Genome Rearrangements Can Make and Break Small RNA Genes

Rahul Raghavan[1],*, Fenil R. Kacharia[1], Jess A. Millar[1], Christine D. Sislak[1], and Howard Ochman[2]

[1]Department of Biology and Center for Life in Extreme Environments, Portland State University

[2]Department of Integrative Biology, The University of Texas at Austin

*Corresponding author: E-mail: rahul.raghavan@pdx.edu.

Accepted: July 12, 2015

Data deposition: This project has been deposited at NCBI Sequence Read Archive (SRA) under the accession SRP044074 and SRP009089.

Abstract

Small RNAs (sRNAs) are short, transcribed regulatory elements that are typically encoded in the intergenic regions (IGRs) of bacterial genomes. Several sRNAs, first recognized in *Escherichia coli*, are conserved among enteric bacteria, but because of the regulatory roles of sRNAs, differences in sRNA repertoires might be responsible for features that differentiate closely related species. We scanned the *E. coli* MG1655 and *Salmonella enterica* Typhimurium genomes for nonsyntenic IGRs as a potential source of uncharacterized, species-specific sRNAs and found that genome rearrangements have reconfigured several IGRs causing the disruption and formation of sRNAs. Within an IGR that is present in *E. coli* but was disrupted in *Salmonella* by a translocation event is an sRNA that is associated with the FNR/CRP global regulators and influences *E. coli* biofilm formation. A *Salmonella*-specific sRNA evolved de novo through point mutations that generated a σ^{70} promoter sequence in an IGR that arose through genome rearrangement events. The differences in the sRNA pools among bacterial species have previously been ascribed to duplication, deletion, or horizontal acquisition. Here, we show that genomic rearrangements also contribute to this process by either disrupting sRNA-containing IGRs or creating IGRs in which novel sRNAs may evolve.

Key words: sRNA, *E. coli*, *Salmonella*, intergenic regions, gene origination.

Introduction

RNAs that do not code for proteins are critical to gene regulation in all domains of life. In bacteria, small RNAs (sRNAs) are typically 50–200 nucleotides (nt) in length and are usually encoded in genomic regions between protein-coding genes (intergenic regions or IGRs). They can control gene expression by modulating transcription, translation, or mRNA stability (Storz et al. 2011). The application of technologies that interrogate entire transcriptomes has revealed unexpectedly large numbers of sRNAs in bacterial genomes (Raghavan, Groisman, et al. 2011; Kroger et al. 2012). But unlike protein-coding genes, the mechanisms by which new sRNA genes arise and the forces that shape the sRNAs contents of genomes are not well understood (Gottesman and Storz 2011). Some sRNAs, such as 6S RNA (Wassarman and Storz 2000), are broadly conserved among bacteria, whereas several others are species- or even strain-specific (Gottesman and Storz 2011; Skippington and Ragan 2012). The sRNA transcriptomes of the enterics *Escherichia coli* and *Salmonella enterica* show substantial overlap; however, some of the orthologous IGRs display different patterns of expression and several sRNAs are present in only one of the species (Raghavan et al. 2012).

Differences in sRNA gene contents among bacteria can arise from lineage-specific loss or from the emergence of new sRNAs through duplication (Lenz et al. 2004; Wilderman et al. 2004) or horizontal acquisition (Pichon and Felden 2005; Sittka et al. 2008). An examination of the distribution of sRNAs within the *E. coli/Shigella* complex showed that the variation in the presence of known sRNAs was dominated by gene loss through deletions (Skippington and Ragan 2012). However, because this study focused only on those sRNAs that were originally characterized in a single strain of *E. coli*, it was biased toward the recognition of deletion events as it could not detect unique sRNAs in the genomes of other strains. Applying a broader phylogenetic perspective, homologs of a dual-function sRNA, SgrS, have been detected in distantly related Gammaproteobacteria (Horler and

Vanderpool 2009), and an exhaustive survey of sRNAs revealed that most *E. coli* sRNAs originated after *Enterobacteriales* split from other Gammaproteobacteria (Peer and Margalit 2014). This lineage-specific sRNA accumulation seems to be related to the evolution of the RNA-binding protein Hfq; however, the mechanisms by which new bacterial sRNAs emerge or are lost remain largely unknown.

In eukaryotes, there are cases where novel regulatory RNAs have evolved through gene duplication, by de novo origination from noncoding sequences, and from the degradation of protein-coding genes (Kaessmann 2010); but in bacteria, the mechanisms by which new regulatory RNAs arise are much less clear. Because new genes can form through the chimeric assembly of fragments from various sources—one well-known example of this is the *jingwei* gene of *Drosophila* (Long and Langley 1993)—we first adopted a structural genomics and RNA sequencing (RNA-seq)-based approach to identify new sRNA genes and then tested for sRNA functions. This combination of comparative and experimental analyses identified several previously unrecognized sRNAs and uncovered the sources of these differences in sRNA repertoires. We find that genome rearrangements have disrupted and formed IGRs containing functional sRNAs, thereby causing disparity in the sRNA contents of related bacterial species.

Materials and Methods

RNA Sequencing

For sRNA discovery, *E. coli* K-12 MG1655 (GenBank NC_000913.2) and *S. enterica* subsp. *enterica* serovar Typhimurium str. 14028S (GenBank NC_016856.1) were grown in lysogeny broth (LB) to $OD_{600} \approx 0.5$ and then harvested by centrifugation. Total RNA was extracted from bacterial pellets using TRI reagent (Life Technologies), and cleaned on RNeasy columns (Qiagen) to remove spurious transcripts, transfer RNAs and 5S ribosomal RNA (rRNA). Genomic DNA was degraded by DNase treatment (Life Technologies) and 16S and 23S rRNAs were removed with a MICROBExpress kit (Life Technologies). Strand-specific RNA-seq libraries were synthesized (Raghavan et al. 2012), and each library was sequenced on the Illumina GA II platform (35 cycles) at the Yale Center for Genome Analysis.

Mapping Sequencing Reads

To identify sRNAs, sequencing reads were mapped onto the published *E. coli* (NC_000913.2) or *Salmonella* Typhimurium (NC_016856.1) genomes using MAQ (Li et al. 2008) and examined with Artemis (Rutherford et al. 2000), as described previously (Raghavan, Groisman, et al. 2011; Raghavan, Sage, et al. 2011; Raghavan et al. 2012). Those previously uncharacterized sRNAs identified in *E. coli* are numbered and given the prefix EcsR (*E. coli* sRNA) and those in *Salmonella*, SesR (*S. enterica* sRNA).

