UNIVERSITÀ DEGLI STUDI DI NAPOLI

"FEDERICO II"

Scuola di Dottorato in Medicina Molecolare

Dottorato di Ricerca in Genetica e Medicina Molecolare

Intergenic sequences in prokaryotes: structure, genome organization and functional properties

Coordinatore:
Prof. Carmelo Bruno Bruni

Candidato:
Dott. Giustina Silvestro

Anno

2007

UNIVERSITÀ DEGLI STUDI DI NAPOLI

"FEDERICO II"

Dipartimento di Biologia e Patologia Cellulare e Molecolare "L. Califano"

Dottorato di Ricerca in Genetica e Medicina Molecolare

Coordinatore Prof. Carmelo Bruno Bruni

Sede amministrativa:
Dipartimento di Biologia e Patologia Cellulare e Molecolare "Luigi Califano"

UNIVERSITÀ DEGLI STUDI DI NAPOLI

"FEDERICO II"

Dipartimento di Biologia e Patologia Cellulare e Molecolare "L. Califano"

Tesi di Dottorato di Ricerca in Genetica e Medicina Molecolare
XIX ciclo

Intergenic sequences in prokaryotes: structure, genome organization and functional properties

Candidato: Giustina Silvestro

Docente guida: Prof. Pier Paolo Di Nocera

INDEX

INTRODUCTION

Large scale genome-sequencing projects, run over the last fifteen years by different laboratories, made possible to elucidate the complete nucleotide sequence of one or more strains of over 300 bacterial species. The availability of wholly-sequenced genomes also stimulated investigations aimed to understand the functional organization of the bacterial chromosomes, and reconstruct how they have been remodeled in evolution by processes of gain and loss of genetic material (Frank et al., 2002; Achaz et al., 2003; Audit et al., 2003; Rocha and Danchin, 2003; Chain et al., 2004; Rocha, 2006). In prokaryotes, approximately 90% of the DNA is codogenic. A significant fraction of intergenic DNA, which separate ORFs (open reading frames), is composed by repetitive DNA sequences. In many species, repetitive DNA significantly contribute to the genetic landscape, and changes in the distribution of members of repeated DNA families among strains are routinely exploited for diagnostic purposes (van Belkum et al., 1999).

The main class of repetitive DNA sequences, located in prokaryotic genomes, is represented by IS (insertion sequences), genetic mobile elements feature by long terminal inverted repeats (TIRs) ranging in size from 10 to 40 bp. The ISs vary in size from 750 to 2500 bp and are capable to codify for a transposase, necessary for their transposition. The IS integration often determine a target site duplication; accordingly insertion site analyses show that ISs are always flanked by short direct repeats (DR), ranging in size from 2 to 14 bp. The ISs are able to provoke different types of genetic arrangements, like deletions and inversions, which determine assembly of genes with specialized

functions in clusters. Moreover ISs may be involved in activation or inhibition of genetic expression (Mahillon et al., 1999).

Prokaryotic genomes contain different classes of repetitive DNA sequences which have a more shorter size (40-300 bp) than ISs. Some laboratories are interested, in the past, in characterizing these sequences, defining their genomic organization and asking themselves on possible functional roles that these sequences may have. The biological effects caused by DNA repeats are different. Knowledge on the function of sequence repeats derive mostly from analyses carried out on the short (40 bp) palindromic repeats, found in *E. coli* and other enterobacteriaceae, called REPs. REPs can work both as DNA and RNA elements. As DNA elements. REPs are targets for the DNA gyrase (Yang and Ames, 1988), and may act by alleviating torsional stresses accumulated in the chromosome by transcriptionally induced positive supercoiling. REPs may also act as determinants of segmental RNA stability, by protecting from degradation the upstream segments of mRNAs in which they are embedded (Higgins et al., 1988).

In the last few years it has been identificated, in a great number of vertebrate and invertebrate genomes, mobile DNA sequences called MITEs (Miniature Inverted-repeat Transposable Elements; Feschotte et al., 2002). MITEs are transposable elements non autonomous, i.e. are inable to codify for a transposase but have in *cis* sequences necessary for their transposition. MITEs are <600 bp long and are featured by TIRs. It is noted that RUP (repeat unit of pneumococcus) elements, located in the genome of *S. pneumoniae*, are an example of MITEs in prokaryotes. The mobilization of these elements may be

mediate by the transposase codified by IS630-Spn1 (Oggioni and Claverys, 1999). Another example of MITEs is represented by NEMIS (Neisseria Miniature Insertion Sequences) repeat family, which has been well analyzed and characterized in the laboratory where I have spend my period of doctorate. NEMIS elements are short DNA elements of 108-158 bp featured by 26-27 bp long TIRs and represented 2% of *N. meningitidis* genome. NEMIS elements are frequently inserted at 5' or 3' of coding regions and are co-transcribed with genes. At RNA level, these elements fold into stem-loop structures (SLSs) and the hairpins formed by TIRs of NEMISs are processed by ribonuclease III (RNaseIII). This interaction regulate the expression of genes, which are flanked by NEMIS repeats, at the post-trascriptional level (Mazzone et al., 2001; De Gregorio et al., 2002, 2003).

In the last few years, we have been interested in defining the organization and the possible role of abundant families of small sequence repeats punctuating the chromosomes of Yersiniae. On the basis of *in vitro* and *in vivo* assays, we reached the conclusion that most members of the *Y. enterocolitica* ERIC (Enterobacterial Repetitive Intergenic Consensus Sequence) and YPAL (*Yersinia* palindromic elements) families function as cis-acting RNA regulatory sequences, controlling at the post-transcriptional level the rate of expression of neighbouring genes (De Gregorio et al., 2005; 2006). The folding into SLSs is crucial for the functioning of all these sequences as RNA control elements. On the base of these results, we thought of interest to investigate on the occurrence of families of intergenic prokaryotic sequences able to fold into SLSs in a systematic fashion. To this end, wide-genome

analyses have been carried out in a representative set of 40 prokaryotic genomes (Petrillo et al., 2006). Populations of isolated SLSs have been subsequently screened by a combination of criteria, pruning procedures leading to the identification of SLS subsets which share similarities in sequence and a predicted secondary structure (Cozzuto et al., 2007). Most of these sequences appear to be members of repeated DNA families. Some correspond to already known families, but a number of them represents novel ones (Cozzuto et al., 2007).

This study provides information on structural organization, patterns of chromosomal interspersion, folding aptitudes and evolutionarily relatedness of the predominant SLS types identified in the prokaryotic genomes analyzed. Results integrate data emerging from earlier surveys on repetitive sequences carried out on bacterial chromosomes and add knowledge to the field in many respects. The functional roles that specific SLS-containing repeats may play in different genomes are also discussed.

AIM OF THE THESIS

In the present study, I have analyzed and characterized 27 repeated DNA families able to fold, at RNA level, into SLSs. The analyzed SLS-containing sequences are been searched by bioinformatic analyses carryed out with a systematic manner in 40 wholly-sequenced bacterial genomes, constituting a representative sample of the prokaryotic world in terms of evolutionary distance, genome complexity and GC content.

Some families are repetitive DNA elements already described in literature, others are novel. On the base of SLS size and folding type, we have distinct SLS-containing repeats in two main groups: type-I and type-II. SLSs formed by type-I elements vary in size from 40 to 60 bp. On the base of their structural organization these elements are been subdivided in two subtypes: type-IA and type-IB. Type-IA are capable to fold into a single hairpin, while type-IB elements, in virtue of their dimeric organization, may form a secondary structure featuring two hairpins. Type-II group include an heterogeneous set of SLS-containing repeats, varying in size from ˜50 to ˜300 bp long. On the base of their potentially formed secondary structure, type-II elements are been distinct in type-IIA and type-IIB. The folding of type-IIA elements is directed by TIRs complementarity and therefore these repeats fold into stable canonical SLSs. In contrast, the two type-IIB elements, BruRS and EFAR, are able to fold into a peculiar secondary structure featured by two SLSs connected by a 15-20 bp spacer.

On the base either of literature data either of our results obteined by experimental analyses, carry out on two repetitive DNA families, ERIC and

YPAL, we belive that the folding into SLSs is essential for the functional role that these elements assumed as modulators of RNA decay.

Hairpins formed by transcripts ERIC+, according to their orientations and their relative position in mRNA, may increase or decrease the degradation of trancripts ERIC+ from degradative machinery of bacterial RNAs, known as degradosome. Similarly hairpins formed by mRNA YPAL+ are processed by RNase III. The cleavage may modulate tournover of transcripts YPAL+, and expression levels of homologous trancripts YPAL+ and YPAL- in different *Y. enterocolitica* strains are in fact quite different.

In contrast to ERICs and YPALs, EFAR sequences lack TIRs. The secondary structure of EFAR sequences are featured by a short SLS1 and a long SLS2 separated by a 20 bp long unfolded segment. Through *in silico* analyses, the identification of EFAR sequences corrisponding only to the short SLS corroborate the hypothesis that EFAR repeats derived from the fusion of indipendent SLSs.

MATERIALS AND METHODS

Genomic sequence data

Complete genomic sequences and their annotations about CDS, rRNA and tRNA were downloaded from the online repository made available at The Institute for Genomic Research (TIGR). Automatic annotations have been stored into a SQL database (SLS-DB), for further analysis.

Constraints established for SLS identification

SLS identification was performed by using the program *rnamotif* of the package RNAMOTIF, version 2.1.2 (Macke et al., 2001) according to the following rules:

-GU pairing in the stem was allowed

-the minimal stem length was 12 bp

-loop length could vary from 5 to 100 nt

-1 bulged or 1 mispaired base, at least two matches away from the ends of the stem, was allowed.

As a consequence of the constraints imposed, the smallest SLS that could be found is 29 bp.

The Gibbs free energy (*dG*) of each SLS containing region was calculated by calling the built-in function *efn2* of *rnamotif*. The minimum free energy with no constraint for SLS formation was obtained by running the program *mfold* developed by Zuker and coworkers (Mathews et al., 1999) on the SLS sequences.

Refinement of the initial clustering

Out of the SLSs previously identified in 40 bacterial genomes (Petrillo et al., 2006) only those predicted to fold with a free energy <–5 Kcal/mol were selected to look for sequence repeat families (Cozzuto et al., 2007).

For each genome, selected SLSs were clustered according to a procedure based on BLAST and MCL programs (Altschul et al., 1990, Enright et al., 2002). In order to identify all family members of each cluster, a pipeline was developed, based on cycles of alignment by PCMA and search on the genome by HMMER package tools (Bateman et al., 1999).

Aptitude to form a stable secondary structure

SLS ability to fold into a reliable secondary structure was analyzed by using RANDFOLD (Bonnet et al., 2004), which compares the predicted minimum folding energy (MFE) of a sequence with those of a large number of random shuffles of the same sequence. Conserved secondary structures were predicted by RNAz analyses (Washietl et al., 2005).

RESULTS

A systematic analysis of a representative set of bacterial genomes produced a large collection of sequences, potentially able to fold into high stability SLSs (Petrillo et al., 2006). Some of these SLSs were shown to have strong similarity with each other, and a pipeline combining sequence clustering and Hidden Markov Model (HMM) based searches, was developed. This strategy led to the definition of 92 families of SLS-containing sequences exhibiting sequence similarity. One third of the families had no shared secondary structure, and has been detected mostly because of the incidental presence of SLSs within large repeated sequences. Some of the structured families were found within CDSs, where the formation of secondary structures in the RNA is expected to be limited by the translation machinery. Others resulted from the recurrence, in the genomes of *Pasteurella multocida*, *Haemophilus influenzae* and *Neisseria meningitidis*, of short oligonucleotides called DUS (for DNA uptake sequence) located at close distances in head-to-head orientation. All such families were not analyzed in this study. The SLS-containing sequences analyzed here are mostly located in the intergenic space. Consequently, except for those located at the boundary of genes, the folding of such sequences as RNA elements is presumably unaffected by translating ribosomes.

