From these genomes, all IGRs potentially encompassing a 5′-UTR were extracted and grouped according to the COG annotation of a downstream ORF. The groups were analyzed using CMfinder ("cmfinder -f 0.2 -n 5 -s 1 -m 30 -M 100," "cmfinder -f 0.2 -n 5 -s 2 -m 40 -M 100," followed by CombMotif) and the resulting motifs were scored (sp*lc*bp/sid, where sp is the number of sequences forming a motif, lc is the number of conserved

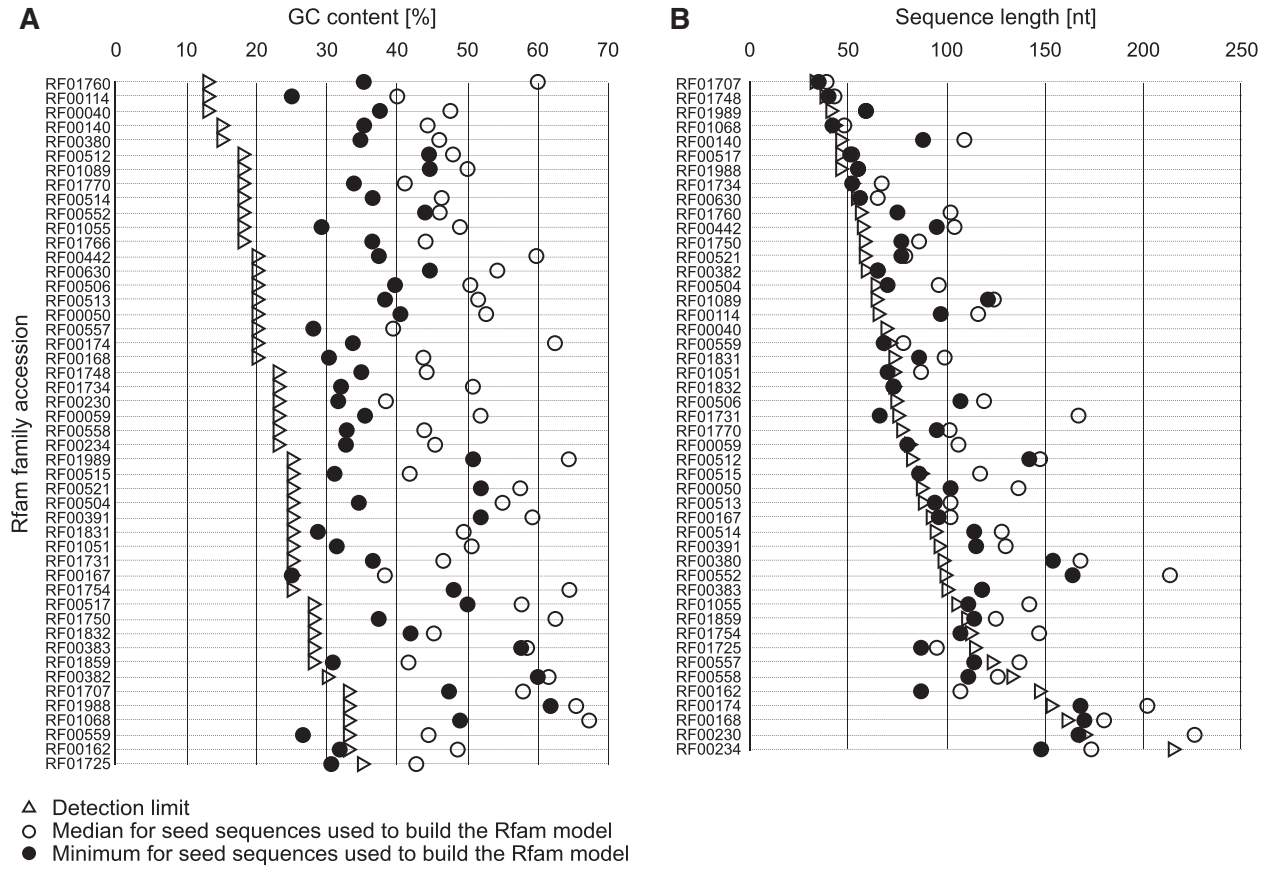

Fig. 1.—GC content (*A*) and sequence length (*B*) detection limits of individual Rfam *cis*-regulatory families. Triangles represent minimal GC content (*A*) and sequence length (*B*), which allow for the detection of Rfam representatives in at least 90% of test sequences with cmsearch, using model-specific bit-score gathering thresholds. The test sequences were derived from randomly mutated native family representatives from free-living relatives of reduced genomes. Black and transparent circles correspond to the minimal and median, respectively, GC content (*A*) and sequence lengths (*B*) of the seed alignments that were used to create Rfam models.

positions, bp is the number of base pairs, and sid is the average sequence identity of the sequences). The top 5% of the motifs were used as queries to scan all IGRs using cmscan (*E*-value < 0.01). For further analyses (fig. 5), we considered only the nonredundant motifs (<50% common matching IGRs), not covered by Rfam, and localized upstream of a conserved ORF (at least 25% of the matches belong to the same COG).

Assignment of Homologous IGR Pairs

For each 5′-UTR *cis*-regulatory ncRNA that is absent in a given reduced genome, we assigned the most closely related organism possessing at least one instance of this ncRNA. The evolutionary distance between organisms was calculated as the sum of lengths of connecting branches in the phylogenetic tree. For each ncRNA instance, its flanking protein-coding genes were extracted and their orthologs in the reduced genome were identified using bidirectional best Basic Local Alignment Search Tool (BLAST) hits approach (Wolf and Koonin 2012). If both orthologous genes in the reduced genome were adjacent, the IGR between them was considered to be a "certain" (two-sided) homolog of the IGR containing a *cis*-regulatory ncRNA (fig. 2A). If only an ortholog of a gene downstream to ncRNA was present, its upstream IGR was assigned as "potential" (one-sided) homolog.

To assess the statistical significance of changes in IGRs that have lost ncRNA elements, we built a background set of homologous IGR pairs from the reduced and nonreduced genomes, where neither of the two IGRs contains an ncRNA. As the bidirectional BLAST best hits approach would be very time consuming for such a number of genes to be compared (over 85,000 BLAST searches), we decided to use COG assignments to denote homology between proteins. Consequently, "potential" (one-sided) homologs are these IGRs that are localized upstream to protein-coding genes from the same COG, occurring once in each genome, and "certain"