To characterize regions that are differentially expression, sequencing reads were mapped onto *E. coli* (NC_000913.2) using the CLC Genomics Workbench. Genes with at least one read per sample and at least 20 total reads across all samples were chosen based on raw gene read counts from CLC mapping. Differential expression analysis of genes was performed using the DESeq R package (Anders and Huber 2010). Genes were chosen for downstream analysis based on significance ($P < 0.05$, FDR-corrected). Gene Ontology (GO) terms were found using Database for Annotation, Visualization and Integrated Discovery (DAVID) and the GO FAT filter (Huang et al. 2009). GO-term enrichment tests were also performed with DAVID. GO-terms overrepresented among differentially expressed genes were chosen based on the level of statistical significance ($P < 0.05$, Benjamini-corrected).

sRNA Target Identification

To identify sRNA-regulated genes, EcsR1 was cloned into the *Nhe*I and *Hin*dIII sites behind the arabinose-inducible promoter on plasmid pBAD using the polymerase chain reaction (PCR) primers 5′-CCG CTA GCG TTT TAG TAT CCG CAT AAA GTG TAA C-3′ and 5′-CTA AGC TTT CCT GCC CGC TGT TAT GGC G-3′. *Escherichia coli* or *Salmonella*, transformed with either the empty pBAD vector (control) or pBAD+ EcsR1 (test), were grown in LB to $OD_{600} \approx 0.5$ and induced with 0.2% arabinose for 15 min, as previously described (Durand and Storz 2010). RNA was extracted and processed for Illumina sequencing as above. Four Illumina mRNA-seq libraries (two control samples and two test samples) were prepared for each bacterium and multiplexed into a single lane of an Illumina HiSeq 2000 (101 cycles) at the Genomic Sequencing and Analysis Facility at University of Texas at Austin.

Measuring Hfq Stabilization of sRNAs

An *Hfq*-deleted strain of *E. coli* (JW4130-1) and its isogenic parent strain (BW25113) (Baba et al. 2006) were obtained from Yale Coli Genetic Stock Center and grown to mid-log phase ($OD_{600} \approx 0.5$) in LB. Total RNA was DNase-treated, and 1 μg used as template for preparing cDNA. The abundances of EcsR1 in the wild-type and *Hfq*-deleted strains were determined by quantitative PCR (primers: 5′-TTT TTG TGT AAT GAC GGA GTT CA-3′, and 5′-GCG GGC TTT TTC TGC TTA TT-3′), and calculated from Ct (threshold cycle) values.

Identification of Unique IGRs

Orthologous genes common to *E. coli* and *Salmonella* were identified using a reciprocal BLAST best-hit approach (Raghavan et al. 2012). Gene order of each orthologous gene-pair was determined with GeneOrder 4.0 (Mahadevan and Seto 2010), and in cases where the genomic locations were not syntenic in the two species, we searched for gene-pairs with adjacent novel IGRs using Artemis.

Identification of -10 Promoter Elements and sRNA Homologs

Transcriptional start sites (TSSs) for novel sRNAs were identified from RNA-seq data as described previously (Raghavan, Sage, et al. 2011; Raghavan et al. 2012). TSSs for flanking genes were identified as above and were confirmed using published data (Kroger et al. 2012; Keseler et al. 2013). The σ^{70} -10 motif has a 6-bp consensus sequence TATAAT; however, promoters often have imperfect matches to the consensus and can be located anywhere in a window ranging from approximately 4 to 18 bp upstream of the TSS (Huerta and Collado-Vides et al. 2003). To identify potential -10 elements associated with sRNAs, we searched this 15-bp window for any hexamers that matched at least 4 of the 6 bp in the consensus sequence including the two most highly conserved positions, A2 and T6 (Huerta et al. 2006). Bacterial genomes were queried for homologs of sRNAs identified in this study by analyzing a combination of sequence identity, secondary structure conservation, and genomic location as described previously (Raghavan, Groisman, et al. 2011).

Detection of sRNA 3′-Ends

A modified Rapid Amplification of cDNA Ends (RACE) procedure (Raghavan, Groisman, et al. 2011) was used to determine the 3′-ends of sRNAs as follows: Total RNA, depleted of 16S and 23S rRNA using a MICROBExpress kit (Life Technologies), was dephosphorylated with alkaline phosphatase (NEB), and a short oligonucleotide adapter (5′-P-UCG UAU GCC GUC UUC UGC UUG UidT-3′) was ligated to 3′-ends using T4 RNA ligase (NEB). The 3′ adapter-ligated RNA was reverse-transcribed using a primer complementary to the adapter (5′-CAA GCA GAA GAC GGC ATA CGA-3′), and the resulting cDNA was used as template in PCR reactions using primers specific to sRNAs (EcsR1: 5′-AGA TGA CAC TTT TTG TGT AAT GAC G-3′; EcsR2: 5′-TAT CGC GCT ACT TCA GGA TGA TGT A-3′) along with the adapter-complementary primer. Amplicons were resolved on 3% low-range ultra agarose (Bio-Rad) gels to determine their lengths, and their nucleotide sequences were determined by Sanger sequencing.

Biofilm Assay

EcsR1-deletion strain of *E. coli* was constructed using λ Red-mediated recombination (Datsenko and Wanner 2000). *Escherichia coli* or *Salmonella* strains grown overnight at 37 °C in LB (or LB with 100 μg/ml ampicillin) were diluted 1:100 in fresh media and grown in 96-well microtiter plates for 48 h at 28 °C without shaking. Planktonic growth (OD_{600}) of *E. coli* and *Salmonella* strains measured on a Victor X5microplate reader (Perkin Elmer) did not significantly differ from each other. Supernatants containing nonadhered cells were discarded, and samples were washed twice with distilled water and the attached biofilm in each well was stained with 0.1% crystal violet for 30 min. Unbound stain was removed by washing with distilled water. To quantify biofilm production, the crystal violet associated with biofilms was dissolved in 100% ethanol and absorbance (A_{600}) was measured, and normalized to the OD_{600} value of each strain, as described previously (Gualdi et al. 2008). Average intensity of biofilm formation for each strain was generated from at least four replicate experiments.

Results

Genome Rearrangements Form Unique IGRs

To identify IGRs that are unique to either *E. coli* or *Salmonella*, we compared the genomic locations of all orthologous genes in the two genomes. Because *E. coli* and *Salmonella* genomes are largely syntenic, the majority of IGRs situated between orthologous gene-pairs in the two genomes are also syntenic. However, there are several instances where orthologous protein-coding genes are situated at different relative locations in each genome (apparent as data points that do not lie along the diagonal in supplementary fig. S1, Supplementary Material online). After examining each of these cases, we identified chimeric IGRs present in either *E. coli* or *Salmonella* that were generated through the rearrangement of 68 genes (supplementary table S1, Supplementary Material online).