On the basis of SLS size and folding aptitude, SLS-containing sequences may be sorted into two main categories or sequence types (Table I). Type-I elements have the potential to form secondary structures consisting of 10-20 bp double-stranded stems, which delimit single-stranded loops ranging between 4-15 nt. On the basis of their genomic organization, the elements belonging to this group can be further subdivided into type-IA and type-IB

subtypes, with type-IA repeats typically scattered along the chromosome as single units, and type-IB frequently organized as dimers. Elements of type-IB families may often be found in larger arrays of up to 7 to 10 copies.

Type II elements are longer than type I, and their termini can be delimited by the duplication of bases at the insertion site. On the basis of their folding characteristics, these sequences have been sorted into two main subsets. Type-IIA elements fold into the canonical one hairpin structure, while type-IIB elements fold into more complex, two hairpin, structures.

Properties of type-I and type-II elements are discussed below.

Type-I elements, group A

Consensus sequence and predicted secondary structure of type IA elements are shown in Fig. 1. Except for REP 2 (Parkhill et al., 2000), all these elements have not been previously described. Pae-4 and Myt-1 define moderately abundant (40 to 60 units) DNA families, accounting for ˜0,05% of the whole DNA of *P. aeruginosa* and *M. tuberculosis*, respectively. Homology searches revealed that Pae-4 elements are restricted to the *P. aeruginosa* genome, while small Myt-1 families, including only 21 and 9 intact members, have been found in the sequenced *M. avium* paratuberculosis and *M. smegmatis* MC2 strains, respectively. Pae-4 family members feature 12-13 bp long stems, which result from the adjoining of an outer AT-rich and an inner GC-rich segment, separated by 4 nt loops (Fig. 1), and can be viewed as a sequence-specific variant of short SLSs over-represented in the genomes of all low-GC firmicutes, which are characterized by terminal A_4 and T_4 runs (Petrillo et al., 2006). Myt-1 elements

feature stems ranging in size from 20 to 26 bp, and loops varying in size from 18 to 8 nt (Fig. 1). The abundance of the two Myt-1 sequence types is comparable. Myt-1 are distributed throughout the *M. tuberculosis* genome, and differ from already known repetitive sequences characterized in tubercle bacilli, as they are not components of the mosaic RLEP nor of REPLEP repeats (Cole et al., 2001), and do not exhibit relatedness to MIRUs, small intergenic sequences clustered at a few chromosomal sites extensively exploited as strain-specific markers in M. tuberculosis (Supply et al., 2000).

Bhal-1 and Clot-1 elements feature SLSs embedded in relatively conserved unstructured regions (Fig. 1). The REP 2 elements are internal to the homonymous 154 bp long sequence previously annotated in the genome of the Z2491 *N. meningitidis* strain (Parkhill at al., 2000).

Type-I elements, group B

This group includes elements which have been extensively characterized at the experimental level, such as the *E. coli* REPs (Higgins et al., 1988), and elements annotated, but not functionally analyzed, such as the *P. putida* REP (Aranda-Olmedo et al., 2002), the *E. coli* BoxC (Bachellier et al., 1999), the *N. meningitidis* dRS3 (Parkhill et al., 2000) and the *R. conorii* RPE6 (Ogata et al., 2001) sequences, and two novel families emerging in this study, Bor-1 and Pae-1, found in the chromosomes of *Bordetellae* and *Pseudomonas aeruginosa*, respectively. For clarity, in this study the abundant sequence repeats already described in *E. coli, Salmonellae* and *P. putida* as REPs have been renamed

EPU and PPU, for *Enterobacterial* and *Pseudomonas* palindromic units, respectively.

Bor-1 elements contribute to ˜0.1% of the genome in all *Bordetella* species analyzed. In contrast, Pae-1 are species-specific units which make up ˜0.07% of the P. *aeruginosa* chromosome. Homology searches failed to identify Pae-1 units in the genomes of other Pseudomonaceae. Consensus sequences and predicted secondary structures of type IB elements are shown in Fig. 2.

Compensatory nucleotide changes at specific positions do not interfere with the base pairing of complementary segments, and allow sorting members of most type-IB element families into two to three predominant sequence subtypes (Fig. 2). The Hofnung's lab nomenclature referring to Z and Y repeats has been conserved for EPU sequences (Bachellier et al., 1999). For the other repeats, the predominant sequence subtypes are referred to as **a** and **b** (Fig. 2). PPU repeats are extremely heterogeneous, nearly half of them fitting altogether the **a** (35%) and **b** (11%) consensus sequences reported in Fig. 2, the remaining elements the minor **c** to **h** variants (not shown).

A relevant fraction of type-IB elements is organized as dimers, i.e. repeats reiterated at a ˜10-100 bp distance. This configuration creates the opportunity for dimers to fold, rather than as two separate hairpins, as an alternative, larger SLS in which the stem results from the pairing of the two individual units (Fig. 2).

In a fraction of the dimers, units are separated by one or a few specific spacer sequences. In other instances, the spacer varies both in length and base composition. In the same genome, different classes of dimers may coexist (Fig.

3). The dimerized elements belong to the same or different sequence subtype. Bor-1 and RPE6 repeat families feature both classes of dimers. In contrast, all EPU dimers result from the combination of Y and either Z1 or Z2 subtype elements (Fig. 3). PPU, BoxC and Pae-1 dimers are also exclusively formed by units of different subtypes. Dimers may be formed by units in either head-to-head (convergent elements) or tail-to-tail (divergent elements) configuration. According to the nomenclature in use (Bachellier et al., 1999) the distal end (or head) of EPU features the tetranucleotide CTAC, not engaged in base pairing (Fig. 2), and the convergently oriented Y and Z1 units, and the divergently oriented Y and Z2 units make up the known EPU dimers described as BIME-1 and BIME-2 repeats, respectively (Espeli and Boccard, 1997). Dimers formed by either convergent or divergent Pae-1 and PPU, which also feature the CTAC tetranucleotide at one terminus, have been identified.

Within each species, a fraction of type-IB elements are clustered as arrays. Single repeats, reiterated in tandem in a head-to-tail configuration, and dimers, are found clustered at different *loci*. Pae-1 dimers are invariably reiterated together with a variable (20 to 120 bp) amount of flanking DNA. In clusters formed by elements of the EPU and Bor-1 families, in contrast, just the sequence of the repeat is duplicated. Yet, flanking DNA is reiterated with EPU sequences at three *loci* in the *E. coli* K 12 chromosome (http://www.pasteur.fr/recherche/unites/pmtg/repet/index.html), suggesting that repeats of the same type may undergo different patterns of amplification.

The relative abundance of single, dimeric and clustered elements varies widely among type-IB repeat families (Fig. 3). Single units constitute a minor fraction

of the EPU family, most family members being found as dimers (44%) and clusters (41%). In contrast, PPU elements are never found in clusters, and the repeat family is formed by only single (25%) and dimeric (75%) repeats. *R. conorii* RPE6 and *N. meningitidis* dRS3 repeats are, but for a few changes in the loop region, identical (Fig. 2). However, the two repeat families exhibit a quite different genomic organization. RPE6 are found as single units or dimers featuring a few alternative spacer sequences (Fig. 3). In contrast, about 90% of dRS3 elements are found interleaved with >30 different RSs (repeat sequences), the most frequently ones being RSs 13, 14, 17, 25 and 26, in the large composite intergenic regions called NIMEs (Parkhill et al., 2000). Actually, NIMEs can be viewed as a scrambled collection of dRS3 dimers, featuring specific repeats as spacers. Taking into account the relatedness to RPE6, dRS3 are larger than the units annotated in the genome of the Z2491 *N. meningitidis* strain (Parkhill et al., 2000), and the sizes of their flanking repeats accordingly reduced.

Type-II elements, group A

Type-IIA elements typically feature a long base paired region, interrupted by one or more internal loops or bulges, closed by 15-18 bp TIRs. This structure is almost invariably framed by target site duplications (TSDs) of variable length, indicative of their origin as mobile sequences. Members of most type-IIA families characteristically retain the same termini, but often vary at the closed end of the stem loop, resulting in different sizes because of deletion/insertions of specific internal segments. Most type-IIA repeats have been extensively

analyzed (Oggioni and Claverys, 1999; Mazzone et al., 2001; Brugger et al., 2002; Mràzek et al., 2002; Claverie and Ogata, 2003; Okstad et al., 2004; De Gregorio et al., 2005; De Gregorio et al., 2006; Knutsen et al., 2006).

In our study, we have identified three novel type-IIA-like elements: Sal-1, Cod-1, and Clop-1. Sal-1 repeats feature 15 bp long stems, and range in size from 151 to 212 bp, length heterogeneity being due to the presence/absence of a 61 bp long central DNA segment (Fig. 4). In this figure is reported the folding of 119 bp long Sal-1 consensus sequence because of no base conservation in all Sal-1 members at right terminus. Sal-1 repeats are restricted to the genus Salmonella, with the largest (20 members) Sal-1 family found in *S. typhimurium* LT2 strain. The alignment of homologous chromosomal regions carrying a member of the Sal-1 family in *S. typhimurium*, but lacking it in *S. typhi*, led to delimit the ends of Sal-I elements, and establish that their insertion induces the formation of 3 bp long TSDs, which match the consensus sequence TWA (data not shown).

Cod-1 and Clop-1 do not feature TSDs, and are restricted to the genomes of C. diphteriae and C. perfringens, respectively. Cod-1 elements consist of a 95 bp SLS featuring a 19-20 bp long stem. Careful manual analysis of Clop-1 family revealed that, in addition to the previously reported 26 elements, which measure 67 bp, a larger number (33 elements) may be observed which include additional segments bringing the total size to 199-265 bp. The insertion consists of two conserved 66 and 55 bp segments, separated by two alternative DNA tracts (lowercase residues in brackets in Fig. 5 measuring 78 bp (11 elements) or 12 bp (22 elements). These insertion modules

are unable to fold into a secondary structure and disrupt the overall Clop-1 structure.

Type-II elements, group B

This repeat subset includes only two sequence types, Bru-RS and EFAR, both predicted to fold as a complex structure, which includes two long, base paired structures of different sequence, joined by an unstructured connector (Fig. 6). Interestingly, these structures tend to have short inverted repeats at the very ends, which are not part of the predicted structure. Such terminal repeats have previously been noted in the original description of the Bru-RS family (Halling and Bricker, 1994). Bru-RS elements come in two 65% homologous variants, Bru-RS1 and Bru-RS2, which measure 103 and 105 bp, respectively. Some Bru-RS2 feature 42 bp DNA insertions (see large Bru-RS2 in Fig. 7). The Bru-RS elements found in the genome of most Brucella species are homogeneously distributed on the two chromosomes, two-thirds being located on chromosome I (˜2,1 Mb), and the rest on chromosome II (1.1 Mb). Although Brucellae share sequence identity over 94% (Chain et al. 2005), the genome of *B. melitensis* 16M strain contains an unusual number of BRU-RS elements, about two times higher than in the other species. This excess is mostly concentrated in chromosome I, and consists of a selective over-representation of the BRU-RS2 subtype; BRU-RS1 elements show only a minimal increase.

EFAR repeats are predicted to fold into a complex secondary structures, which results from the combination of a stable short SLS with a long double-stranded region. The identification in silico of EFAR units spanning just

the short SLS supports the notion that EFAR derive from the fusion of independent SLSs (Venditti et al., 2007).