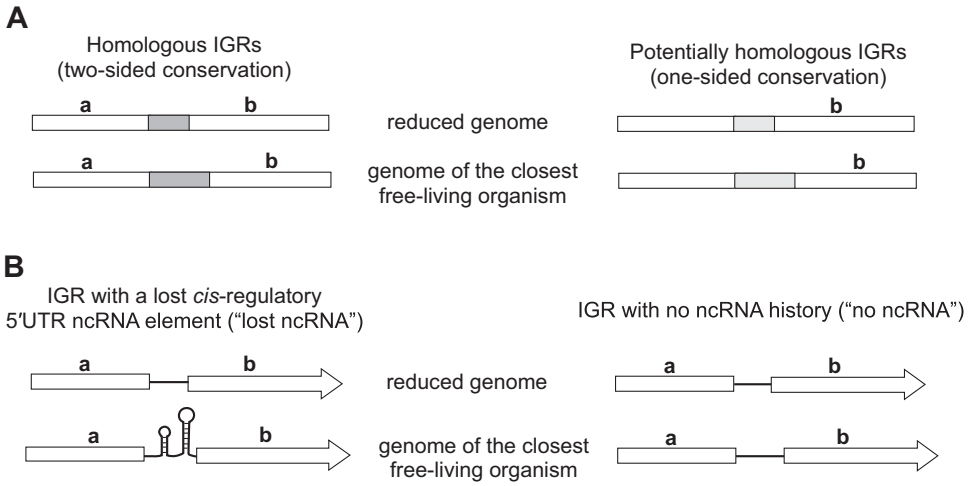

Fig. 2.—Homologous IGR pairs' definition procedure. (*A*) Definition of homologous IGRs based on the conservation of their flanking protein-coding genes. If both of them are conserved, the IGR is "two-sided" conserved. If only one of them (downstream to the potential 5′-UTR) is conserved, the IGR is "one-sided" conserved. (*B*) Definition of homologous IGRs that either have lost a *cis*-regulatory 5′-UTR ncRNA element upon genome reduction (lost ncRNA) or presumably did not contain ncRNA in the last common ancestor of the reduced and nonreduced organisms (no ncRNA).

(two-sided) homologs are these IGRs that are flanked by protein-coding genes of the same COGs, each of which occur once in each genome.

Comparison of Homologous IGR Pairs

The homologous IGR pairs were divided into two sets: Background (homologous IGR pairs where both sequences do not contain a *cis*-regulatory ncRNA) and test (homologous IGR pairs where a sequence from the reduced genome does not contain a *cis*-regulatory ncRNA, and a sequence from the free-living genome contains an ncRNA). In the background set, we included homologous IGR pairs that fulfill the following conditions: 1) Both IGRs do not contain an ncRNA, 2) both IGRs are shorter than 2,000 nt, 3) the length of an IGR from the nonreduced genome equals or is greater than 100 nt, 4) the flanking genes are localized on the same strand or on opposite strands (in the latter case, the 5′-ends must face each other), and 5) in the case of comparison of GC content, the length of an IGR from the reduced genome equals or is greater than 10 nt. In the test set, we included IGR pairs that fulfill the following conditions: 1) The IGR from the reference genome contains a *cis*-regulatory ncRNA, whereas the IGR from the reduced one does not contain any ncRNA; 2) both IGRs are shorter than 2,000 nt; 3) the length of an IGR from the reference genome equals or is greater than 100 nt; and 4) in the case of GC content comparison, the length of an IGR from the reduced genome equals or is greater than 10 nt.

The background and test sets were further subdivided into clades according to the taxonomy of the reduced genomes and closest free-living organisms: Five clades of γ-proteobacteria, two clades of Mollicutes, and one clade of *Rickettsiales*. For each clade, we calculated two distributions. In the first one, a sample was the length difference between an IGR from the reduced genome and its homolog from the free-living organism. In the second one, a sample was a corresponding difference in the GC content. Boxplots of pairs of the distributions are plotted in figure 3.

Distributions of IGR length differences and GC content differences from test and background sets from the individual clades were independently compared using a *t*-test. Cases where one of the distributions contains an insufficient number of samples were discarded. Significantly different distributions are indicated with stars near the top of each box in figure 3 ($P < 0.05$).

Experimental Determination of 5′-UTR Length

The RNA-seq data on *Pantoea stewartii* DC283 transcriptome were downloaded from NCBI SRA (accession SRX529441) (Ramachandran et al. 2014). Reads were mapped on scaffolds of *P. stewartii* DC283 genome (NCBI Genome accession numbers NZ_AHIE01000001–NZ_AHIE01000065) using megablast ($E < 1e^{-10}$) (Camacho et al. 2009). The gene annotations were downloaded from NCBI Genome (Acland et al. 2014), and mapping coverage of the genes of interest was visualized with in-house scripts. 5′-UTR lengths were estimated based on the relative coverage of the reads upstream of the annotated coding sequences.

Total RNA was isolated from *Pantoea ananatis* LMG 2665 strain using standard protocols (phenol–chloroform extraction). The 3 μg of total RNA was subjected to the 5′-RACE with a "5′ RACE System for Rapid Amplification of cDNA