Unique IGRs Contain Novel sRNAs

We performed a directional RNA-seq analysis on *E. coli* and *Salmonella* grown under identical conditions to determine whether any of the species-specific IGRs contained highly transcribed regions. After mapping sequencing reads onto the respective genomes, we detected "transcriptional peaks," which usually indicate the presence of sRNAs, in four of the species-specific IGRs, two in *E. coli* and two in *Salmonella* (fig. 1). Transcripts mapping to the corresponding locations in the *E. coli* and *Salmonella* genomes have been observed in previous studies (Tjaden et al. 2002; Dornenburg et al. 2010; Kroger et al. 2013) further verifying their transcriptional status, and there were no potential open reading frames (ORFs) of substantial length within these transcripts indicating that they represent sRNAs.

TSSs and 3′-ends of the transcribed sequences detected in these IGRs were identified from RNA-seq data (Raghavan et al. 2012), and a modified 3′-RACE procedure (Raghavan, Groisman, et al. 2011) was used to confirm the sRNAs in *E. coli* (supplementary fig. S2, Supplementary Material online), yielding the following results: 1) The sRNA (EcsR1) within the IGR between *uspF* and *ompN* genes in *E. coli* is 126 nt (genomic location 1433654–1433779), 2) the sRNA (EcsR2) within the IGR between *yagU* and *ykgJ* genes in *E. coli* is 166 nt (genomic location 302905–303070), 3) the sRNA (SesR1) within the IGR between STM14_1512 and

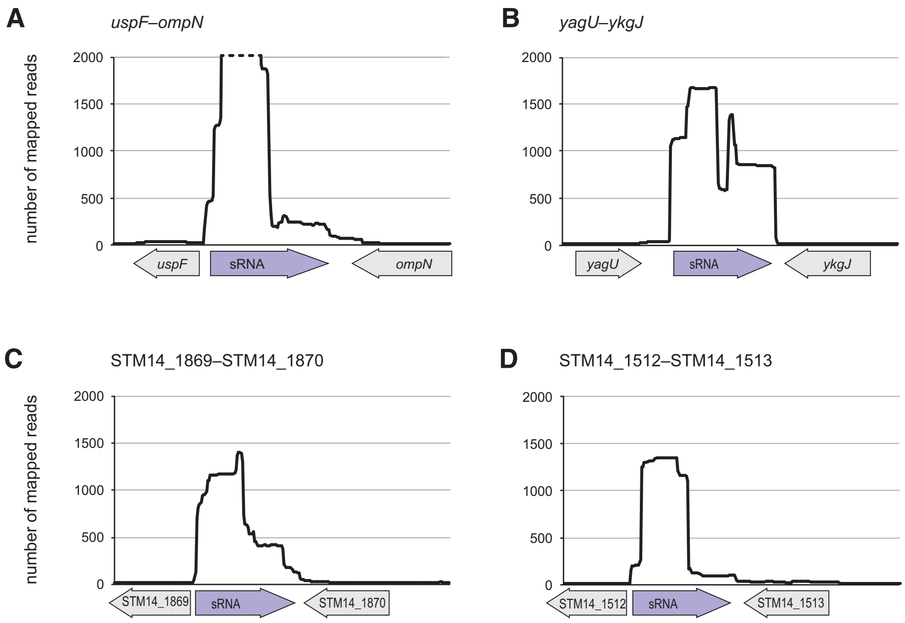

Fig. 1.—Expression profiles within nonsyntenic IGRs. Putative sRNAs were detected by RNA-seq analysis of transcript levels within IGRs in *E. coli* (*A*, *B*) and *Salmonella* (*C*, *D*). For uniformity, the number of sequencing reads mapped to IGRs is limited to 2,000 (dashed line). Arrows showing the orientation of ORFs and putative sRNAs are not drawn to scale.

STM14_1513 genes in *Salmonella* is 105 nt (genomic location 1347963–1348067), and 4) the sRNA (SesR2) within the IGR between STM14_1869 and STM14_1870 genes in *Salmonella* is 111 nt (genomic location 1636380–1636490).

Homologs of EcsR1 and EcsR2 are present in all 66 *E. coli* genomes available in the RefSeq database (supplementary table S2, Supplementary Material online). SesR1 homologs were detected in all 44 *S. enterica* and *S. bongori* genomes (supplementary table S3, Supplementary Material online), and a recent study reported an sRNA (STnc1990) at the homologous position in *S. enterica* Typhimurium SL1344 (Kroger et al. 2013). SesR2 is conserved in 20 *S. enterica* genomes and in the two sequenced *S. bongori* strains (supplementary table S3, Supplementary Material online). However, the STM14_1869–STM14_1870 IGR is not maintained in *S. bongori* due to the loss of the STM14_1870 ortholog. Because SesR2 is absent from a few *S. enterica* serovars but present in the *S. bongori* outgroup, this sRNA is ancestral to *Salmonella* and was subsequently lost in some *S. enterica* lineages.

A *Salmonella*-specific IGR Formed Through HGT-mediated Genome Rearrangement

The STM14_1869–STM14_1870 IGR is present in *Salmonella* but not in *E. coli*. *Escherichia coli* possesses a gene, *yjgH*, that is orthologous to STM14_1869, but contained no ortholog for STM14_1870. Further analysis uncovered that STM14_1870 and its neighboring gene STM14_1871 constitute the toxin and antitoxin, respectively, of the StbED toxin–antitoxin (TA) system (Unterholzner et al. 2013) (fig. 2).

Orthologs of *stbED* TA genes are present on several enterobacterial plasmids and prophages, and are horizontally transferred between bacteria (Anantharaman and Aravind 2003; Unterholzner et al. 2013). Additionally, the succeeding gene in the *Salmonella* genome, STM14_1872, is also homologous to a gene of bacteriophage origin (fig. 2), further indicating that the IGR between STM14_1869 and STM14_18670 was created by the introduction of genes through horizontal gene transfer (HGT)-mediated events.

Evolution of a New sRNA in a *Salmonella* IGR

To determine whether SesR2 was introduced along with its horizontally acquired neighboring genes into *Salmonella*, we searched the IGRs downstream of the *stbE* gene in several enterobacterial genomes, but could not detect homologous sRNAs (fig. 2). Because a promoter is required for the new sRNA to be transcribed, we compared the region that contains the sRNA's TSS (5′-end of STM14_1869) with its homologous sequence in *E. coli* (5′-end of *yjgH*) and in other enterics. As shown in figure 3 and supplementary figure S3, Supplementary Material online, a putative σ^{70} promoter (CA TAAT, located -6 to -11 bp from sRNA's TSS) is uniquely present in *Salmonella*, indicating that this sRNA originated de novo in the *Salmonella*-specific IGR. Additionally, the promoter and sRNA sequences are conserved in *S. bongori* and in those *S. enterica* serovars that maintain an intact STM14_1869–STM14_1870 IGR.