Organization and characterization of three SLS families

During the period of doctorate I were interested at study of organization and at possible functional role of two type-IIA families, ERIC and YPAL located in *Yersiniae* and previously described in literature, and one of type IIB family, EFAR sequences, a novel class of DNA repetitive sequences, identificated by our systematic analysis in the *E. faecalis* V583 chromosome. Structural and functional features of these families are summarized below.

ERIC elements comes in three different size (De Gregorio et al., 2005). Unit-length ERICs are 127 bp; shorter elements measure ~70 bp, lacking a 50-bp-long internal segment and larger elements are interrupted at specific sites by three different types of DNA insertions (Fig. 9A). In contrast both YPALs and EFARs have a modular organization, being made by a combination of 9 and 16 modules respectively (De Gregorio et al., 2006, Venditti et al., 2007). Thus they vary in size from 130 to 210 bp and from 42 to 650 bp respectively (Fig. 9B and 9C).

ERICs and YPALs are more similar than EFARs, because they may fold, at RNA level, in stable canonical SLS (Fig. 6 and Fig. 10). All these families partly resemble MITEs, small noncodogenic sequences, which also fold into secondary structures. MITEs feature long TIRs, and their mobilization is mediated by transposases encoded by ISs featuring similar TIRs (Oggioni and Claverys, 1999; Brugger et al, 2002; De Gregorio et al., 2003). ERICs feature

long TIRs and can fold as RNA into single hairpin structures. YPAL repeats do not feature long TIRs, but their preferential insertion into regions of dyad symmetry contributes to stabilize their ends by forming longer complementary tracts at their termini. Intriguingly many YPAL targets overlap rho-independent transcriptional terminator-like sequences. This conclusion was supported both by the presence of runs of thymidines immediately next to palindromes targeted by YPAL and by the finding that most tergets were found located next to the stop codon of annotated ORFs in sequenced *Yersinia* genomes. (De Gregorio et al., 2006). From the alignment of DNA regions from the wholly sequenced *Y. enterocolitica* 8081 and *Y. pestis* CO92 strains, we could establish that insertion of ERIC elements leads to duplication of the dinucleotide TA while the genomic spread of YPALs, occurred via transposition, was supported by the presence at their termini of TSD which vary in both sequence and size (from 3 to 26 bp). In contrast, EFAR lack TIRs and seem to transpose by an unusual cut-and-paste process. Most EFAR elements measure 170 bp, and can fold into peculiar L-shaped structures in which a short SLS1 and a long SLS2 are separated by 20 nt single stranded region. Homologous chromosomal regions lacking or containing EFAR sequences were identified by PCR among 20 *E. faecalis* clinical isolates of different genotypes (Venditti et al., 2007). Like YPAL elements, sequencing of a representative set of ‘empty’ sites revealed that 24–37 bp long sequences, unrelated to each other but all able to fold into SLSs, functioned as targets for the integration of EFAR. In the process, most of the SLSs had been deleted, but part of the targeted stems had been retained at EFAR termini.

Interspersion analyses with coding regions revealed that most elements of the three DNA repetitive families analyzed tend to be inserted closely downstream from the stop codon of flanking genes. The proximity to coding regions suggests that most ERICs, YPALs and EFARs are co-transcribed with flanking genes. To this end experimental procedures were setting up for ERIC and YPAL repeats. Whole *in silico* surveys surprisingly revealed a privileged orientation of ERIC sequences relative to their position in the mRNA. ERICs which either overlap or are located next to stop codons are preferentially inserted in the same (or B) orientation. In contrast, ERICs located far apart from open reading frames are inserted in the opposite (or A) orientation. The expression of genes cotranscribed with A- and B-oriented ERICs has been monitored *in vivo* (De Gregorio et al. 2005). In mRNAs spanning B-oriented ERICs, upstream gene transcripts accumulated at lower levels than downstream gene transcripts. This difference was abolished by treating cells with chloramphenicol. We hypothesize that folding of B-oriented elements is impeded by translating ribosomes. Consequently, upstream RNA degradation is triggered by the unmasking of a site for the RNase E located in the right-hand TIR of ERICs (Fig. 11). A-oriented ERICs may act in contrast as upstream RNA stabilizers or may have other functions. The hypothesis that ERICs act as regulatory RNA elements is supported by analyses carried out in *Yersinia* strains which either lack ERIC sequences or carry alternatively oriented ERICs at specific loci. Similarly to ERICs, YPAL RNAs fold into stable hairpins which may modulate mRNA decay. Accordingly, we found that YPAL-positive transcripts accumulate in *Yersinia enterocolitica* cells at significantly higher

levels than homologous transcripts lacking YPAL sequences in their 3' untranslated region (De Gregorio et al., 2006).

DISCUSSION

In addition to transposons and insertion sequences, different classes of repetitive DNA are present in prokaryotic genomes. Many of these are able to fold into SLSs and their functional role is unknown or partially note. In this study we have interested in the search of repeated DNA sequences in bacterial chromosomes. By our analyses we have observed that the number of predicted SLSs is significantly higher in prokaryotic genomes existing in nature than in random sequences of comparable GC content (Petrillo et al., 2006). This implies that the ability of a variety of sequences to fold into secondary structures, either at the DNA or RNA level, is positively selected in prokaryotic genomes, and may have functional significance. Some SLSs-containing sequences resulted to be members of repeated DNA families (Cozzuto et al., 2007). Some of them have been extensively analyzed experimentally, while others have been only reported, or are completely novel. The identification of repeated palindromic sequences is typically biased by the chosen screening constraints. In our studies, limiting the search to SLSs, featuring stems of defined minimal lengths and mismatches (Cozzuto et al., 2007), prevented the identification of a small number of palindromic sequences such as the *E. coli* 29 bp repeats (Bachellier et al., 1999), and some of the RPEs found in *R. conorii* (Ogata et al., 2001). Likewise, previous searches (Tobes and Ramos, 2005) for abundant intergenic repeats yielded several families reported in this study, but failed to identify the abundant family of Bor-1 elements in Bordetellae. The observed objects vary extensively in size and number, but remarkably may be assigned to a restricted number of sequence classes. According to the features of their secondary structure, the repeats may be grouped into type-I and type-II

elements. The size of the conserved stem-loop contained in both type-I and type-II elements falls in a few, narrow length windows. Type-I elements feature stem loops measuring either 30 to 40, or ˜60 nt. The stem-loops in type-II elements are more heterogeneous in length, but most of them fall into discrete size classes (see Table I). A tighter classification may take into account the relatedness exhibited by specific repeats. dRS3 and RPE6 are likely related by descent (see Fig. 2). The identification of variants of one single type of repeat in the genome of microrganisms as distant as *N. meningitidis* and *R. conorii* mirably illustrates how the lateral gene transfer may have contributed to the reshaping of bacterial chromosomes during evolution. It is plausible to hypothesize that dRS3 elements evolved from repeats imported from *Rickettsiae*, and not *viceversa*, as *Neisseriae* are prone to acquire exogenous DNA by transformation. The event must have occurred prior to *Neisseriae* speciation, since dRS3 are also found in the genomes of *N. gonhorroeae* and *N. lactamica* species. BLAST analyses failed to identify sequences homologous to other *R. conorii* repeats in *Neisseriae* or to other bacterial species.

EPU, PPU and Pae-1 repeats, although exhibiting poor homology to each other, intriguingly share the same tetranucleotide CTAC (or GTAG, on the opposite strand), at one terminus (Fig 2). We speculate that these elements are members of a large super-family of repeats spread in gamma-proteobacteria, and that DNA-binding proteins, able to recognize the terminal CTAC/GTAG sequences, may account for their spreading, and/or provide them a role. EPUs have been shown to interact with the DNA gyrase, but CTAC/GTAG sequences are not targeted by the enzyme (Espeli and Boccard, 1997). The hypothesis that

these elements form a superfamily gains support from the identification in the genomes of several *Pseudomonaceae* and *Xanthomonaceae* (Tobes and Pareja, 2005; Feil et al., 2005, P.P. Di Nocera, unpublished results) of type-IB elements, which are predominantly organized as dimers, and similarly feature CTAC/GTAG sequences at their termini. Bor-1 elements partly recall EPU and Pae-1 at the sequence level, but lack terminal CTAC/GTAG residues. It is difficult therefore to assess whether they are species-specific variants of the supposed superfamily, or represent a type of repeats arisen independently in *Bordetellae*.

A significant fraction of type-IB elements is organized as dimers. Given the limited extent of base-pairing within monomers, the SLSs which have been pinned down by our searches are predominantly hairpins formed by the base-pairing elements forming the dimer. It is intriguing to note that most dimers result from the association of allelic repeats (Fig. 2). We do not have an answer for this finding, but dimers made up by sequence variants cannot fold into perfect stem-loop structures, and this may plausibly preserve them from being erased from the chromosome. As several dimers are located at the 3' end of transcriptional units, and are thus plausibly transcribed, the bulged hairpins formed by allelic units would be protected from cleavage by RNAseIII, an endoribonuclease known to recognize perfect stem-loop structures.

The presence of conserved spacers suggests that some type-I dimers are privilegedly expanded sequences. The spacer of the EPU dimers known as BIME-1 binds IHF, a pleiotropic protein responsible for bending DNA. It had been proposed that IHF may assist EPU-gyrase interactions (Boccard and

Prentki, 1993), but the hypothesis had been ruled out by experiments in which EPUs containing or lacking the IHF site, were compared as topological insulators in vivo (Moulin et al., 2005). Plausibly, IHF has (or had) a role in the transposition of BIME-1, which are framed by target site duplications as typical mobile DNA sequences do, and analogous DNA-binding proteins may interact with the spacers of other conserved dimers.

Type II elements were sorted into two main groups according to their secondary structures. This classification probably underlies their different evolutionary origin. Type-IIA elements are plausibly all deletion-derivatives of large, codogenic ISs. These progenitor elements, with the possible exception of RUPs (Oggioni and Claverys, 1999), are likely extinct, but enzymes encoded by unrelated ISs, able to recognize the TIRs, allow some type-IIA elements to transpose. Type-IIB elements, by contrast, plausibly derive from the adjoining of two independent SLSs. This hypothesis, supported by the identification of genomic intervals spanning single EFAR-SLS modules in *E. faecalis* (Venditti et al., 2007), is validated by the identification of homologous EFAR repeats in *E. faecium* which result from the assembly of alternative, independent SLS modules (De Gregorio, Silvestro and Di Nocera, in preparation).

The pattern of the genomic distribution of SLS elements may provide hints on the functional role that distinct palindromic repeats may play. Known elements are frequently located at short distance from the stop (<30 bp) or the start (<50 bp) codon of flanking ORFs, or both. This implies that many SLS are transcribed, and may fold into secondary structures at the RNA level. The same picture can be drawn by looking at the interspersion pattern of a palindromic

element subset previously undescribed with coding sequences (Fig. 8A). It has been shown that RNA hairpins formed in the 5'- or the 3'-untranslated region may protect mRNAs from degradation, or enhance the process upon cleavage by specific endoribonucleases (De Gregorio et al., 2003; Rouquette-Loughlin, 2004; De Gregorio et al., 2006). The presence of ribosomes may affect the formation of RNA hairpins. In *Y. enterocolitica*, we have observed that several ERIC elements are located between genes transcribed unidirectionally. The distance from the stop codon of the upstream ORF determines whether these elements fold into a secondary structure, and this ultimately influences the relative stability of upstream and downstream mRNA segments (De Gregorio et al., 2005). Similarly YPALs are frequently located at the 3' ends of *Yersinia* genes and often between genes transcribed in a convergent manner. By *in vivo* and *in vitro* assays we have showed that YPALs may impede the degradosome and they might also stimulate degradation of some mRNAs in which they are embedded (De Gregorio et al., 2006).