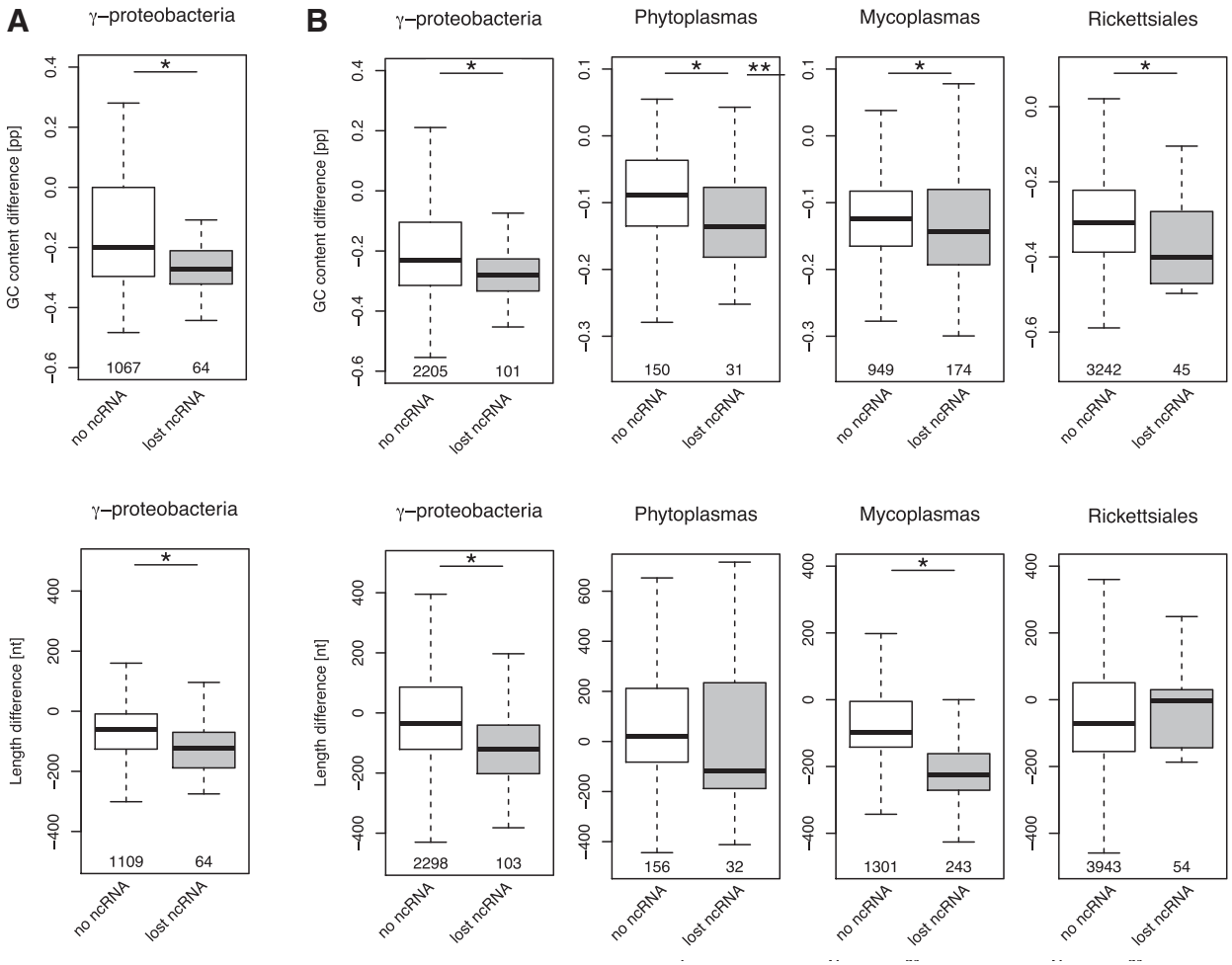

Fig. 3.—Box-and-whisker plots of differences in GC content (upper panels) and length (lower panels) between pairs of homologous IGRs from reduced and free-living genomes. Compared IGR pairs were defined based on the (*A*) two-sided and (*B*) one-sided conservation of flanking protein-coding genes. On the left hand, white plots (no ncRNA) describe the average changes in IGR pairs where neither of the two IGRs contains an ncRNA, whereas on the right hand, gray plots (lost ncRNA) describe the changes in IGR pairs where the IGR from the nonreduced genome contains a *cis*-regulatory 5′-UTR ncRNA and the IGR from the reduced genome does not contain an ncRNA. The upper and lower limits of the boxes represent the first and third quartiles of differences. The horizontal bar in the middle of each box represents the median for each group. The upper (and lower) limits of the whiskers represent upper (lower) quartiles plus (minus) 1.5 times the length of interquartile range. Outliers are not shown.

Ends, version 2.0" (ThermoFisher Scientific) according to the manufacturer's protocol. Briefly, the first-strand cDNA was synthesized from total RNA using a gene-specific primer 1 (GSP1, supplementary table S5, Supplementary Material online) and SuperScript II reverse transcriptase. After cDNA synthesis, the template was removed by the treatment with RNAseH/T1 mix, and the product was purified using S.N.A.P column (provided in the kit). In the next step, the homopolymeric tail (polyC) was added to the 3′-end of the cDNA using terminal deoxynucleotidyl transferase. Subsequently, polymerase chain reaction amplification of the tailed cDNA was performed using a second gene-specific primer 2 (GSP2, nested with respect to the cDNA primer; supplementary table S5, Supplementary Material online) and Abridged Anchor Primer (AAP, provided in the kit).

Results

To get an insight into the repertoire of evolutionarily conserved ncRNAs in bacteria, we scanned 1,156 fully sequenced bacterial genomes (supplementary table S1, Supplementary Material online) using covariance models of 2,208 ncRNA families from Rfam (Burge et al. 2013). Representatives of 506 of these families were found in at least one genome, out of which 111 (22%) were present in at least two distinct phyla. Although most of the Rfam families have narrow

taxonomic distribution, we found that the number of ncRNA families encoded in a genome positively correlates with the genome size (Spearman correlation coefficient, $\rho = 0.53$, $P < 0.05$) as well as with the proteome complexity measured as the number of COGs (Tatusov et al. 2000) into which proteins encoded by that genome can be classified ($\rho = 0.62$, $P < 0.05$, fig. 4 and supplementary fig. S1, Supplementary Material online). These correlations are typically more pronounced when calculated within the taxonomic phyla (e.g., γ-proteobacteria $\rho = 0.71$ for COGs; α-proteobacteria $\rho = 0.84$; and Tenericutes $\rho = 0.66$). It has been known that intracellular bacteria are characterized by a reduced genome size and low proteome complexity (Mira et al. 2001). Our results reveal that they also tend to have considerably less evolutionarily conserved ncRNA families and that this feature can be used to separate bacteria into intracellular and free-living ones (supplementary fig. S2, Supplementary Material online).

Smaller genomes tend to have considerably lower GC content (Moran 2002). This feature, in turn, may hamper the detection of ncRNAs using profiles calculated by using sequences obtained from genomes of free-living bacteria. To address this issue, we carried out a test aiming at verifying the effectiveness of Rfam covariance models in detecting the low-GC content sequences (see Materials and Methods). The obtained results show that 50% (52 of 104) and 89% of Rfam profiles are capable of detecting artificially generated ncRNA sequences with GC content reduced to 20% and 30%, respectively (fig. 1*A* and supplementary fig. S3*a*, Supplementary Material online). We also validated the effectiveness of covariance models in the detection of ncRNA sequences with deletions, which have been shown to occur in some ncRNAs encoded in reduced genomes (e.g., RNase P and tmRNA; see Discussion for details). We found that the vast majority of Rfam covariance models (79/104) is sensitive enough to detect matches in sequences shortened by 10% (fig. 1*B* and supplementary fig. S3*b*, Supplementary Material online). Larger deletions, comprising 20% of nonreduced sequences, are tolerable in a half of the models (50/104). Altogether, these results indicate that low GC content and small deletions, in general, do not prevent usage of covariance models for scanning genomic sequences of reduced bacteria.

Another issue that might hamper the detection of ncRNAs in reduced genomes is the fact that Rfam covariance models are biased toward model organisms (fig. 4) and ncRNAs conserved in the available genomes. Intuitively, the broader the definition of the covariance model (in terms of the evolutionary divergence of the genomes whose sequences were used to build seed alignments, hereinafter referred to as "seed genomes"), the more sensitive it is in finding new homologs (supplementary fig. S4A, Supplementary Material online, Pearson correlation coefficient $r = 0.69$, $P < 0.05$). We found that, for most Rfam families, the evolutionary distance between the seed genomes and reduced genomes is lower than the evolutionary distance between the seed genomes

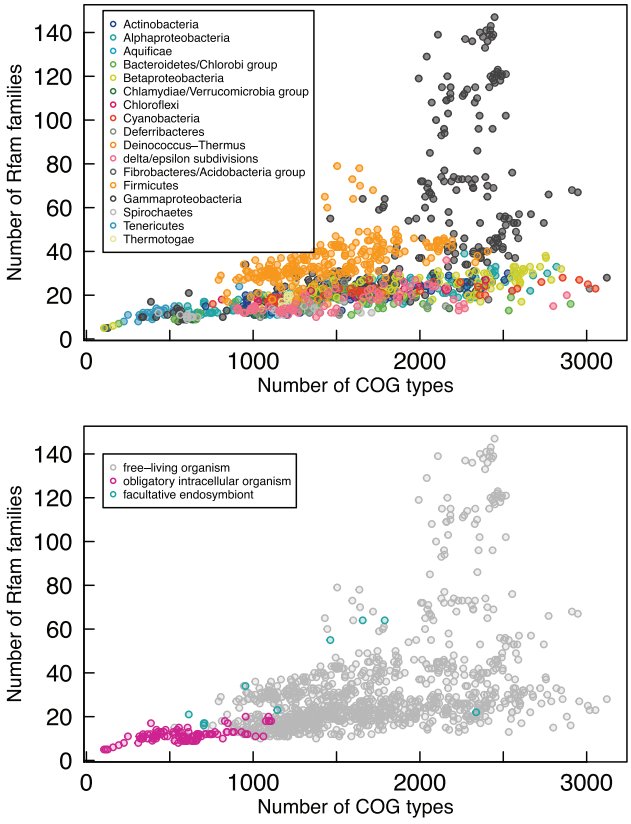

Fig. 4.—The relationship between the number of ncRNA families and the proteome complexity in bacteria. Plot (*A*) is colored according to the taxonomic phyla, whereas plot (*B*) is colored according to the lifestyle.

and genomes in which a covariance model is capable of detecting homologous sequences (supplementary fig. S4*b*, Supplementary Material online). This suggests that the evolutionary distance between the reduced genomes and seed genomes does not hamper the sensitivity of the covariance models and supports the notion that lack of a match to a given covariance model is indicative of ncRNA loss.