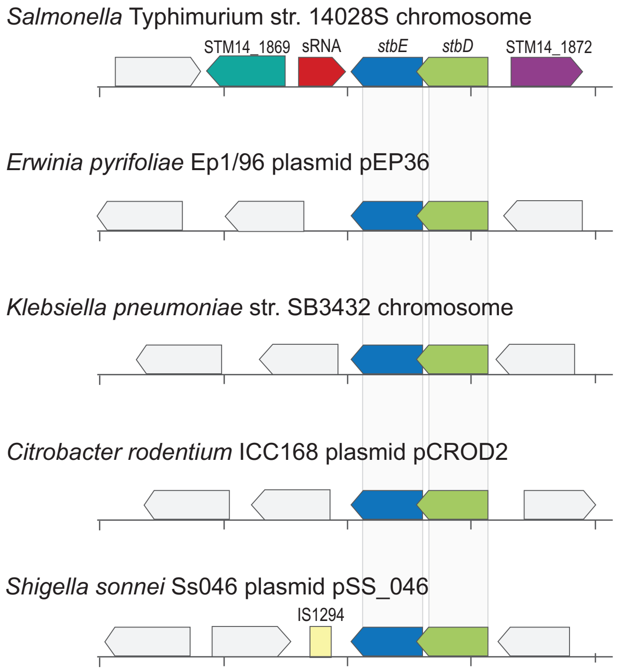
Salmonella Typhimurium str. 14028S chromosome
STM14_1869
sRNA
stbE
stbD
STM14_1872
Erwinia pyrifoliae Ep1/96 plasmid pEP36
Klebsiella pneumoniae str. SB3432 chromosome
Citrobacter rodentium ICC168 plasmid pCROD2
Shigella sonnei Ss046 plasmid pSS_046
IS1294

Fig. 2.—*Salmonella* IGR formed through an HGT-mediated genome rearrangement. Most homologs of STM14_1870 (*stbE*, blue arrow) and STM14_1871 (*stbD,* green arrow) are situated on bacterial plasmids, and STM14_1872 (purple arrow) is a prophage gene. Both the STM14_1869–STM14_1870 IGR and the sRNA (SesR2) are present only in *Salmonella*. In *Shigella*, an insertion sequence (IS1294) flanks the *stbE* gene.

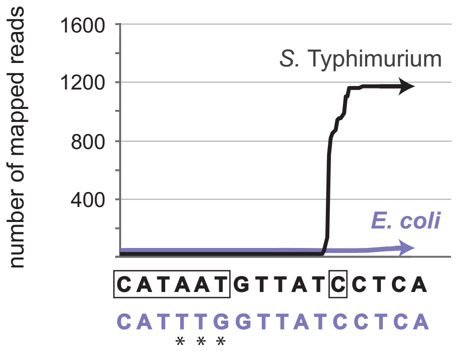

Fig. 3.—Evolution of a new sRNA promoter. Sequences immediately upstream of STM14_1869 (4471230–4471245) and its ortholog *yjgH* in *E. coli* (1636369–1636384) are aligned. Numbers of RNA-seq reads mapping to this region are shown (black, *Salmonella*; blue, *E. coli*). The new *Salmonella* σ^{70} promoter and sRNA (SesR2) transcription start site are boxed. Asterisks indicate point mutations that differentiate the *Salmonella* sequence from the corresponding region in *E. coli*.

Rearrangement-Induced Loss of a *Salmonella* IGR

In *E. coli,* the *ompN* and *uspF* genes are adjacent, separated by a 140-bp IGR that contains EcsR1; whereas in *Salmonella,* the *ompN* gene is in an alternate location, situated between the STM14_1775 and *rstA* genes. To determine the ancestry of these gene arrangements—specifically whether the *uspF–ompN* IGR was gained by *E. coli* or lost by *Salmonella*—we analyzed the organization of the orthologous regions in the genomes of other enteric bacteria. The *uspF–ompN* IGR is present and intact in other enteric species (*Klebsiella pneumoniae*, *Enterobacter aerogenes,* and *Citrobacter koseri*) establishing that this IGR predates the split between *E. coli* and *Salmonella* and was lost in *Salmonella* due to the relocation of *ompN* gene (figs. 4 and 5). As a consequence of this genome rearrangement in *Salmonella*, EcsR1 was split into two fragments located $\approx$200 kb apart, neither of which is transcribed (supplementary fig. S4, Supplementary Material online).

EcsR1 Is Associated with Global Regulators in *E. coli*

The *uspF–ompN* IGR and EcsR1 are present in all strains of *E. coli*, which suggests that it maintains a regulatory function. To identify genes that are potentially under the control of EcsR1, we examined the effect of its overexpression on *E. coli* genes genome-wide, an approach that has been used previously to characterize the regulatory targets of sRNAs (Durand and Storz 2010; Beisel and Storz 2011). When analyzed by RNA-seq, the expression levels of 43 genes were significantly different ($P < 0.05$) in the EcsR1-overexpressing strain when compared with wild-type *E. coli* (supplementary table S4, Supplementary Material online). A GO analysis uncovered bacterial membrane (GO:0031090), carbohydrate catabolic process (GO:0016052), and nitrate metabolic process (GO:0042126) as processes that were significantly enriched ($P < 0.05$) in our data set. Eleven downregulated genes were associated with these GO terms, of which nine were regulated by CRP and/or FNR (supplementary table S5, Supplementary Material online) (Constantinidou et al. 2006; Keseler et al. 2013). In concert with these observations, a 22-nt palindromic sequence with features resembling the consensus CRP-binding site and a putative 15-nt FNR-binding sequence (fig. 5) were identified upstream of EcsR1, indicative of the sRNA being part of the CRP and FNR regulons. Expression of another *E. coli* sRNA, FnrS, is known to be affected by both FNR and CRP (Durand and Storz 2010), showing that CRP and FNR regulons overlap and may control multiple sRNAs. It has been shown previously that the transcriptional regulator CRP can control the expression of both an sRNA and the sRNA's target genes, and this "feed-forward loop" is thought to aid in the efficient modulation of gene expression in *E. coli* (Beisel and Storz 2011).