Several Bru-RS and Pae-1 repeats are at close distance from the stop codons of upstream ORFs (Fig. 8A). The stop codons are actually provided by the *termini* of the elements, which feature TAG triplets, in 25/51 cases for Bru-RS, and 13/44 cases for Pae-1. A ribosome temporarily stalled at the stop codon may interfere with the folding of either one of the two SLSs featured by Bru-RS. This may have relevance in view of the hypothesis, supported by assays carried out with EFAR repeats, that transcripts spanning type-IIB elements may be processed at the SLSs junction (De Gregorio, Silvestro and Di Nocera, in preparation). In the case of Pae-1 elements, which are mostly organized as

dimers, the ribosome, by interfering with the folding of one repeat, may force both repeats to fold into a single, large SLS which may have regulatory function (Fig. 8B).

Future investigations will provide a more articulate knowledge of the structure, folding properties, and pattern of interspersion of palindromic sequence repeats present in the prokaryotic world. Available informations alredy let to design experiments aimed to assess the function of specific SLS-containing repeats, and thus possibly predict the role of a larger number of structurally related sequences scattered, in the same chromosomal environment, as single-copy units.

REFERENCES

Achaz G, Coissac E, Netter P, Rocha EP.

Associations between inverted repeats and the structural evolution of bacterial genomes.

Genetics 2003, 164:1279-1289.

Altschul, SF, Gish, W, Miller, W, Myers, EW, Lipman, DJ.

Basic local alignment search tool.

J. Mol. Biol. 1990, 215:403-410.

Aranda-Olmedo I, Tobes R, Manzanera M, Ramos JL, Marques S.

Species-specific repetitive extragenic palindromic (REP) sequences in Pseudomonas putida.

Nucleic Acids Res. 2002 Apr 15;30(8):1826-33.

Audit B, Ouzounis CA.

From genes to genomes, universal scale-invariant properties of microbial chromosome organisation.

J Mol Biol 2003, 332:617-633.

Bachellier S, Clement JM, Hofnung M.

Short palindromic repetitive DNA elements in enterobacteria: a survey.

Res Microbiol. 1999 150:627-639.

Bateman A, Birney E, Durbin R, Eddy SR, Finn RD, Sonnhammer EL.

Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins.

Nucleic Acids Res. 1999, 27:260-262.

Boccard F, Prentki P.

Specific interaction of IHF with RIBs, a class of bacterial repetitive DNA elements located at the 3' end of transcription units.

EMBO J. 1993 12: 5019-5027.

Bonnet E, Wuyts J, Rouze P, Van de Peer Y.

Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences.

Bioinformatics 2004 20: 2911-2917.

Brugger K, Redder P, She Q, Confalonieri F, Zivanovic Y, Garrett RA.

Mobile elements in archaeal genomes.

FEMS Microbiol Lett. 2002 206:131-141.

Chain PS, Carniel E, Larimer FW, Lamerdin J, Stoutland PO, Regala WM, Georgescu AM, Vergez LM, Land ML, Motin VL, Brubaker RR, Fowler J, Hinnebusch J, Marceau M, Medigue C, Simonet M, Chenal-Francisque V, Souza B, Dacheux D, Elliott JM, Derbise A, Hauser LJ, Garcia E.
Insights into the evolution of Yersinia pestis through whole-genome comparison with Yersinia pseudotuberculosis.
Proc Natl Acad Sci U S A 2004, 101:13826-13831.

Chain PS, Comerci DJ, Tolmasky ME, Larimer FW, Malfatti SA, Vergez LM, Aguero F, Land ML, Ugalde RA, Garcia E.
Whole-genome analyses of speciation events in pathogenic Brucellae.
Infect Immun. 2005 73:8353-8361.

Claverie JM, Ogata H.
The insertion of palindromic repeats in the evolution of proteins.
Trends Biochem. Sci. 2003 28:75-80.

Cole ST, Supply P, Honore N.
Repetitive sequences in Mycobacterium leprae and their impact on genome plasticity.
Lepr Rev. 2001 72:449-461.

Cozzuto, L, Petrillo, M., Silvestro, G, Di Nocera P.P. and Paolella, G. Systematic identification of stem-loop containing sequence families in bacterial genomes.

BMC Genomics 2007, in press

De Gregorio, E., Abrescia, C., Carlomagno. M. S. and P.P. Di Nocera. 2002. The abundant class of nemis repeats provides RNA substrates for ribonuclease III in *Neisseriae*.

Biochim Biophys Acta 2002. **1576**: 39-44.

De Gregorio E, Abrescia C, Carlomagno MS, Di Nocera PP.

Ribonuclease III-mediated processing of specific Neisseria meningitidis mRNAs.

Biochem J. 2003, 374:799-805.

De Gregorio E, Silvestro G, Petrillo M, Carlomagno MS, Di Nocera PP. Enterobacterial repetitive intergenic consensus sequence repeats in Yersiniae: genomic organization and functional properties.

J Bacteriol. 2005 187:7945-7954.

De Gregorio E, Silvestro G, Venditti R, Carlomagno MS, Di Nocera PP. Structural organization and functional properties of miniature DNA insertion sequences in Yersiniae.

J Bacteriol. 2006 188:7876-7884.

Enright AJ, Van Dongen S, Ouzounis CA.

An efficient algorithm for large-scale detection of protein families.

Nucleic Acids Res. 2002, 30:1575-1584.

Espeli O, Boccard F.

In vivo cleavage of Escherichia coli BIME-2 repeats by DNA gyrase: genetic characterization of the target and identification of the cut site.

Mol Microbiol. 1997 26:767-777.

Feil H, Feil WS, Chain P, Larimer F, DiBartolo G, Copeland A, Lykidis A, Trong S, Nolan M, Goltsman E, Thiel J, Malfatti S, Loper JE, Lapidus A, Detter JC, Land M, Richardson PM, Kyrpides NC, Ivanova N, Lindow SE. Comparison of the complete genome sequences of Pseudomonas syringae pv. syringae B728a and pv. tomato DC3000.

Proc Natl Acad Sci U S A. 2005 102:11064-11069.

Feschotte C, Jiang N, Wessler SR.

Plant transposable elements: where genetics meets genomics.

Nat Rev Genet. 2002 3:329-341.

Frank AC, Amiri H, Andersson SG.

Genome deterioration, loss of repeated sequences and accumulation of junk DNA.

Genetica 2002, 115:1-12.

Halling SM, Bricker BJ.

Characterization and occurrence of two repeated palindromic DNA elements of Brucella spp.: Bru-RS1 and Bru-RS2.

Mol Microbiol. 1994 14: 681-689.

Higgins CF, McLaren RS, Newbury SF.

Repetitive extragenic palindromic sequences, mRNA stability and gene expression: evolution by gene conversion? A review.

Gene 1988, 72:3-14.

Knutsen E, Johnsborg O, Quentin Y, Claverys JP, Havarstein LS.

BOX elements modulate gene expression in Streptococcus pneumoniae: impact on the fine-tuning of competence development.

J Bacteriol. 2006 88:8307-8312.

Macke TJ, Ecker DJ, Gutell RR, Gautheret D, Case DA, Sampath R.

RNAMotif, an RNA secondary structure definition and search algorithm. Nucleic Acids Res 2001, 29:4724-4735.

Mahillon J., Leonard C., Chandler M.

IS elements as constituents of bacterial genomes.

Res Microbiol. 1999 150, 675-687.

Mathews DH, Sabina J, Zuker M, Turner DH.

Expanded sequence dependence of thermodynamic parameters provides robust prediction of rna secondary structure.

J Mol Biol 1999, 288:910-940.

Mazzone M, De Gregorio E, Lavitola A, Pagliarulo C, Alifano P, Di Nocera PP.

Whole-genome organization and functional properties of miniature DNA insertion sequences conserved in pathogenic Neisseriae.

Gene. 2001 278:211-222.

Moulin L, Rahmouni AR, Boccard F.

Topological insulators inhibit diffusion of transcription-induced positive supercoils in the chromosome of Escherichia coli.

Mol Microbiol. 2005. 55:601-610.

Mràzek J, Gaynon LH, Karlin S.

Frequent oligonucleotide motifs in genomes of three streptococci. Nucleic Acids Res. 2002 30:4216-4221.

Ogata H, Audic S, Renesto-Audiffren P, Fournier PE, Barbe V, Samson D, Roux V, Cossart P, Weissenbach J, Claverie JM, Raoult D.

Mechanisms of evolution in Rickettsia conorii and R. prowazekii.

Science. 2001 293 :2093-2098

Oggioni MR, Claverys JP.

Repeated extragenic sequences in prokaryotic genomes: a proposal for the origin and dynamics of the RUP element in Streptococcus pneumoniae. Microbiology. 1999 145 :2647-2653.

Okstad OA, Tourasse NJ, Stabell FB, Sundfaer CK, Egge-Jacobsen W, Risoen PA, Read TD, Kolsto AB.

The bcr1 DNA repeat element is specific to the Bacillus cereus group and exhibits mobile element characteristics.

J Bacteriol. 2004 186:7714-7725.

Parkhill J, Achtman M, James KD, Bentley SD, Churcher C, Klee SR, Morelli G, Basham D, Brown D, Chillingworth T et al:

Complete DNA sequence of a serogroup A strain of Neisseria meningitidis Z2491.

Nature 2000, 404:502-506.

Petrillo M, Silvestro G, Di Nocera PP, Boccia A, Paolella G.

Stem-loop structures in prokaryotic genomes.

BMC Genomics. 2006 7:170

Rocha EP, Danchin A.

Gene essentiality determines chromosome organisation in bacteria.

Nucleic Acids Res 2003, 31:6570-6577.

Rocha EP.

Inference and analysis of the relative stability of bacterial chromosomes.

Mol Biol Evol. 2006, 23:513-522.

Rouquette-Loughlin CE, Balthazar JT, Hill SA, Shafer WM.

Modulation of the mtrCDE-encoded efflux pump gene complex of Neisseria meningitidis due to a Correia element insertion sequence.

Mol Microbiol. 2004 54:731-741.

RicBase Rickettsia genome database

[http://igs-server.cnrs-mrs.fr/mgdb/Rickettsia]

Supply P, Mazars E, Lesjean S, Vincent V, Gicquel B, Locht C.

Variable human minisatellite-like regions in the Mycobacterium tuberculosis genome.

Mol Microbiol. 2000 36:762-771

Tobes R, Ramos JL.

REP code: defining bacterial identity in extragenic space.

Environ Microbiol. 2005 7:225-228.

Tobes R, Pareja E.

Repetitive extragenic palindromic sequences in the Pseudomonas syringae pv. tomato DC3000 genome: extragenic signals for genome reannotation.

Res Microbiol. 2005 156:424-433.

van Belkum A, van Leeuwen W, Scherer S, Verbrugh H.

Occurrence and structure-function relationship of pentameric short sequence repeats in microbial genomes.

Res Microbiol 1999, 150:617-626.

Venditti R, De Gregorio E, Silvestro G, Bertocco T, Salza MF, Zarrilli R, Di Nocera PP.

A novel class of small repetitive DNA sequences in Enterococcus faecalis. FEMS Microbiol Lett. 2007 271:193-201.

Washietl S, Hofacker IL, Stadler PF.

Fast and reliable prediction of noncoding RNAs.

Proc. Natl. Acad. Sci. U. S. A. 2005, 102:2454-2459.

Yang Y, Ames GF.