We categorized ncRNA families into groups according to their function (supplementary table S2, Supplementary Material online) and mapped their occurrences onto the bacterial phylogenetic tree. Besides the ubiquitous ribosomal RNA and transfer RNA, tmRNA (RF01850/RF01849), SRP RNA (RF00169/RF01854), and RNase P RNA (RF00010/RF00011) are also present in almost all bacteria regardless of their lifestyle (supplementary table S3, Supplementary Material online). As reported by others (Willkomm et al. 2002), conserved RNase P components are missing in the Aquificaceae family, despite *Aquifex aeolicus* was shown to exhibit RNase P-like trimming of tRNA precursors (Marszalkowski et al. 2008). We found that besides Aquificaceae, both RNase P components are also missing in thermophilic organisms *Thermocrinis albus* DSM 14484, *Hydrogenobaculum* sp.

Y04AAS1, *Thermodesulfovibrio yellowstonii* DSM 11347, as well as in extremely reduced endosymbionts *Ca. Carsonella ruddii* PV and *Ca. Zinderia insecticola* CARI. Similarly, a few reduced genomes lack both RNA and protein components of the SRP: *Ca. Hodgkinia cicadicola* Dsem, *Ca. Tremblaya princeps* PCIT, *Ca. Zinderia insecticola* CARI, *Ca. Carsonella ruddii* PV, *Ca. Blochmannia vafer* str. BVAF, and *Ca. Sulcia muelleri* DMIN, as well as *Elusimicrobium minutum* Pei191 and *Dehalococcoides* spp.

In accordance with our aforementioned observations, the *cis*-regulatory Rfam ncRNA families (i.e., thermometers, riboswitches, and ribosomal leaders) are also generally absent in bacteria that depend on eukaryotic host cells. Widely distributed families, such as cobalamin riboswitch (RF00174), pfl RNA (RF01750), and ykkC–yxkD leader (RF00442), were independently lost in the multiple clades. The other, less ubiquitous, ncRNA families (e.g., flavin mononucleotide [FMN; RF00050], lysine [RF00168], and Moco [RF01055] riboswitches) are also typically absent in the three groups. In contrast to the above, the thiamine pyrophosphate (TPP) riboswitch (RF00059), S-adenosylmethionine riboswitch (RF00162), purine riboswitch (RF00167), T-box leader (RF00230), and several ribosomal protein regulators such as L10, L20, and L21 leaders (RF00557, RF00558, and RF00559, respectively) are preserved in some of the reduced genomes.

To gain insight into the process of *cis*-regulatory ncRNAs' loss, we analyzed in greater detail three distinct bacterial clades that include reduced genomes: 1) The γ-proteobacterial symbionts of insects, a model group for comparative studies of a bacterial genome reduction; 2) *Rickettsiales*, a group of obligate intracellular α-proteobacteria and closest extant relatives of the mitochondria ancestor; and 3) Mollicutes, cell wall-less obligate parasites including mycoplasmas, ureaplasmas, and phytoplasmas (*Tenericutes phylum*). The trend of the loss of *cis*-regulatory Rfam families is apparent in each of the three clades (fig. 5, for complete census see supplementary fig. S5, Supplementary Material online). Given the fact that Rfam database is biased toward nonreduced genomes, we decided to complement the initial scan with Rfam covariance models with de novo predictions using CMfinder (see Materials and Methods section). The pattern of loss of Rfam-defined and predicted *cis*-regulatory ncRNA families at the transition from free-living to intracellular lifestyle is very similar in γ-proteobacteria ($r = 0.97$, $P < 0.05$) and *Rickettsiales* ($r = 0.94$, $P < 0.05$). Such a correlation is less pronounced in Mollicutes because in this group, we predicted only eight motifs, which are mostly conserved both in reduced and nonreduced genomes. This is presumably due to the limited number of available sequenced nonreduced genomes that are closely related to mycoplasmas. As in the case of Rfam, we found many predicted motifs that are specifically absent in the reduced genomes (see supplementary fig. S6, Supplementary Material online) and only a few that are preserved (e.g., *rpsF* motif [Matelska et al. 2013] is present in many reduced γ-

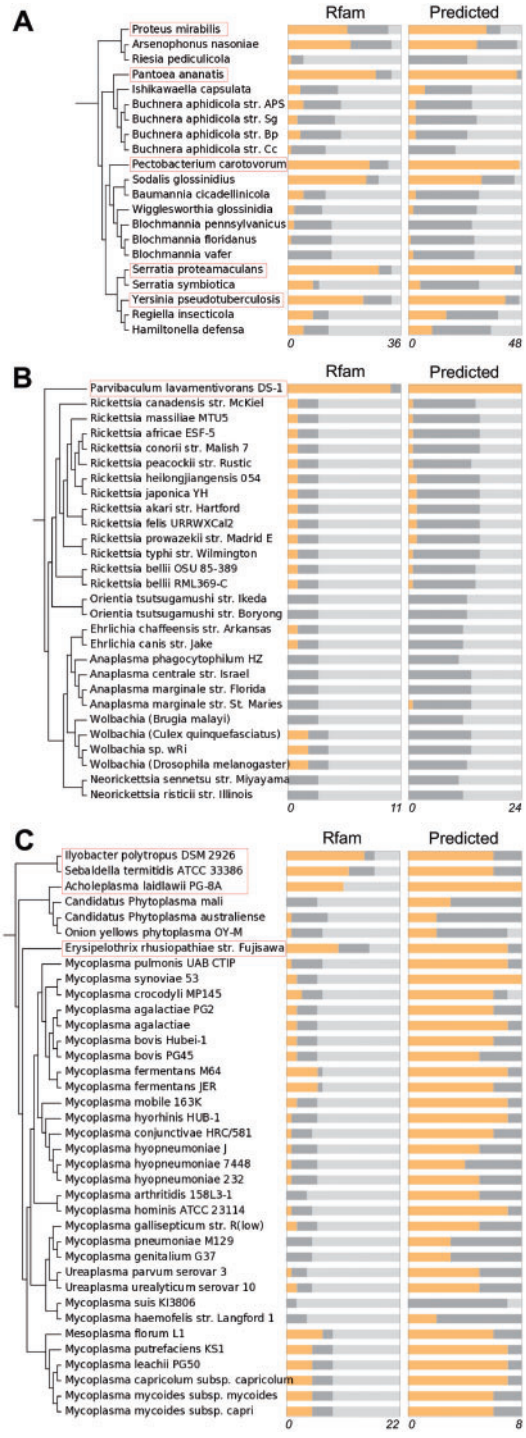

Fig. 5.—Distribution of Rfam and predicted 5′-UTR *cis*-regulatory ncRNA families in (*A*) γ-proteobacterial endosymbionts, (*B*) α-proteobacteria (*Rickettsiales*), and (*C*) Mollicutes (mycoplasmas, phytoplasmas, and ureaplasmas) and their closest free-living relatives. Orange bars indicate the number of ncRNAs in a given organism. Dark gray bars indicate the number of cases in which an ncRNA is absent, but its downstream protein-coding gene is conserved, whereas light gray bars indicate the number of cases in which both ncRNA and its genomic surrounding are not conserved. Nonreduced genomes are outlined with red boxes.

proteobacterial and Mollicutes genomes). The consistency of the results obtained with Rfam and CMfinder implies that the observed tendencies are not a result of Rfam bias. Moreover, the fact that the predictions did not yield motifs specific for reduced genomes (i.e., conserved in these genomes, but not similar to any motif present in nonreduced genomes) shows that the absent *cis*-regulatory ncRNAs were not replaced with novel motifs.