Because many sRNAs in *E. coli* associate with and are stabilized by the RNA-binding protein Hfq (De Lay et al. 2013), we examined whether Hfq stabilizes EcsR1. We measured the abundance of EcsR1 in wild-type and Hfq-deficient strains of *E. coli,* and found it to be significantly more abundant in the

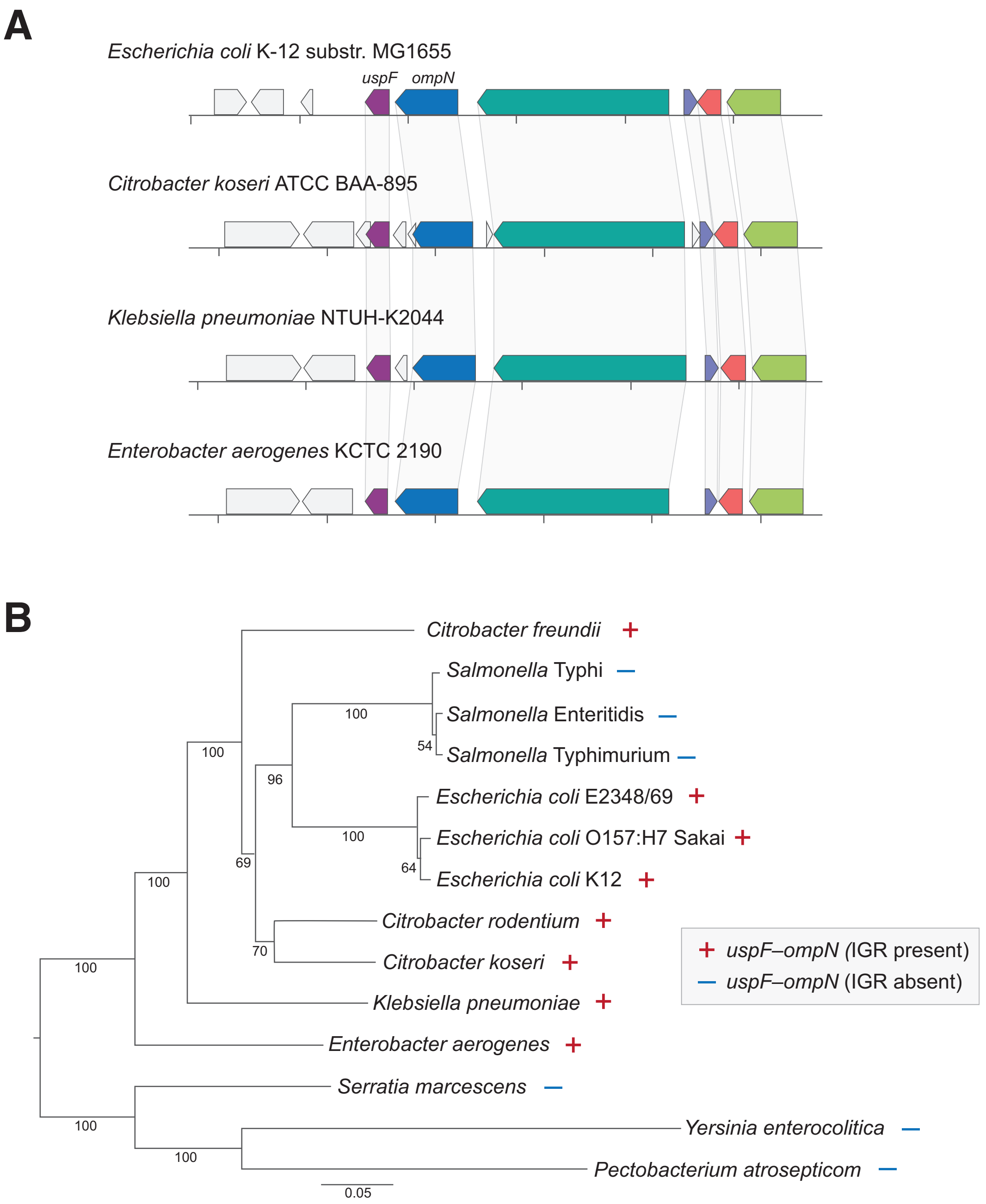

Fig. 4.—Distribution of the *uspF–ompN* IGR among enteric species. (*A*) Alignment of genomic regions containing the *uspF–ompN* IGR in *E. coli* and three other enteric species. Note that in both *Citrobacter koseri* and *Klebsiella pneumoniae*, small ORFs (gray arrows situated between *uspF* [purple] and *ompN* [blue]) have been predicted to occur in this IGR. (*B*) Phylogenetic tree (modified from Petty et al. 2010) showing the presence or absence of the *uspF–ompN* IGR among species.

wild-type strain (supplementary fig. S5, Supplementary Material online), reinforcing its identity as an sRNA.

EcsR1 Impacts Biofilm Formation

Because CRP and FNR control carbohydrate and nitrate metabolism during biofilm formation (Van Alst et al. 2007; Karatan and Watnick 2009), we constructed EcsR1-deletion and EcsR1-overexpression *E. coli* strains and measured the impact of this sRNA on biofilm formation. As shown in figure 6, biofilm production increased significantly ($P < 0.0001$) in the EcsR1-deleted strain when compared with wild-type *E. coli*. Reintroduction of a plasmid-borne copy of EcsR1 into the deletion strain reduced biofilm