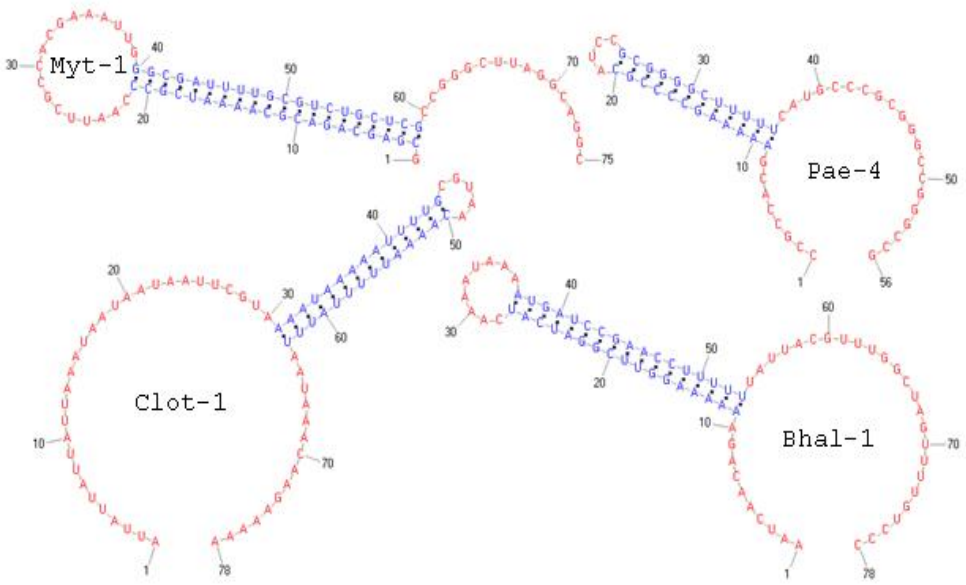
DNA gyrase binds to the family of prokaryotic repetitive extragenic palindromic sequences.

Proc. Natl. Acad. Sci. U S A. 1988, 85:8850-8854.

TABLES AND FIGURES

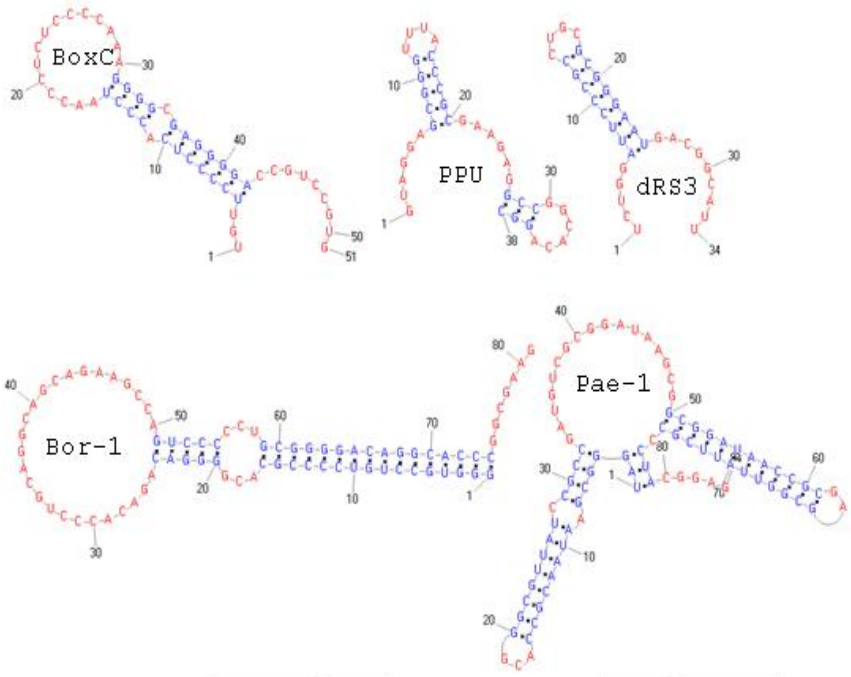
			HMM size	SLS size	TIR	TSD	copy number	A
type-IA	Myt-1	*M. tuberculosis*	72	57-61	20-28	-	70	s
	Pae-4	*P. aeruginosa*	52	29-40	12-13	-	74	s
	Clot-1	*C. tetanii*	74	35	14-15	-	22	s
	Bhal-1	*B. halodurans*	74	45	17-18	-	35	s
type-IB	REP 2	*N. meningitidis*	65	39	12	-	23	s
	REP	*E. coli*	108	35-40	10-13	-	485	s
		S. typhimurium	78				82	s
	BoxC	*E. coli*	50	39	12	-	32	c
	REP	*P. putida*	39	31	-	-	623	-
	dRS3	*N. meningitidis*	33	30	8	-	770	c
	RPE6	*R. conorii*	108	30	8	-	168	-
type-IIA	Bor-1	*B. bronchiseptica*	117	28-32	13	-	220	s
		B. pertussis	93				140	s
	Pae-1	*P. aeruginosa*	84	33	12	-	154	s
	ERIC	*E. coli*	140	69-127	26	2	21	s
		Y. pestis	115				167	s
		V. cholerae	103				80	-
	NEMIS	*N. meningitidis*	46	108-158	26-27	2	250	s
	ATR	*N. meningitidis*	206	183	17	3	13	s
	RUP	*S. pneumoniae*	63	108	26	2	108	-
	Bcr1	*B. anthracis*	167	157	-	5	12	s
	YPAL	*Y. pestis*	136-168	130-169	-	3-25	80	s
	BOX	*S. pneumoniae*	84	100-200	-	1	127	-
	RPE-4	*R. conorii*	100	103	30	-	19	s
	RPE-5	*R. conorii*	115	130	-	-	26	s
	RPE-7	*R. conorii*	123	100-136	-	-	36	s
	Sal-1	*S. typhimurium*	115	150-212	15	3	20	c

Table I. Families of SLS-containing repeated sequences. The size of repeats, TIRs and TSDs are in bp. Repeats may fold into simple (s) or complex (c) structure (column A)



Myt-1	gCGAGCAGACGCAAAATCGCCcaatt	CGCCACGA	aattgGGCGATTTTGCGTCTGCTCGc
Pae-4	gAAAAAGCCCCGC	ATCC	GCGGGGCTTTTTc
Clot-1	tAAATAAAAATTTTG	CGTAA	CAAAATTTTTATTTa
Bhal-1	aAAAAAGGTTCGGATCAT	CAAAATAAA	ATGATCCGAACCTTTTTt
Rep2	CAGGACGATACA	AAAAACGTCGCCCTA	TGTGCCGTCCTG

Figure 1. Type-IA elements. The conserved secondary structures and the consensus sequences of the five type-IA elements are shown.
Complementary bases contributing to the stem of each element are highlighted. Lowercase letters refer to complementary TIR residues present in family members subsets.



EPU	Y	MWTGCCGGATGCGGCGT	RA	ACGCCTTATCCGGC**CTAC**
	Z1	AATGCCTGATGCGC---	TAC	---GCTTATCAGGC**CTAC**
	Z2	RWYGCCTGATGCGACGC	TKGC	GCGTCTTATCAGGC**CTAC**
BoxC	a	TCCCCTCACCCT	AACCCTCTCCCCAAA	GGGGCGAGGGGA
	b	TCCCCTCACCCC	GGCCCTCTCCCCAAA	GGGGCGAGGGGG
PPU	a	MYGGCCTCTTCGCGGGt	AA	aCCCGCTC**CTAC**
	b	CCGGCCCTTCCGCGGGT	RA	ACCCGCTC**CTAC**
dRS3	a	YCGTCATTCCCGC	GCAG	GCGGGAATCYAGG
	b	YCGTCATTCCCAC	GAAA	GTGGGAATCYAGG
RPE6	a	ATGTCATTCCCGC	GTAG	GCGGGAATCYAGW
	b	ATGTCATTCCCGC	GAAA	GCGGGAATCYAGW
Bor-1	a	GTGTGCCTGTCCC	NGCG	GGGACAGGCACAC
	b	G-GTGCCTGTCCC	NGCG	GGGACAGATAC-C
Pae-1	a	GGCGGATAACGCC	CGT	GGCGTTATACGCC**CTAC**
	b	GGCGGATAACCGC	GA	GCGGTTATTCGCC**CTAC**

Figure 2. Type-IB elements. The conserved secondary structures and the consensus sequences of type-IB elements are shown. Highlighting and lowercase letters are as in the legend to Fig. 1. According to IUB codes Y=C or T; R=A or G; H=A, C or T; K=G or T; M=A or C; S=G or C; W=A or T; N=any nucleotide. Dashes were introduced to maximize homologies among families units. The terminal CTAC residues in EPU, PPU and Pae-1 repeats are in boldface.

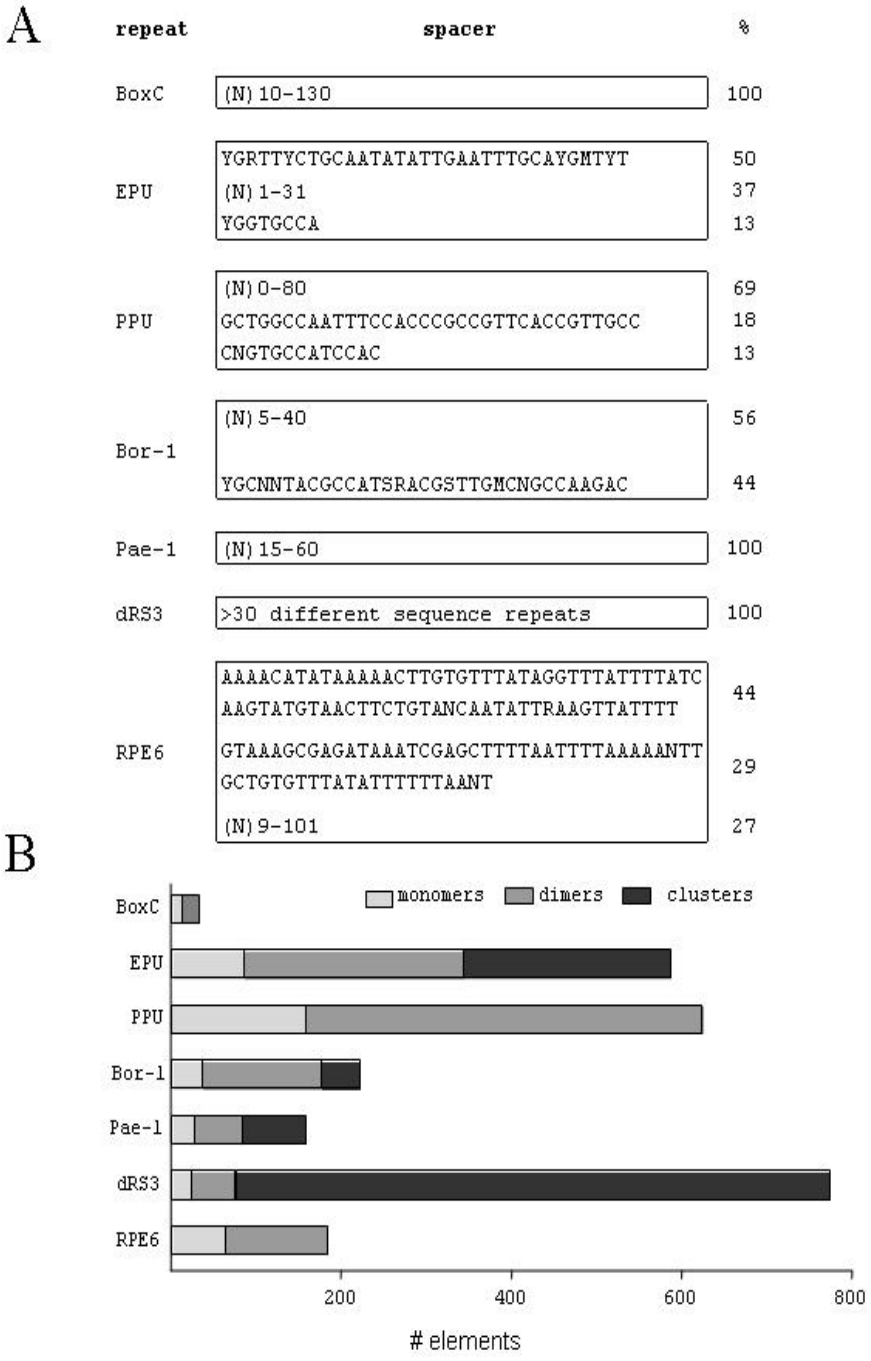


Figure 3. Type-IB dimeric elements. A) Spacer sequences separating members of type-IB element families arranged as dimers. The relative amount of each spacer type is shown. B) relative amount of type-IB elements found as single units, dimers or arranged in clusters.