Interestingly, secondary facultative γ-proteobacterial symbionts (*Arsenophonus nasoniae*, *S. glossinidius*, *Se. symbiotica*, *Regiella insecticola*, and *Hamiltonella defensa*) tend to contain more *cis*-regulatory elements when compared with obligatory primary symbionts (fig. 5). Moreover, we observe that the number of *cis*-regulatory elements decreases along with the evolutionary distance from a reference free-living genome. For example, in nonreduced genome of *Proteus mirabilis*, we detected 19 Rfam and 33 predicted motifs, in its closest reduced relative, secondary endosymbiont *A. nasoniae*, 20 and 29, and in primary endosymbiont *Riesia pediculicola*, only 1 and 0. A similar pattern can be observed in *Pectobacterium carotovorum* (26/47), a related secondary endosymbiont *S. glossinidius* (25/31), and a primary endosymbiont *Baumania cicadellinicola* (5/3). The limited number of closely related sequenced genomes that represent different stages of genome reduction hampers the possibility to estimate robustly the rate of *cis*-regulatory elements' loss. However, the abovementioned observations suggest that the number of preserved *cis*-regulatory ncRNAs is correlated with the degree of host-dependence.

As expected, all *cis*-regulatory ncRNAs preserved in the reduced genomes are localized upstream of the same protein-coding genes as their homologs from the free-living bacteria. In turn, for many of the lost *cis*-regulatory ncRNAs, we could not detect close homologs of their flanking protein-coding genes in the free-living bacteria (fig. 5, light gray bars), suggesting that these RNAs and their neighboring genes were lost together. However, we identified ncRNAs that are preferentially lost, but their genomic neighborhood (i.e., flanking protein-coding genes) is preserved (fig. 5, dark gray bars; supplementary fig. S5, Supplementary Material online, dark and light gray boxes). As RNAs with the *cis*-regulatory function are known (or predicted) to regulate the expression of their cognate downstream protein-coding regions, it is reasonable to assume that the genomic arrangement of these two elements is conserved. Here, we assume that this conservation persists even upon the loss of ncRNA, and thus the "residual" nonfunctional IGRs from reduced genomes are homologous to the structured *cis*-regulatory RNA elements from nonreduced genomes. To determine such homology relationships, we used two approaches: A "two-sided" approach where both protein-coding genes flanking a given IGR are conserved and a "one-sided" approach where only a downstream protein-coding gene is conserved (for details, see also Assignment of Homologous IGR Pairs in Materials and Methods). IGR pairs defined by the conservation of both flanking protein-coding genes could be determined only in γ-proteobacteria.

The comparison of "ncRNA remnant" sequences from the reduced genomes to their RNA-containing homologs from the reference free-living genomes revealed that they have a lower GC content (−27 percentage points [pp] on average) and are predominantly shorter (−121 nt on average; fig. 3*A*, "lost ncRNA" panels). To check whether these effects are characteristic for the *cis*-regulatory ncRNA loss process, we defined a background set of homologous IGR pairs, where neither of the two IGRs contains a candidate ncRNA. Comparison of such 1067 IGR pairs revealed significantly ($P < 0.05$, one-sided *t*-test) lower level of GC content reduction (−20 pp on average, fig. 3*A*, "no ncRNA" panels). We also found that the length reduction (−62 nt on average) is significantly less pronounced in the background IGR pairs ($P < 0.05$, one-sided *t*-test).

Because the "two-sided" homologous IGR pairs could be determined only in γ-proteobacteria, we investigated IGR pairs defined in a less reliable manner; that is, by conservation of a downstream protein-coding gene only. Also for the one-sided conservation data set, the average GC content reduction level in γ-proteobacteria endosymbionts, phytoplasmas, mycoplasmas, and *Rickettsiales* is only slightly (but significantly) higher for the IGRs, which have lost their ncRNA than in those that have never contained any ncRNA (−28/−23, −14/−9, −14/−12, and −40/−31 pp, respectively, see fig. 3*B*). Analogously, in γ-proteobacterial endosymbionts and mycoplasmas, the average IGR length decrease is significantly higher for IGRs ($P < 0.05$, one-sided *t*-test), which have lost their ncRNA than in those that have never contained any ncRNA (−120/−36 and −225/−97 nt, respectively). These results suggest that IGR length and GC content reduction correlating with the *cis*-regulatory ncRNA loss can be observed not only in γ-proteobacteria but also in other taxonomic groups.

The fact that many *cis*-regulatory ncRNAs are lost without the cognate protein-coding genes and that their loss is associated with the shrinkage of the corresponding IGRs prompted us to investigate the possibility that this is caused by the shortening of 5′-UTRs. To this end, we compared the lengths of experimentally confirmed 5′-UTRs of five homologous genes from reduced genome of *B. aphidicola* str. Sg (Hansen and Degnan 2014) and two closely related nonreduced genomes of *P. stewartii* (Ramachandran et al. 2014) and *P. ananatis* (this study). In both *Pantoea* genomes, 5′-UTRs of all five genes are long enough to contain structured *cis*-regulatory motifs (RF00127 and RF00050 in *rpsB* and *ribB*, respectively, and predicted motifs in *rpmB*, *rplN*, and *rpmE*; see supplementary fig. S7, Supplementary Material online). In *Buchnera*, all the corresponding 5′-UTRs are considerably shorter and neither contained Rfam nor predicted structured motifs (table 1).

Table 1
Experimentally Determined 5′-UTR Lengths of Five Genes That Lost Their *Cis*-Regulatory RNA Elements

	rplN (~145)[a]	rpmB (~143)	rpmE (~78)	rpsB (~96)	ribB(~136)
Buchnera aphidicola str. Sg (RNA-seq)	22[b]/–[c]	48/–	31/–	42/–	78/–
Pantoea stewartii (RNA-seq)	~145/+	~225/+	~90/+	nd/+	~390/+
Pantoea ananatis (5′-RACE)	nd/+	~210/+	nd/+	~211/+	~348/+

Note.—nd, no data.
[a]*Cis*-regulatory RNA element length (nt).
[b]Length of 5′-UTR (nt).
[c]*Cis*-regulatory RNA element present (+) or absent (–).

Discussion

The size and coding capacity of a bacterial genome reflect a balance between the acquisition and loss of genetic material (Mira et al. 2001). In the highly reduced genomes of endosymbionts and intracellular parasites, the loss of genetic material is a dominating process. Cases of independent reductive evolution have been described for groups as diverse as α-proteobacteria (*Rickettsia* spp.), γ-proteobacteria (*Buchnera* spp.), and Tenericutes (*Mycoplasma* spp.). Although they originate from different ancestors, further observations suggested that mechanisms governing their genome reduction can be generalized (McCutcheon and Moran 2012).

As a consequence of coevolution with a host, the genomes of intracellular bacteria changed considerably. On the one hand, reduced levels of purifying selection on a large number of genes are observed (Shigenobu et al. 2000). On the other, some genes are preserved but frequently lack systems of expression regulation (Wilcox et al. 2003; Moran et al. 2005; Dale and Moran 2006). Also, bioinformatics analysis showed that the obligate symbionts have the smallest number of transcription factors and their genetic networks are smaller and less modular in comparison to species from other types of environments (Parter et al. 2007).