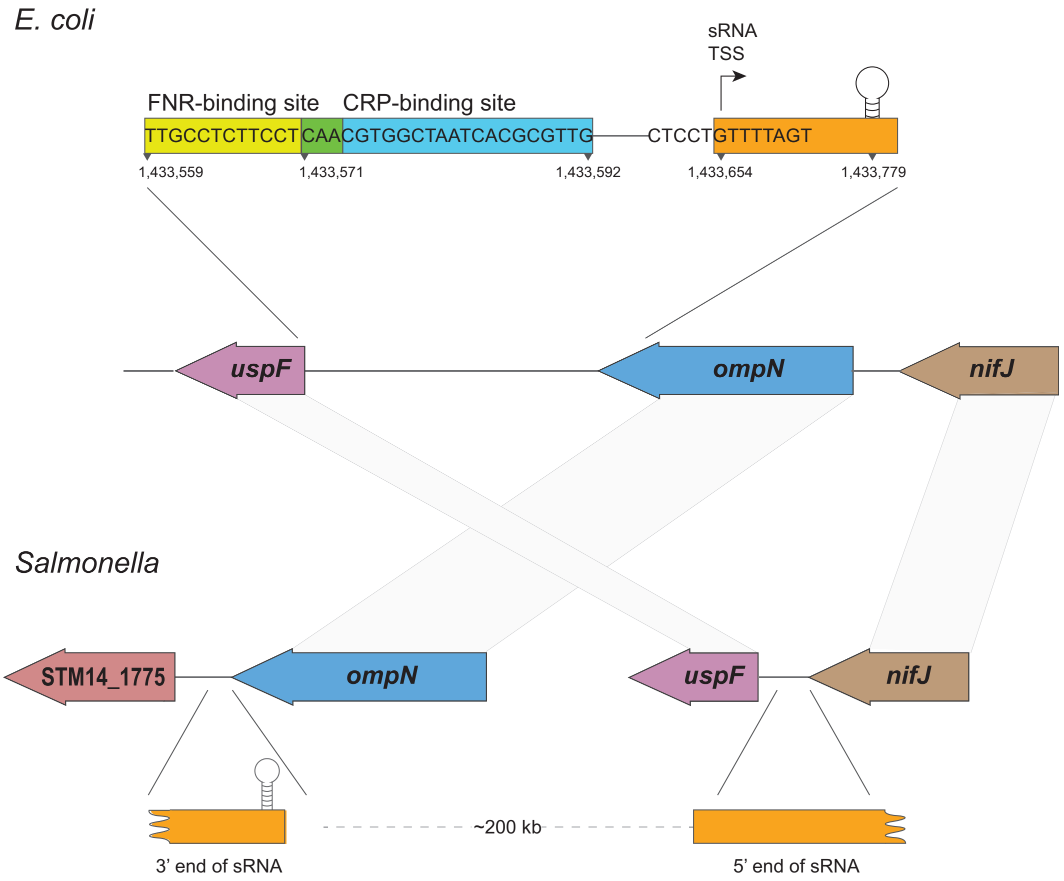

Fig. 5.—Loss of *uspF–ompN* IGR through genome rearrangement. The *uspF–ompN* IGR of *E. coli* was fragmented in *Salmonella* due to the translocation of *ompN* to a site adjacent to STM14_1775. The predicted FNR- and CRP-binding sites (yellow and blue, respectively; overlapping region in green) upstream of the sRNA (EcsR1) transcription start site (sRNA TSS) are shown. A predicted Rho-independent terminator (stem-loop structure) situated 3′ of the sRNA is also depicted.

formation to the same level as that of the wild-type strain (fig. 6), indicating that biofilm-inhibition is an sRNA-specific phenotype.

Expression of *E. coli* EcsR1 in *Salmonella* Activates Invasion-Associated Genes

Biofilm production is important to virulence of enteric pathogens, so we tested the effects of EcsR1 on biofilm production in *Salmonella* by reintroducing the sRNA in an expression vector. There was no significant difference in biofilm production between the wild-type and EcsR1-overexpression strains; the overexpression of EcsR1 in *Salmonella* alters the expression of 128 genes genome-wide (supplementary table S6, Supplementary Material online). GO analysis revealed nine processes (representing 27 genes) that were significantly enriched within this gene set, with "pathogenesis" (GO:0009405) being the most highly significant (supplementary table S7, Supplementary Material online). Among genes regulated by this sRNA, 22 are known to promote *Salmonella* invasion of host cells, most of which are situated within SPI-1 pathogenicity island (Fàbrega and Vila 2013).

Discussion

Our search for species-specific sRNAs was directed toward IGRs that were unique to either *E. coli* or *Salmonella* because most bacterial regulatory sRNAs are contained within these noncoding regions, although 3′-untranslated regions (UTRs) of mRNAs and promoters within mRNAs can also give rise to sRNAs (Chao et al. 2012; Guo et al. 2014). We found that genome rearrangements have altered IGRs and, in doing so, caused disparity in sRNA contents of these two species. A newly discovered sRNA (EcsR1), situated within the IGR between the *uspF–ompN* genes in *E. coli*, is absent from *Salmonella* due to the translocation of a genomic segment containing the *ompN* gene. This sRNA is associated with the FNR and CRP regulons, and its expression impacts *E. coli* biofilm formation. Additionally, we identified an sRNA (SesR2) unique to *Salmonella* that evolved de novo in an IGR that

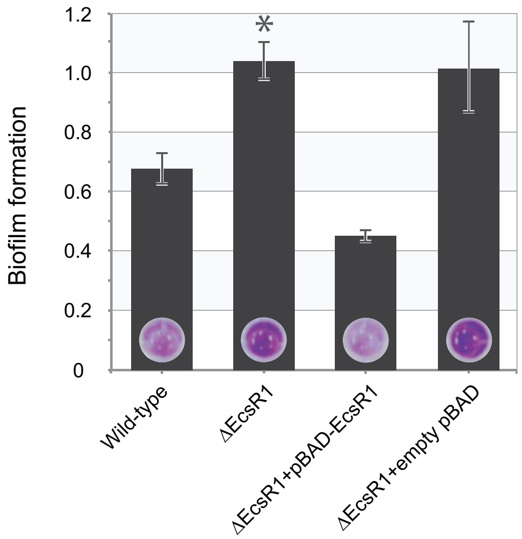

Fig. 6.—Biofilm formation is influenced by EcsR1. *Escherichia coli* biofilms stained with crystal violet were measured (A_{600}) after 48-h growth at 28 °C and normalized to OD_{600} value. A wild-type strain, an EcsR1-deleted strain (ΔEcsR1), a ΔEcsR1 strain containing pBAD with cloned EcsR1 (ΔEcsR1-pBAD-EcsR1), and a ΔEcsR1 strain containing empty pBAD (ΔEcsR1-empty pBAD) were tested. Asterisks indicate a statistically significant difference between wild-type and ΔEcsR1 strains ($P < 0.0001$).

was formed through a phage-mediated genome rearrangement. Although disparities in genome architectures are common among related species, these are the first known cases where rearrangements have caused the generation and destruction of sRNAs.

The main source of rearrangement events in bacterial genomes is homologous recombination across identical sequences. *Escherichia coli* and *Salmonella* contain numerous classes of repeat elements that can serve as templates for exchange (Rocha 2004). In addition, recombination between bacteriophage sequences in a genome can result in altered genome architectures in related bacteria (Brüssow et al. 2004); large proportions of both *E. coli* and *Salmonella* genomes consist of prophage genes (Bobay et al. 2013). Notwithstanding the large number of targets for homologous exchange, the gene order of *E. coli* and *Salmonella* has been well conserved despite an estimated 100-Myr divergence (Ochman et al. 1999). The major difference in their genome architectures involves a large 600-kb inversion spanning the replication terminus and approximately 50 small-scale translocation events (supplementary fig. S1, Supplementary Material online). This contrasts the situation in many bacteria, such as *Yersinia* and *Portiera* (Parkhill et al. 2001; Sloan and Moran 2013), in which there have been substantial changes in gene arrangement among closely related strains. The source of this variation has been ascribed not only to the differences among species in their repertoires of DNA recombination and repair enzymes (Tamas et al. 2002) but also to selection on gene order and position (Suyama and Bork 2001; Ballouz et al. 2010; Treangen and Rocha 2011). Our analyses show that some fraction of the rearrangements that shuffle IGRs may affect organismal fitness by disrupting or generating regulatory elements.