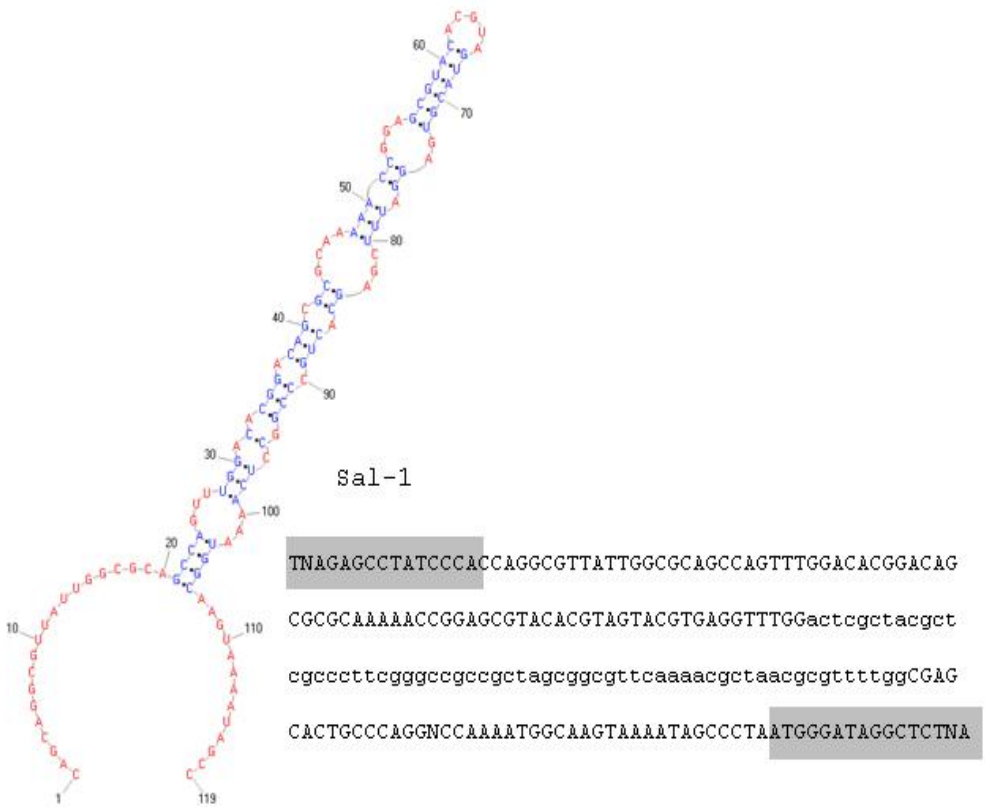
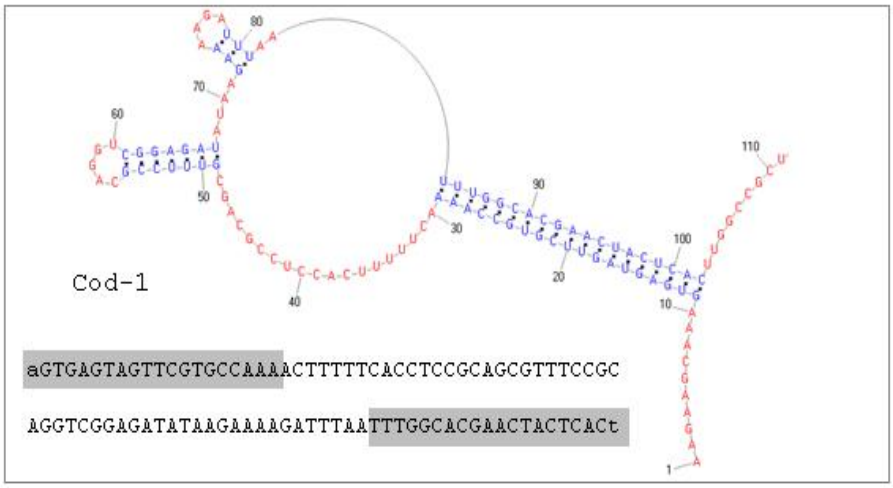


Figure 4. Sequence and secondary structure of Sal-1 repeats. TIRs are highlighted. Lowercase letters denote the 61 bp long DNA segment inserted in a subset of elements.



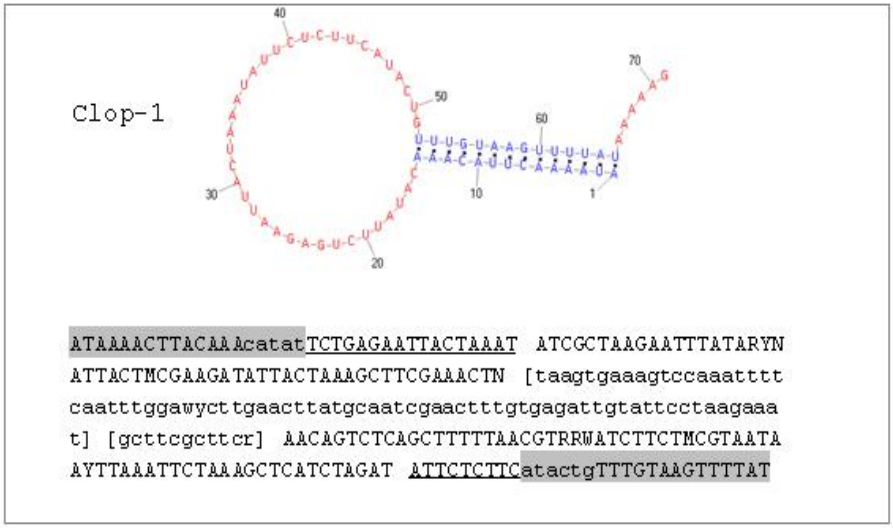


Figure 5. Sequence and secondary structure of Cod-1 and Clop-1 repeats. TIR residues are highlighted. Unit-length Clop-1include the TIRs and the underlined residues. Additional body segments may be in turn interrupted by alternative segments of foreign DNA (bracketed residues; see Text).

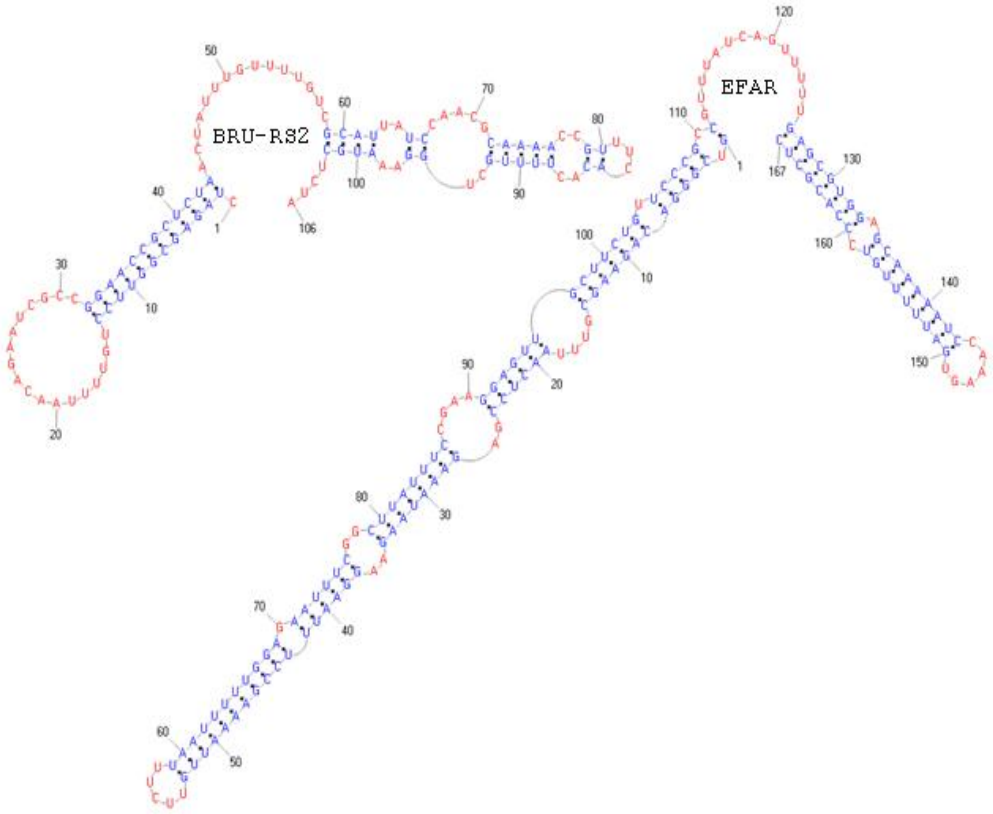


Figure 6. SLSs formed by type IIB elements. Two-hairpins structures formed by EFAR and BRU-RS2 repeats

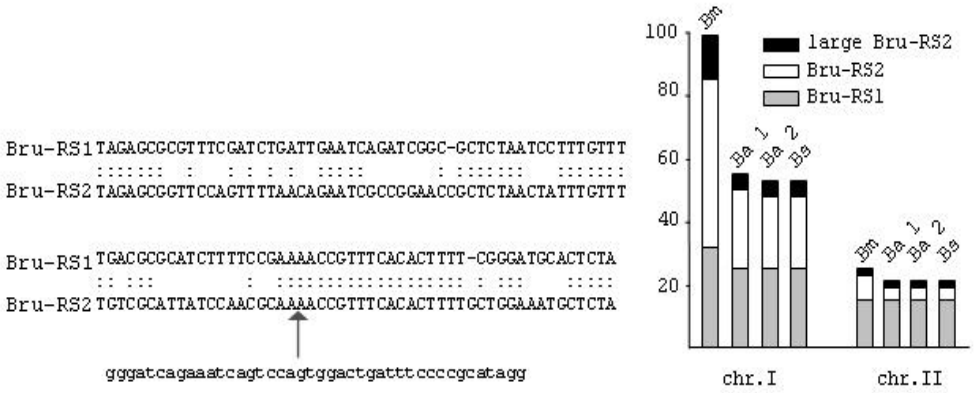


Figure 7. Bru-RS1 and Bru-RS2 sequence alignment. Homology is highlighted by dots. The sequence and the site of insertion of a 42 bp long DNA segment interrupting Bru-RS2 repeats are shown. The number of Bru-RS1 and Bru-RS2 elements found in the two chromosomes of different Brucellae is diagrammed. Bm, B. melitensis strain 16M; Ba 1, B. abortus strain 2308; Ba 2, B. abortus biovar 1 strain 9-941; Bs, B. suis strain 1330.