Here, we aimed at characterizing the process of the loss of evolutionarily conserved ncRNA in intracellular bacteria from three distinct phyla. In particular, we focused on a group of conserved *cis*-regulatory ncRNA elements localized within 5′-UTRs of protein-coding genes. Using covariance models representing sequences and secondary structures of conserved ncRNA families, we showed that most ncRNA families, which are present in genomes of free-living bacteria, are absent in the reduced genomes of the three distinct phyla.

It is important to emphasize that the Rfam covariance models used in this study were calculated mainly based on RNA sequences from free-living organisms (not always including close relatives of bacteria with reduced genomes), raising the possibility that their sensitivity will be low for detecting ncRNA homologs in genomes of host-restricted bacteria. First, to check whether our observations could result from GC content bias (reduced genomes tend to be AU-rich), we evaluated the sensitivity of the covariance models against artificially AU-enriched sequences (see Materials and Methods). For each Rfam family, we computed the GC content detection limit, which denotes the minimal GC content of artificial sequences that allow for the detection of family representatives in at least half of the sequences. The obtained results show that for 93 of the available 104 Rfam families (absent in at least one of the three groups considered in this study), 30% GC content is not a limitation in detecting homology (fig. 1*A*). The computed GC content detection limits are significantly lower than minimal GC content of the native ncRNA sequences found in bacteria: For example, RNase P RNA (RF00010)—30% (in *Ureaplasma parvum* serovar 3 str. ATCC 27815) versus calculated 18%, SRP RNA (RF00169)—28% (*Ca. Phytoplasma mali*) versus 23%. This result suggests that most of Rfam families that are not detected in the reduced genomes were indeed lost and that regardless of the average GC content of the seed sequences, the sensitivity of their covariance models is good even for the artificial target sequences in which the GC content was decreased by approximately 18 pp (fig. 1*A*). Second, the length reduction has been observed in some of ncRNAs that are preserved in the reduced genomes, for example, RNase P (Siegel et al. 1996) and tmRNA (Gueneau de Novoa and Williams 2004). To assess whether covariance models can theoretically detect such truncated sequences, we estimated minimal sequence lengths allowing for the ncRNA detection with Rfam covariance models (see Materials and Methods). The theoretical sequence limits for RNase P (RF00010) and tmRNA (RF00023) are considerably lower than the lengths of the reported shortest sequences from the reduced genomes: 213 versus 276 nt (Siegel et al. 1996) and 166 versus 262 nt (Gueneau de Novoa and Williams 2004), respectively. In general, our results suggest that most of the Rfam models are sensitive enough to detect matches in the sequences with deletions comprising 18% of the native starting sequence. Third, we checked the possibility that ncRNAs have not been detected in the reduced genome due to the evolutionary distance between the reduced genomes and genomes from which the covariance models' seed sequences were obtained (seed genomes). We found that most covariance models, corresponding to the lost ncRNA families, are capable of detecting ncRNAs in genomes, where the evolutionary distance from seed genomes is larger than the evolutionary distance between seed genomes and

reduced genomes (supplementary fig. S4, Supplementary Material online). Finally, we used CMfinder for de novo prediction of ncRNA motifs in reduced and closely related nonreduced genomes. The fact that the distribution of the predicted motifs over the phylogenetic trees resembles the one obtained with Rfam models supports the notion on structured ncRNA loss in reduced genomes.

Some of the Rfam families (e.g., RF00504, RF00059, RF00174, RF01750, and RF00442) have members in the vast majority of nonreduced genomes analyzed in this study (see supplementary fig. S5, Supplementary Material online). Thus, their absence in the related reduced genomes is clearly a result of the independent loss events. A similar pattern is seen in the γ-proteobacterial symbionts, which do not form a monophyletic group, but rather have evolved from several free-living bacteria (Husník et al. 2011). We showed that these free-living bacteria share many conserved ncRNA families (supplementary fig. S5*b*, Supplementary Material online) and that most of them were independently lost in the separate clades of γ-proteobacterial endosymbionts. Moreover, we observed that "early-stage" γ-proteobacterial secondary symbionts with relatively large genomes (e.g., *A. nasoniae* and *S. glossinidius*) retained many of the ncRNAs found in their closely related free-living bacteria.

The repertoire of riboswitches, which bind small ligands and typically negatively regulate expression of genes involved in biosynthesis, catabolism, and transport of metabolites, is greatly reduced in γ-proteobacterial endosymbionts. This may result from either the absence of pathways involving proteins encoded by these genes or from the simplification of their expression regulation systems (Shigenobu et al. 2000; Moran et al. 2005). We observed that the majority of riboswitches is lost together with their cognate protein-coding genes, suggesting that the first scenario is more common. For example, the TPP riboswitch, which in *P. ananatis* regulates expression of genes for transport and biosynthesis of thiamine, disappears in *Buchnera* genomes together with the thiamine biosynthesis pathway. The second scenario can be exemplified by the FMN riboswitch, which in *Pantoea* regulates expression of *ribB*, a gene involved in the riboflavin synthesis, and is lost without the cognate gene in *Buchnera* (table 1), as *Buchnera* needs this pathway to supply riboflavin to the host cells (Nakabachi and Ishikawa 1999). Overall, we observed over 100 cases in which the conserved *cis*-regulatory element is lost without its cognate protein-coding gene, and it can be speculated that some of these loss events occurred due to selection. For example, the removal of *cis*-regulatory negative regulators of amino acid metabolism operons (supplementary fig. S5*b*, Supplementary Material online) is presumably beneficial as it ensures a constant overproduction of amino acids that are provided to the host (Moran et al. 2005). Analogously, the loss of the aforementioned FMN riboswitch, which represses riboflavin biosynthesis, could also be attributed to the selection. The case of *cis*-regulatory elements' loss in the *ibpA* and *rne* genes (encoding for a heat-shock protein and ribonuclease E, respectively) is less clear. In *E. coli*, the expression of *ibpA* is induced upon heat stress through RpoH sigma factor and ROSE thermometer (RF01832) (Kortmann and Narberhaus 2012). In *B. aphidicola, ibpA* is also under control of RpoH (Wilcox et al. 2003) but lacks the ROSE *cis*-regulatory element (supplementary fig. S5*b*, Supplementary Material online). The *rne* genes lack RpoH recognition sequences in their promoters both in *B. aphidicola* and *E. coli*, although *rne* from *Buchnera* shows response to heat stress (Wilcox et al. 2003) while it lacks a negative regulator (RF00040). The loss of these regulators is not apparently beneficial, thus the removal could also be ascribed to the drift.