Although the IGR between *yagU* and *ompN* was disrupted and split in the *Salmonella* genome, some portions of it—approximately 70 nt of the 5′-end of EcsR1 and 20 nt of its 3′-end—are still recognizable adjacent to the *uspF* and *ompN* genes in *Salmonella* (fig. 5). It is likely that these sRNA segments are not transcribed and are not functional in *Salmonella* (supplementary fig. S4, Supplementary Material online, and Kroger et al. 2013) because nucleotide substitutions in the putative CRP- and FNR-binding regions (supplementary fig. S6, Supplementary Material online) have rendered them inactive. The reintroduction of EcsR1 into *Salmonella* did not affect biofilm production but instead triggered the increased expression of several virulence genes, particularly those within the SPI-1 pathogenicity island (supplementary table S7, Supplementary Material online). Multiple factors, including the biofilm machinery, are known to regulate the expression of these invasion-associated genes (Fàbrega and Vila 2013), suggestive of links between the different phenotypes produced by this sRNA in *E. coli* and *Salmonella*. Additional experiments are necessary to understand how EcsR1 induces diverse phenotypes in the two species; nevertheless, our findings demonstrate the potential of sRNAs to influence bacterial adaptation and evolution.

In addition to losing the biofilm-reducing sRNA (EcsR1), *Salmonella* has gained, again by a rearrangement event, an IGR that contains an sRNA (SesR2) that is not present in other enteric species. Because none of the corresponding regions in *E. coli* displays any appreciable transcript production, the evolution of this new sRNA in *Salmonella* also required the de novo formation of a new promoter sequence (fig. 3 and supplementary fig. S3, Supplementary Material online). In bacterial genomes, σ^{70} promoter-like sequences are able to arise spontaneously through point mutations, especially in IGRs (Stone and Wray 2001; Huerta et al. 2006; Mendoza-Vargas et al. 2009), and transcription can originate from newly evolved σ^{70} promoters (Mendoza-Vargas et al. 2009; Raghavan et al. 2012). Therefore, it is most likely that an incipient promoter in the newly formed STM14_1869–STM14_1870 IGR gave rise to the transcript that evolved into SesR2. An alternate possibility is that SesR2 was introduced into *Salmonella* with the HGT event that brought in the entire STM14_1870–STM14_1872 region, as has been proposed for other sRNAs located close to transposon insertion sites in *Salmonella* (Sittka et al. 2008). However, no similar sRNA is detectable in the homologous regions found on various enteric plasmids and genomes, making this scenario less likely. Finally, because the first 55 nt of this sRNA is complementary to the 5′-UTR of STM14_1869 (TSS of STM14_1869 is located 63 bp upstream of coding region), it could be functioning as an antisense RNA to regulate STM14_1869 expression, as shown previously for other genes in *Salmonella* (Lee and Groisman 2010).

In bacteria, differences in protein-coding gene contents between closely related species are either due to new genes that arose by gene duplication or HGT (Lerat et al. 2005; Blount et al. 2012; Nasvall et al. 2012), or due to gene loss through pseudogenization and deletion (Mira et al. 2001; Kuo and Ochman 2010). Although the mechanisms that shape bacterial sRNA gene repertoires are not well understood, duplication, deletion, and HGT have also been attributed to this process (Gottesman and Storz 2011). In this report, we show that genome rearrangements that create and disrupt IGRs can result in the gain or loss of sRNA genes in bacteria. Because sRNAs regulate myriad metabolic processes, this disparity in sRNA repertoires between closely related bacteria might also contribute to niche adaptation and speciation events.

Supplementary Material

Supplementary figures S1–S6 and tables S1–S7 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank Cindy Barlan, Daniel Phillips, Todd Smith and Abraham Moses for technical help, and Kim Hammond for assistance with figures. This work was supported by Portland State University and by grants from the National Institute of Health (GM74738) and the Templeton Foundation (23536) to H.O.

Literature Cited

Anantharaman V, Aravind L. 2003. New connections in the prokaryotic toxin-antitoxin network: relationship with the eukaryotic nonsense-mediated RNA decay system. Genome Biol. 4:R81.

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. Genome Biol. 11:R106.

Baba T, et al. 2006. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol. 2: 1216–1226.

Ballouz S, Francis AR, Lan R, Tanaka MM. 2010. Conditions for the evolution of gene clusters in bacterial genomes. PLoS Comput Biol. 6:e1000672.

Beisel CL, Storz G. 2011. The base-pairing RNA spot 42 participates in a multioutput feedforward loop to help enact catabolite repression in *Escherichia coli*. Mol Cell. 41:286–297.

Blount ZD, Barrick JE, Davidson CJ, Lenski RE. 2012. Genomic analysis of a key innovation in an experimental *Escherichia coli* population. Nature 489:513–518.

Bobay LM, Rocha EP, Touchon M. 2013. The adaptation of temperate bacteriophages to their host genomes. Mol Biol Evol. 30:737–751.

Brüssow H, Canchaya C, Hardt WD. 2004. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. Microbiol Mol Biol Rev. 68:560–602.

Chao Y, Papenfort K, Reinhardt R, Sharma CM, Vogel J. 2012. An atlas of Hfq-bound transcripts reveals 3′ UTRs as a genomic reservoir of regulatory small RNAs. EMBO J. 31:4005–4019.

Constantinidou C, et al. 2006. A reassessment of the FNR regulon and transcriptomic analysis of the effects of nitrate, nitrite, NarXL, and NarQP as *Escherichia coli* K12 adapts from aerobic to anaerobic growth. J Biol Chem. 281:4802–4815.

Datsenko KA, Wanner B. 2000. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. Proc Natl Acad Sci U S A. 97:6640–6645.

De Lay N, Schu DJ, Gottesman S. 2013. Bacterial small RNA-based negative regulation: Hfq and its accomplices. J Biol Chem. 288:7996–8003.

Dornenburg JE, Devita AM, Palumbo MJ, Wade JT. 2010. Widespread antisense transcription in *Escherichia coli*. mBio 1:e00024–10.

Durand S, Storz G. 2010. Reprogramming of anaerobic metabolism by the FnrS small RNA. Mol Microbiol. 75:1215–1231.

Fàbrega A, Vila J. 2013. *Salmonella enterica* serovar Typhimurium skills to succeed in the host: virulence and regulation. Clin Microbiol Rev. 26: 308–341.

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

Gualdi L, et al. 2008. Cellulose modulates biofilm formation by counteracting curli-mediated colonization of solid surfaces in *Escherichia coli*. Microbiology 154:2017–2024.

Guo MS, et al. 2014. MicL, a new σE-dependent sRNA, combats envelope stress by repressing synthesis of Lpp, the major outer membrane lipoprotein. Genes Dev. 28:1620–1634.