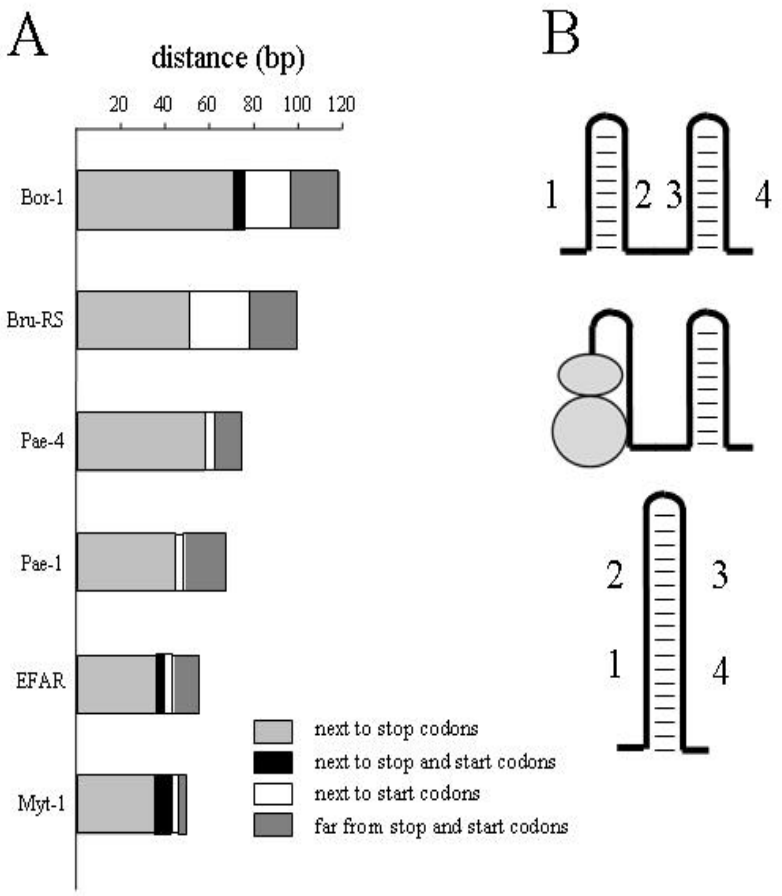


Figure 8. (A) Interspersion with annotated ORFs of a subset of repeated DNA families analyzed in this study. Most elements of diagrammed families are close to (<30 bp) from the stop codon of flanking ORFs.
(B) A ribosome temporarily stalled at the stop codon may interfere with the folding of either one of the two SLSs of both dimeric type-IB and type IIB elements. This situation, in the case of type-IB elements may provoke the formation of a single large SLS, in which one monomer fold to each other.

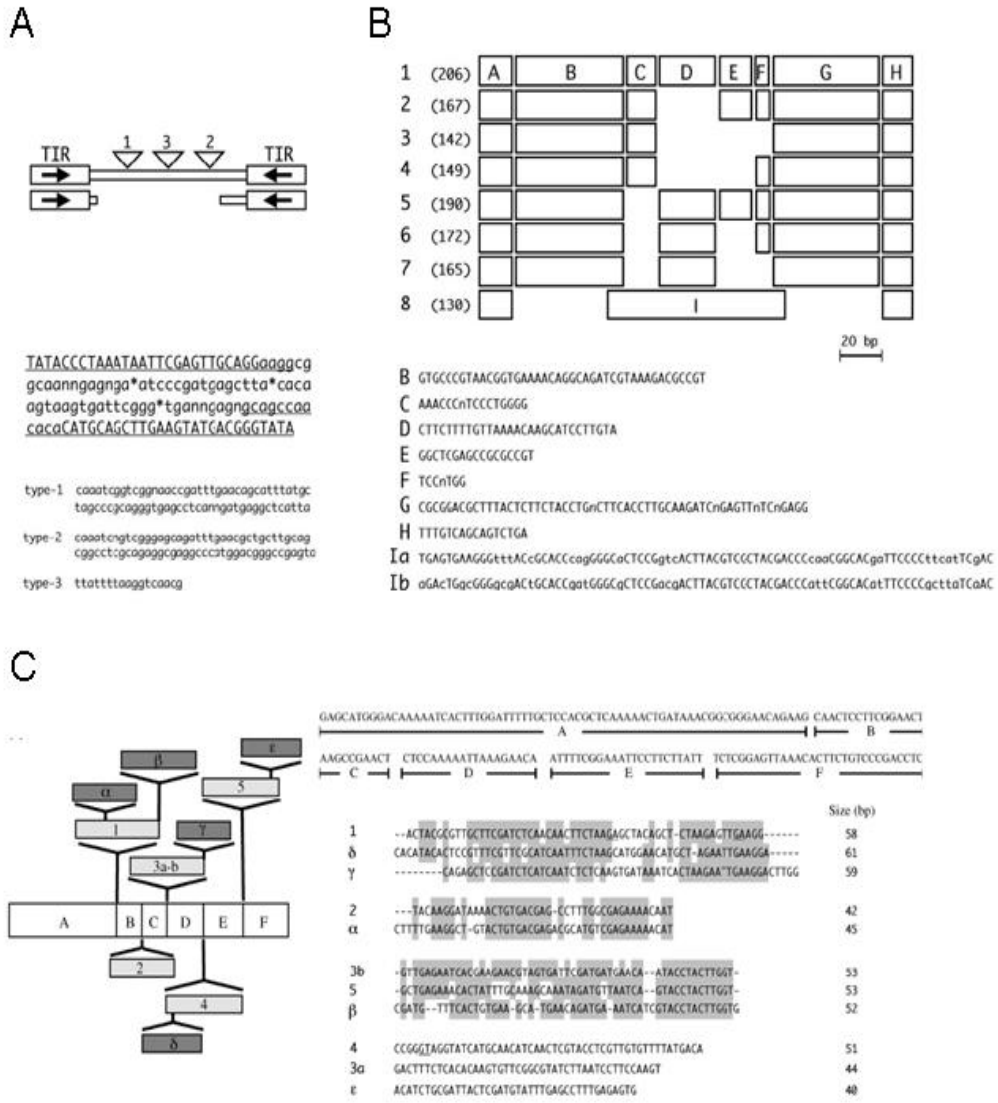


Figure 9. Structural organization of ERIC, YPAL and EFAR repeats. (A) Boxed arrows mark the TIRs of ERICs. Triangles mark type 1-to-type 3 insertions interrupting ERICs. In the bottom of panel A is reported the consensus sequence of the 127-bp unit-length ERICs. TIR residues are in capital letters. Underlined residues mark sequences conserved in the internally rearranged 70-bp-long ERICs. The integration sites of type 1-to-type 3 insertions are denoted by asterisks.The consensus sequences of the three types of intervening sequences found to interrupt ERICs are shown.

(B) The nine modules labeled A to I found in YPAL elements are shown. Numbers to the left (1 to 8) denote YPAL subfamilies. The length in bp of the elements of each subfamily is given in parenthesis. The consensus sequence content of the nine modules is shown at the bottom. Two versions of the module I (Ia and Ib) are reported. Uppercase residues denote sequence identity.

(C) The A–F modules, the sites of integration of primary insertions 1–5, and secondary insertions a–e of EFAR repeats are shown; 3a and 3b are inserted at the same site. Beside the cartoon is reported consensus sequences of EFAR modules and insertions. Sequence relatedness is highlighted. Dashes have been introduced to maximize homologies. Underlined residues mark the sites of integration of secondary insertions. Numbers to the right denote the size in bp of each insertion.

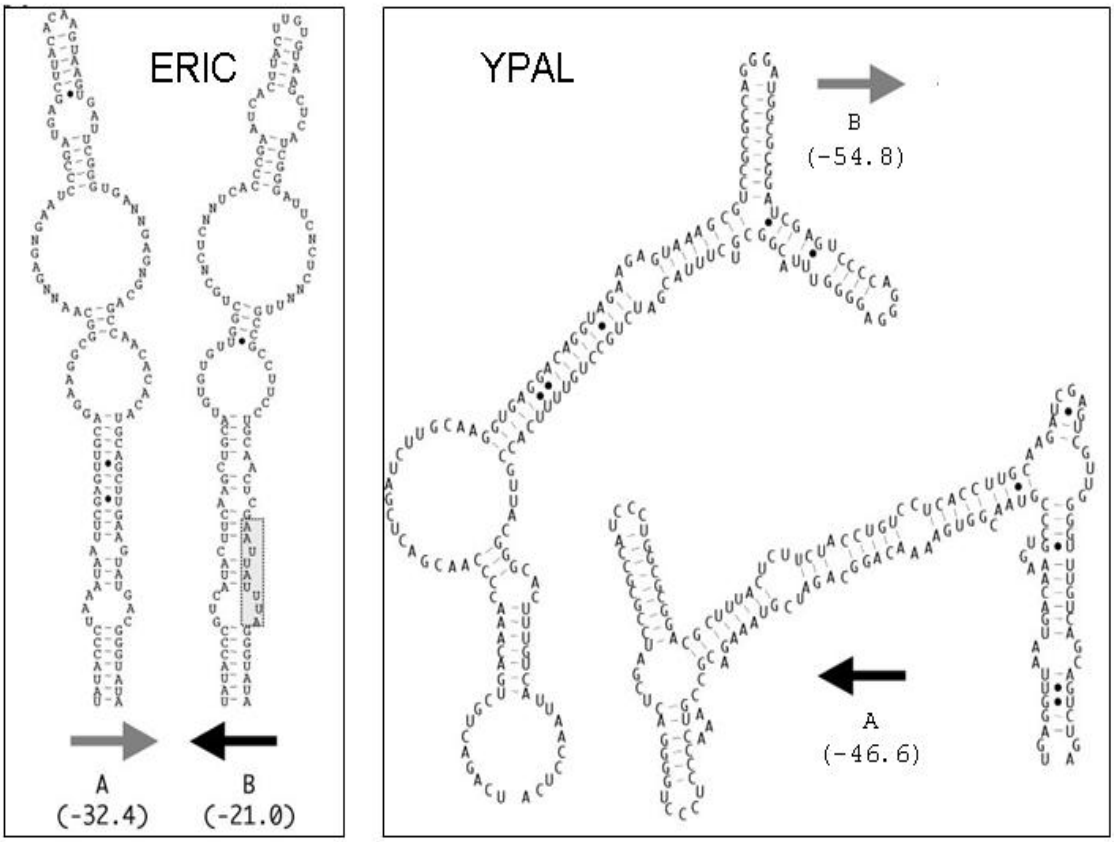


Figure 10. Secondary structure of ERIC and YPAL repeats. (A) Predicted RNA foldings and relative calculated free energies at 37°C of unit-length ERIC consensus sequences inserted in A and B orientations is reported. Non-Watson-Crick base pairings are highlighted by dots. The hypothesized cleavage site for RNase E, present in B-oriented ERICs, is boxed.
(B) Hairpin structures formed by a representative 167 bp long YPAL element inserted in the mRNA in the two possible A and B orientations are shown. GU pairing is highlighted by dots. The Gibbs free energy of each hairpin is indicated

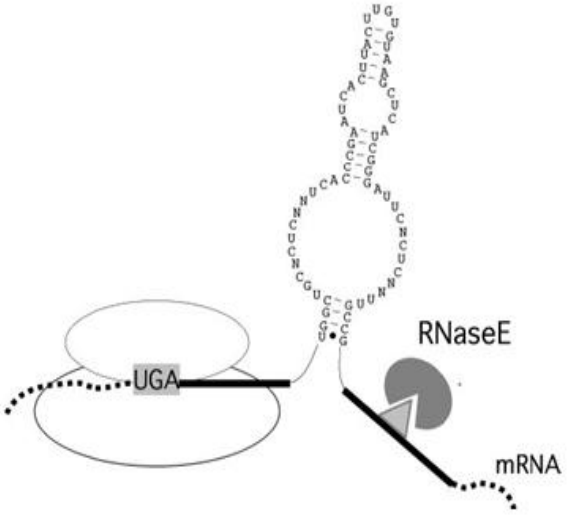


Figure 11. Ribosomes interfere with folding of ERIC-positive RNA. In mRNA-spanning B-oriented elements that are inserted close to the translational stop codon, the translating ribosome covers most of the ERIC left-hand TIR, unmasking the RNase E site (sketched as a triangle) located in the ERIC right-hand TIR.