Only a few of the *cis*-regulatory ncRNAs are preserved in reduced genomes and many of them are found in the context of ribosomal proteins operons (e.g., CMfinder motifs were detected upstream of S1, S6, and S10 genes in γ-proteobacteria; S2, S6, S15, and L11 in Tenericutes; and S2 and L11 in α-proteobacteria), suggesting that the regulation of their expression remains essential in the reduced genomes. In *E. coli*, ribosomal protein S1 *cis*-regulatory element is responsible for the translation initiation in the absence of Shine–Dalgarno motif and autogenous expression regulation. The presence of S1 regulator in *Buchnera* str. APS and Sg was suggested by Tchufistova et al. (2003). Considering the fact that the S1 motif in *Buchnera* is only partially transcribed (Hansen and Degnan 2014) and its regulatory abilities could not be shown in *E. coli* (Tchufistova et al. 2003), one can speculate that it participates in translation initiation only. It is however important to note that presumably untranscribed part of the motif is conserved in *Buchnera*, including the highly conserved GGA motif in the apical loop I (Tchufistova et al. 2003), suggesting that the S1 motif in *Buchnera* might be fully functional but diverged. Another example of preserved *cis*-regulatory element is *cspA* thermoregulator (RF01766), which regulates cold shock response (Kortmann and Narberhaus 2012) and is detectable in nearly all the γ-proteobacterial endosymbionts (supplementary fig. S5, Supplementary Material online), sometimes in two copies (supplementary table S4, Supplementary Material online). Its homologs were identified everywhere but *B. aphidicola* str. Sg and *Bl. vafer*.

To elucidate the process of the removal of *cis*-regulatory structural RNA elements from IGRs, we used an approach in which we compared the syntenic regions across the genomes, similar to those previously applied for protein-coding genes (Silva et al. 2001). In the γ-proteobacterial clades, we defined 1,173 homologous IGR pairs based on the conservation of the flanking protein-coding genes. In each pair, one IGR originates from a reference nonreduced genome (red names in supplementary fig. S5, Supplementary Material online) and the other from one of 16 related reduced genomes. The comparison of these IGR pairs indicated that the decrease in GC content and length is more pronounced in IGRs that lost an ncRNA than in

IGRs, which never contained any ncRNA (fig. 3). The fact that the IGRs that lost an ncRNA shrank and have reduced GC content is not surprising, as partially disrupted, nonfunctional RNA structures are of no use and are not constrained toward elevated GC content (Rivas and Eddy 2000). These findings are also in agreement with the observation that loss of *cis*-regulatory elements is associated with the shortening of the corresponding 5′-UTRs (table 1).

We showed that many of the structured ncRNAs conserved in free-living bacteria become obsolete in host-restricted bacteria, expanding our knowledge on the genomic regions under weakened selection forces after the host restriction. Most of the *cis*-regulatory ncRNAs localized in the 5′-UTRs of protein-coding genes were lost, regardless of whether the associated proteins were preserved or lost, and only a few are still detectable in the reduced genomes we analyzed in this work. The fact that many proteins, which are regulated in nonreduced genomes and are conserved in reduced genomes, lost their *cis*-regulatory RNAs suggests that the presence of such a "fine-tuning" regulation is optional and not indispensable in the stable environment provided by a host. This hypothesis is further supported by the observation that genes that have lost their *cis*-regulatory elements have also considerably shorter 5′-UTRs (table 1). However, it must be emphasized that the loss of conserved regulatory ncRNAs does not necessarily imply that reduced genomes use ncRNAs to a lesser extent. First, recent research clearly indicates an abundance of asRNAs in the genomes of *Mycoplasma* and *Buchnera* (Güell et al. 2009; Hansen and Degnan 2014). It has been even suggested that the genome size negatively correlates with the number of antisense transcripts (i.e., small genomes tend to contain more asRNAs) (Qiu et al. 2010). However, the transcriptomics analyses indicate that asRNA expression is comparable between the genomes of the host-restricted and free-living bacteria (Güell et al. 2011). Second, reduced genomes, such as *M. pneumoniae*, show surprising adaptation capabilities, comparable to those of free-living bacteria, which are suggested to involve ncRNAs (Yus et al. 2009). Finally, it has been shown that in *Buchnera*, many 5′-UTRs are potentially structured (Degnan et al. 2011; Hansen and Degnan 2014). Our de novo CMfinder predictions are generally in agreement with the aforementioned study, but at the same time, we observe that many *cis*-regulatory motifs conserved in *Pantoea* are absent in *Buchnera* (see supplementary figs. S5 and S6, Supplementary Material online). Moreover, CMfinder did not reveal any structured ncRNA motifs that are specifically conserved in the reduced genomes, further supporting the notion of the loss of structured ncRNA. Thus, we may observe two overlapping effects: On the one hand, loss of conserved ncRNAs, including specific loss of *cis*-regulatory elements without the protein-coding genes they regulate; on the other, emergence of new ncRNAs, which are undetectable by the approach we used due to the lack of sequence and/or structural conservation. The latter effect is clearly reflected in the results of a recent comparative study, which takes advantage of a growing repertoire of transcriptomics data (Lindgreen et al. 2014). The authors show that many ncRNAs are only conserved at approximately the genera level, implying that the emergence of new ncRNAs is to be expected.

Supplementary Material

Supplementary figures S1–S7 and tables S1–S5 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was supported by Polish National Science Centre (NCN) (2011/03/D/NZ8/03011 to S.D.H.) and Foundation of Polish Science (FNP) (TEAM/2009-4/2 to J.M.B.). This was also supported by a fellowship for outstanding young scientists from the Polish Ministry of Science and Higher Education (to S.D.H.) and the "Ideas for Poland" fellowships from the FNP (to J.M.B.).

Literature Cited

Acland A, et al. 2014. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 42:D7–D17.

Belda E, Moya A, Silva FJ. 2005. Genome rearrangement distances and gene order phylogeny in gamma-Proteobacteria. Mol Biol Evol. 22:1456–1467.

Bennett GM, Moran NA. 2013. Small, smaller, smallest: the origins and evolution of ancient dual symbioses in a Phloem-feeding insect. Genome Biol Evol. 5:1675–1688.

Burge SW, et al. 2013. Rfam 11.0: 10 years of RNA families. Nucleic Acids Res. 41:D226–D232.

Burke GR, Moran NA. 2011. Massive genomic decay in *Serratia symbiotica*, a recently evolved symbiont of aphids. Genome Biol Evol. 3:195–208.

Camacho C, et al. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.

Cole ST, et al. 2001. Massive gene decay in the leprosy bacillus. Nature 409:1007–1011.

Dale C, Moran NA. 2006. Molecular interactions between bacterial symbionts and their hosts. Cell 126:453–465.

Degnan PH, Ochman H, Moran NA. 2011. Sequence conservation and functional constraint on intergenic spacers in reduced genomes of the obligate symbiont *Buchnera*. PLoS Genet. 7:e1002252.

Eddy SR. 2001. Non-coding RNA genes and the modern RNA world. Nat Rev Genet. 2:919–929.

Fu Y, Deiorio-Haggar K, Anthony J, Meyer MM. 2013. Most RNAs regulating ribosomal protein biosynthesis in *Escherichia coli* are narrowly distributed to Gammaproteobacteria. Nucleic Acids Res. 41:3491–3503.

Gottesman S, Storz G. 2011. Bacterial small RNA regulators: versatile roles and rapidly evolving variations. Cold Spring Harb Perspect Biol. 3:a003798.

Güell M, et al. 2009. Transcriptome complexity in a genome-reduced bacterium. Science 326:1268–1271.

Güell M, Yus E, Lluch-Senar M, Serrano L. 2011. Bacterial transcriptomics: what is beyond the RNA horiz-ome? Nat Rev Microbiol 9:658–669.