Horler RS, Vanderpool CK. 2009. Homologs of the small RNA SgrS are broadly distributed in enteric bacteria but have diverged in size and sequence. Nucleic Acids Res. 37:5465–5476.

Huang DW, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nat Protoc. 4:44–57.

Huerta AM, Collado-Vides J. 2003. Sigma70 promoters in *Escherichia coli*: specific transcription in dense regions of overlapping promoter-like signals. J Mol Biol. 333:261–278.

Huerta AM, Francino MP, Morett E, Collado-Vides J. 2006. Selection for unequal densities of sigma70 promoter-like signals in different regions of large bacterial genomes. PLoS Genet. 2:e185.

Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. Genome Res. 20:1313–1326.

Karatan E, Watnick P. 2009. Signals, regulatory networks, and materials that build and break bacterial biofilms. Microbiol Mol Biol Rev. 73: 310–347.

Keseler IM, et al. 2013. EcoCyc: fusing model organism databases with systems biology. Nucleic Acids Res. 41:D605–D612.

Kroger C, et al. 2012. The transcriptional landscape and small RNAs of *Salmonella enterica* serovar Typhimurium. Proc Natl Acad Sci U S A. 109:E1277–E1286.

Kroger C, et al. 2013. An infection-relevant transcriptomic compendium for *Salmonella enterica* serovar Typhimurium. Cell Host Microbe. 14: 683–695.

Kuo CH, Ochman H. 2010. The extinction dynamics of bacterial pseudogenes. PLoS Genet. 6:e1001050.

Lee EJ, Groisman EA. 2010. An antisense RNA that governs the expression kinetics of a multifunctional virulence gene. Mol Microbiol. 76: 1020–1033.

Lenz DH, et al. 2004. The small RNA chaperone Hfq and multiple small RNAs control quorum sensing in *Vibrio harveyi* and *Vibrio cholerae*. Cell 118:69–82.

Lerat E, Daubin V, Ochman H, Moran NA. 2005. Evolutionary origins of genomic repertoires in bacteria. PLoS Biol. 3:e130.

Li H, Ruan J, Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 18: 1851–1858.

Long M, Langley CH. 1993. Natural selection and the origin of jingwei, a chimeric processed functional gene in *Drosophila*. Science 260:91–95.

Mahadevan P, Seto D. 2010. Rapid pair-wise synteny analysis of large bacterial genomes using web-based GeneOrder4.0. BMC Res Notes. 3:41.

Mendoza-Vargas A, et al. 2009. Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in *E. coli*. PLoS One 4:e7526.

Mira A, Ochman H, Moran NA. 2001. Deletional bias and the evolution of bacterial genomes. Trends Genet. 17:589–596.

Nasvall J, Sun L, Roth JR, Andersson DI. 2012. Real-time evolution of new genes by innovation, amplification, and divergence. Science 338: 384–387.

Ochman H, Elwyn S, Moran NA. 1999. Calibrating bacterial evolution. Proc Natl Acad Sci U S A. 96:12638–12643.

Parkhill J, et al. 2001. Genome sequence of *Yersinia pestis*, the causative agent of plague. Nature 413:523–527.

Peer A, Margalit H. 2014. Evolutionary patterns of *Escherichia coli* small RNAs and their regulatory interactions. RNA 20:994–1003.

Petty NK, et al. 2010. The *Citrobacter rodentium* genome sequence reveals convergent evolution with human pathogenic *Escherichia coli*. J Bacteriol. 192:525–538.

Pichon C, Felden B. 2005. Small RNA genes expressed from *Staphylococcus aureus* genomic and pathogenicity islands with specific expression among pathogenic strains. Proc Natl Acad Sci U S A. 102:14249–14254.

Raghavan R, Groisman EA, Ochman H. 2011. Genome-wide detection of novel regulatory RNAs in *E* coli. Genome Res. 21:1487–1497.

Raghavan R, Sage A, Ochman H. 2011. Genome-wide identification of transcription start sites yields a novel thermosensing RNA and new cyclic AMP receptor protein-regulated genes in *Escherichia coli*. J Bacteriol. 193:2871–2874.

Raghavan R, Sloan DB, Ochman H. 2012. Antisense transcription is pervasive but rarely conserved in enteric bacteria. mBio 3:e00156–12.

Rocha EP. 2004. Order and disorder in bacterial genomes. Curr Opin Microbiol. 7:519–527.

Rutherford K, et al. 2000. Artemis: sequence visualization and annotation. Bioinformatics 16:944–945.

Sittka A, et al. 2008. Deep sequencing analysis of small noncoding RNA and mRNA targets of the global post-transcriptional regulator, Hfq. PLoS Genet. 4:e1000163.

Skippington E, Ragan MA. 2012. Evolutionary dynamics of small RNAs in 27 *Escherichia coli* and *Shigella* genomes. Genome Biol Evol. 4: 330–345.

Sloan DB, Moran NA. 2013. The evolution of genomic instability in the obligate endosymbionts of whiteflies. Genome Biol Evol. 5:783–793.

Stone JR, Wray GA. 2001. Rapid evolution of cis-regulatory sequences via local point mutations. Mol Biol Evol. 18:1764–1770.

Storz G, Vogel J, Wassarman KM. 2011. Regulation by small RNAs in bacteria: expanding frontiers. Mol Cell. 43:880–891.

Suyama M, Bork P. 2001. Evolution of prokaryotic gene order: genome rearrangements in closely related species. Trends Genet. 17:10–13.

Tamas I, et al. 2002. 50 million years of genomic stasis in endosymbiotic bacteria. Science 296:2376–2379.

Tjaden B, et al. 2002. Transcriptome analysis of *Escherichia coli* using high-density oligonucleotide probe arrays. Nucleic Acids Res. 30: 3732–3738.

Treangen TJ, Rocha EP. 2011. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. PLoS Genet. 7: e1001284.

Unterholzner SJ, Hailer B, Poppenberger B, Rozhon W. 2013. Characterisation of the stbD/E toxin-antitoxin system of pEP36, a plasmid of the plant pathogen *Erwinia pyrifoliae*. Plasmid 70:216–225.

Van Alst NE, Picardo KF, Iglewski BH, Haidaris CG. 2007. Nitrate sensing and metabolism modulate motility, biofilm formation, and virulence in *Pseudomonas aeruginosa*. Infect Immun. 75:3780–3790.

Wassarman KM, Storz G. 2000. 6S RNA regulates *E. coli* RNA polymerase activity. Cell 101:613–623.

Wilderman PJ, et al. 2004. Identification of tandem duplicate regulatory small RNAs in *Pseudomonas aeruginosa* involved in iron homeostasis. Proc Natl Acad Sci U S A. 101:9792–9797.

Associate editor: Ruth Hershberg