Gueneau de Novoa P, Williams KP. 2004. The tmRNA website: reductive evolution of tmRNA in plastids and other endosymbionts. Nucleic Acids Res. 32:D104–D108.
Han K, et al. 2013. Extraordinary expansion of a *Sorangium cellulosum* genome from an alkaline milieu. Sci Rep. 3:2101.
Hansen AK, Degnan PH. 2014. Widespread expression of conserved small RNAs in small symbiont genomes. ISME J. 8:2490–2502.
Hoeppner MP, Gardner PP, Poole AM. 2012. Comparative analysis of RNA families reveals distinct repertoires for each domain of life. PLoS Comput Biol. 8:e1002752.
Hudson CM, Lau B, Williams KP. 2014. Ends of the line for tmRNA-SmpB. Front Microbiol. 5:421.
Husník F, Chrudimský T, Hypša V. 2011. Multiple origins of endosymbiosis within the Enterobacteriaceae (γ-Proteobacteria): convergence of complex phylogenetic approaches. BMC Biol. 9:87.
Kortmann J, Narberhaus F. 2012. Bacterial RNA thermometers: molecular zippers and switches. Nat Rev Microbiol. 10:255–265.
Kuo C-H, Moran NA, Ochman H. 2009. The consequences of genetic drift for bacterial genome complexity. Genome Res. 19:1450–1454.
Lamelas A, et al. 2011. *Serratia symbiotica* from the aphid *Cinara cedri*: a missing link from facultative to obligate insect endosymbiont. PLoS Genet. 7:e1002357.
Lindgreen S, et al. 2014. Robust identification of noncoding RNA from transcriptomes requires phylogenetically-informed sampling. PLoS Comput Biol. 10:e1003907.
Marszalkowski M, Willkomm DK, Hartmann RK. 2008. 5′-end maturation of tRNA in *Aquifex aeolicus*. Biol Chem. 389:395–403.
Matelska D, et al. 2013. S6:S18 ribosomal protein complex interacts with a structural motif present in its own mRNA. RNA 19:1341–1348.
McCutcheon JP, Moran NA. 2012. Extreme genome reduction in symbiotic bacteria. Nat Rev Microbiol. 10:13–26.
Merhej V, Royer-Carenzi M, Pontarotti P, Raoult D. 2009. Massive comparative genomic analysis reveals convergent evolution of specialized bacteria. Biol Direct. 4:13.
Mira A, Ochman H, Moran NA. 2001. Deletional bias and the evolution of bacterial genomes. Trends Genet. 17:589–596.
Molina N, van Nimwegen E. 2008. Universal patterns of purifying selection at noncoding positions in bacteria. Genome Res. 18:148–160.
Moran NA. 2002. Microbial minimalism: genome reduction in bacterial pathogens. Cell 108:583–586.
Moran NA, Dunbar HE, Wilcox JL. 2005. Regulation of transcription in a reduced bacterial genome: nutrient-provisioning genes of the obligate symbiont *Buchnera aphidicola*. J Bacteriol. 187:4229–4237.
Moran NA, Mira A. 2001. The process of genome shrinkage in the obligate symbiont *Buchnera aphidicola*. Genome Biol. 2: RESEARCH0054.
Nakabachi A, et al. 2006. The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. Science 314:267.
Nakabachi A, Ishikawa H. 1999. Provision of riboflavin to the host aphid, *Acyrthosiphon pisum*, by endosymbiotic bacteria, *Buchnera*. J Insect Physiol. 45:1–6.
Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29:2933–2935.
Ochman H, Moran NA. 2001. Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. Science 292:1096–1099.
Parter M, Kashtan N, Alon U. 2007. Environmental variability and modularity of bacterial metabolic networks. BMC Evol Biol. 7:169.
Qiu Y, et al. 2010. Structural and operational complexity of the *Geobacter sulfurreducens* genome. Genome Res. 20:1304–1311.
Ramachandran R, Burke AK, Cormier G, Jensen RV, Stevens AM. 2014. Transcriptome-based analysis of the *Pantoea stewartii* quorum-sensing regulon and identification of EsaR direct targets. Appl Environ Microbiol. 80:5790–5800.
Rivas E, Eddy SR. 2000. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. Bioinformatics 16:583–605.
Rosenblad MA, Larsen N, Samuelsson T, Zwieb C. 2009. Kinship in the SRP RNA family. RNA Biol. 6:508–516.
Sayers EW, et al. 2012. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 40:D13–D25.
Serganov A, Patel DJ. 2008. Towards deciphering the principles underlying an mRNA recognition code. Curr Opin Struct Biol. 18:120–129.
Sesto N, Wurtzel O, Archambaud C, Sorek R, Cossart P. 2013. The excludon: a new concept in bacterial antisense RNA-mediated gene regulation. Nat Rev Microbiol. 11:75–82.
Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H. 2000. Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. Nature 407:81–86.
Siegel RW, Banta AB, Haas ES, Brown JW, Pace NR. 1996. *Mycoplasma fermentans* simplifies our view of the catalytic core of ribonuclease P RNA. RNA 2:452–462.
Silva FJ, Latorre A, Moya A. 2001. Genome size reduction through multiple events of gene disintegration in *Buchnera* APS. Trends Genet. 17:615–618.
Storz G. 2002. An expanding universe of noncoding RNAs. Science 296:1260–1263.
Storz G, Vogel J, Wassarman KM. 2011. Regulation by small RNAs in bacteria: expanding frontiers. Mol Cell. 43:880–891.
Tatusov RL, Galperin MY, Natale DA, Koonin EV. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 28:33–36.
Tchufistova LS, Komarova AV, Boni IV. 2003. A key role for the mRNA leader structure in translational control of ribosomal protein S1 synthesis in gamma-proteobacteria. Nucleic Acids Res. 31:6996–7002.
Wallace DC, Morowitz HJ. 1973. Genome size and evolution. Chromosoma 40:121–126.
Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 10:57–63.
Waters LS, Storz G. 2009. Regulatory RNAs in bacteria. Cell 136:615–628.
Wernegreen JJ, Ochman H, Jones IB, Moran NA. 2000. Decoupling of genome size and sequence divergence in a symbiotic bacterium. J Bacteriol. 182:3867–3869.
Wilcox JL, Dunbar HE, Wolfinger RD, Moran NA. 2003. Consequences of reductive evolution for gene expression in an obligate endosymbiont. Mol Microbiol. 48:1491–1500.
Will S, Joshi T, Hofacker IL, Stadler PF, Backofen R. 2012. LocARNA-P: accurate boundary prediction and improved detection of structural RNAs. RNA 18:900–914.
Williams LE, Wernegreen JJ. 2010. Unprecedented loss of ammonia assimilation capability in a urease-encoding bacterial mutualist. BMC Genomics 11:687.
Willkomm DK, Feltens R, Hartmann RK. 2002. tRNA maturation in *Aquifex aeolicus*. Biochimie. 84:713–722.
Woese CR. 1987. Bacterial evolution. Microbiol Rev. 51:221–271.
Wolf YI, Koonin EV. 2012. A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. Genome Biol Evol. 4:1286–1294.
Wu M, Eisen JA. 2008. A simple, fast, and accurate method of phylogenomic inference. Genome Biol. 9:R151.
Wurtzel O, et al. 2012. Comparative transcriptomics of pathogenic and non-pathogenic *Listeria* species. Mol Syst Biol. 8:583.
Yao Z, Weinberg Z, Ruzzo WL. 2006. CMfinder—a covariance model based RNA motif finding algorithm. Bioinformatics 22:445–452.
Yus E, et al. 2009. Impact of genome reduction on bacterial metabolism and its regulation. Science 326:1263–1268..

Associate editor: Daniel Sloan