WORKS *IN EXTENSO*

JOURNAL OF BACTERIOLOGY, Dec. 2005, p. 7945–7954 Vol. 187, No. 23
0021-9193/05/$08.00+0 doi:10.1128/JB.187.23.7945–7954.2005

Enterobacterial Repetitive Intergenic Consensus Sequence Repeats in Yersiniae: Genomic Organization and Functional Properties

Eliana De Gregorio,[1] Giustina Silvestro,[1] Mauro Petrillo,[2] Maria Stella Carlomagno,[1] and Pier Paolo Di Nocera[1]*

Dipartimento di Biologia e Patologia Cellulare e Molecolare, Facoltà di Medicina, Università Federico II, Via S. Pansini 5, 80131 Napoli, Italy,[1] and CEINGE Biotecnologie Avanzate s.c.a.r.l., Via Comunale Margherita n. 482, 80131 Napoli, Italy[2]

Received 29 July 2005/Accepted 21 September 2005

Genome-wide analyses carried out in silico revealed that the DNA repeats called enterobacterial repetitive intergenic consensus sequences (ERICs), which are present in several *Enterobacteriaceae*, are overrepresented in yersiniae. From the alignment of DNA regions from the wholly sequenced *Yersinia enterocolitica* 8081 and *Yersinia pestis* CO92 strains, we could establish that ERICs are miniature mobile elements whose insertion leads to duplication of the dinucleotide TA. ERICs feature long terminal inverted repeats (TIRs) and can fold as RNA into hairpin structures. The proximity to coding regions suggests that most *Y. enterocolitica* ERICs are cotranscribed with flanking genes. Elements which either overlap or are located next to stop codons are preferentially inserted in the same (or B) orientation. In contrast, ERICs located far apart from open reading frames are inserted in the opposite (or A) orientation. The expression of genes cotranscribed with A- and B-oriented ERICs has been monitored in vivo. In mRNAs spanning B-oriented ERICs, upstream gene transcripts accumulated at lower levels than downstream gene transcripts. This difference was abolished by treating cells with chloramphenicol. We hypothesize that folding of B-oriented elements is impeded by translating ribosomes. Consequently, upstream RNA degradation is triggered by the unmasking of a site for the RNase E located in the right-hand TIR of ERIC. A-oriented ERICs may act in contrast as upstream RNA stabilizers or may have other functions. The hypothesis that ERICs act as regulatory RNA elements is supported by analyses carried out in *Yersinia* strains which either lack ERIC sequences or carry alternatively oriented ERICs at specific loci.

Transposable elements (TEs) are widely distributed in prokaryotic and eukaryotic genomes. TEs are broadly divided into two classes according to their transposition intermediates. Class 1 elements transpose by means of an RNA intermediate and feature either long terminal direct repeats or a poly(A) tract at one end. Class 2 elements transpose by means of a DNA intermediate, and most have terminal inverted repeats (TIRs). Integration of most TEs frequently determines the duplication of target sites of fixed lengths (20).

DNA repeats which recall class 2 elements in terms of the presence of TIRs but have no coding capacity are found in many organisms. These nonautonomous mobile elements are commonly referred to as MITEs (miniature inverted transposable elements). First recognized as a predominant sequence type in plants, MITEs have been subsequently identified in many invertebrate and vertebrate genomes (14). A few MITE families have been characterized in archaeal genomes (5, 34) and in eubacteria. *Streptococcus pneumoniae* contains ~100 copies of a 107-bp-long miniature insertion sequence called the repeat unit of pneumococcus (RUP) (29). The 106- to 158-bp-long DNA elements known as Correia or neisseria miniature insertion sequences (NEMIS) make up 1 to 2% of the genome in pathogenic neisseriae (6, 10, 22, 24). RUP and NEMIS feature similar TIRs, and both induce the duplication of the TA dinucleotide upon genomic insertion. Most NEMIS are cotranscribed with neighboring genes, and NEMIS-positive mRNAs fold into hairpins formed by NEMIS termini, which are targeted by RNase III (9, 11). Genome-wide analyses carried out in silico predict that the expression levels of 80 to 100 *Neisseria meningitidis* genes may be tuned by RNase III-dependent processing at NEMIS RNA hairpins (10, 11).

The 127-bp-long elements known either as intergenic repeat units (38) or as enterobacterial repetitive intergenic consensus sequences (ERICs) (17) structurally recall NEMIS and RUP repeats. ERIC families are made up by 20 to 30 elements in both *Escherichia coli* and *Salmonella enterica* serovar Typhimurium. In this report, we show that ERICs, as anticipated by early genomic analyses by Bachellier and coworkers (3), are overrepresented in yersiniae. In silico analyses performed on the wholly sequenced *Yersinia pestis* CO92 (12, 30) and *Yersinia enterocolitica* 8081 (www.sanger.ac.uk/Projects/Y_enterocolitica) strains establish that ERICs constitute a major DNA family in yersiniae. ERICs are (or have been) mobile DNA sequences which also belong to the MITE superfamily. Most of the 247 elements found in *Y. enterocolitica* are inserted at close distance from flanking coding regions, and it is likely that many are transcribed into mRNA. In this paper, we show that, according to their orientations and relative positions within the mRNA, transcribed ERICs may impede or accelerate the decay of specific mRNA segments.

* Corresponding author. Mailing address: Dipartimento di Biologia e Patologia Cellulare e Molecolare, Facoltà di Medicina, Università Federico II, Via S. Pansini 5, 80131 Napoli, Italy. Phone: 0039-81-7462059. Fax: 0039-81-7703285. E-mail: dinocera@unina.it.

MATERIALS AND METHODS

Bacterial strains and growth conditions. The *Y. enterocolitica* strain Ye161 (serogroup O8) was kindly provided by Ida Luzzi at the Istituto Superiore di Sanità, Rome. The *Y. enterocolitica* strains Ye24 (serogroup O8) and Ye25

TABLE 1. Oligonucleotides used to monitor ERIC-positive RNAs[a]

Primer	Sequence
pex. cheW	TTCGTGACGGTTGCTAGTCCTGCCA
pex. trpB	CCCGAACTCGCCAAAATAGGGATTC
pex. uncE	CGATAAAGAACTGTGTACGCAGCAG
pex. lpdA	GGATACAACCGACATTCAGGCACAC
cheB-for	atttaggtgacactatagaaAACTATCAGGTGCGTATTCATGATG
cheB-rev	taatacgactcactatagggGCTTCGTTTTGTGCAATGGTATAAG
cheY-for	atttaggtgacactatagaaTGGTAGACGATTTTTCGACCATGCG
cheY-rev	taatacgactcactatagggTCGGCATGTTCCAGTCAGAAACCAC
panB-for	atttaggtgacactatagaaGCTAACCAGCTATTGAAAGATGCTC
panB-rev	taatacgactcactatagggGAATATACAGCTTAATGGCAGCACG
panC-for	atttaggtgacactatagaaATTGAAACTTTGCCACTGTTACGCC
panC-rev	taatacgactcactatagggATACTCACGACGACAACATCGGCAC

[a] The oligonucleotides used as primers in RNA extension assays are listed at the top. The pairs of 45-mers shown at the bottom have been used to obtain by PCR DNA amplimers in which copies of the bacteriophage T7 promoter (underlined residues) direct the synthesis of antisense RNA probes (Fig. 6). Uppercase residues mark *Y. enterocolitica* sequences.

(serogroup O9) and the *Y. kristensenii* SS47 strain were provided by Francesca Berlutti at the Istituto di Igiene of the University La Sapienza, Rome. *Yersinia* cells were grown in LB broth at 28°C. When needed, exponentially growing Ye161 cells were exposed either for 12 min to rifampin (final concentration, 200 μg/ml) or for 30 min to chloramphenicol (final concentration, 50 μg/ml) before harvesting.

RNA analyses. Total bacterial RNA was purified on an RNeasy column (QIAGEN). Transcripts spanning the *cheW* (open reading frame [ORF] YE2576), *trpB* (ORF YE2213), *uncE* (ORF YE4221), and *lpdA* (ORF YE0702) genes were monitored by RNA extension analyses as reported previously (9) by using as primers the pex.cheW, pex.trpB, pex.uncE, and pex.lpdA oligonucleotides, respectively. The sequences of the four primers are reported in Table 1. Reverse transcriptase-PCR (RT-PCR) analyses were carried out by reverse transcribing 200 nanograms of total *Y. enterocolitica* RNA by random priming. The resulting cDNA was amplified by using pairs of gene-specific oligonucleotides (Table 2). The melting temperature (T_m) of each oligomer (Table 2) was determined by using the Oligo 4.0 primer analysis software (35). In several instances, RT-PCR coamplifications were carried out with alternative pairs of primers. One oligonucleotide of each pair had been ^{32}P end labeled at the 5′ terminus with the polynucleotide kinase. Comparable yields of amplimers were obtained by labeling either forward or reverse cistron-specific primers. To adequately monitor gene-specific RNA levels by RT-PCR, the cDNA was amplified under nonsaturating cycling conditions, and ad hoc low-cycle-number (6 to 12 cycles) PCR analyses were performed for each set of coamplified genes. Amplimers were electrophoresed onto 6% polyacrylamide-8 M urea gels and quantitated by phosphorimagery.

For RNase protection assays, uniformly ^{32}P-labeled RNA probes were obtained by transcribing in vitro linear DNA templates as described previously (9). Templates were obtained by PCR amplification of Ye161 DNA with the 45-mers shown in Table 1. Within each pair, one oligomer included the sequence of the T7 RNA polymerase promoter in the 5′ region. Twenty micrograms of total RNA were mixed with ^{32}P-labeled antisense RNA probes in 30 μl of hybridization buffer (75% formamide, 20 mM Tris [pH 7.5], 1 mM EDTA, 0.4 M NaCl, 0.1% sodium dodecyl sulfate). Samples were incubated at 95°C for 5 min, cooled down slowly, and kept at 45°C for 16 h. After a 60-min incubation at 33°C with RNase T_1 (2 μg/ml), samples were treated with proteinase K (50 μg/ml) for 15 min at 37°C, extracted once with phenol, precipitated with ethanol, resuspended in 80% formamide, and loaded onto 6% polyacrylamide-8 M urea gels.

Computer analysis. *E. coli* ERIC sequences were used as queries in BLAST searches (2) to fetch homologous DNA segments from the genomes of the *Y. pestis* CO92 (30) and KIM (12) strains and from *Y. enterocolitica* 8081 (www.sanger.ac.uk/Projects/Y_enterocolitica). Species-specific queries allowed the identification of *Yersinia* ERICs evolutionarily distant from *E. coli* homologs. Retrieved DNA sequences were aligned with the CLUSTAL W program (41). Consensus sequences from multiple alignments of ERIC family members were established with the program CONS of the EMBOSS package. Secondary structure modeling was done using the Mulfold program (www.bioinfo.rpi.edu/applications/mfold), which predicts RNA secondary structure by free-energy minimization (45).

RESULTS

Genomic and structural organization of ERICs in yersiniae. The genus *Yersinia* includes 11 species, 3 of which are pathogenic to humans (4). The enteropathogens *Y. pseudotuberculosis* and *Y. enterocolitica* are widely found in the environment. In contrast, *Y. pestis* is a highly virulent blood-borne pathogen which is transmitted by fleas and which rapidly evolved from *Y. pseudotuberculosis* (1, 44). In silico analyses carried out on wholly sequenced strains showed that *Yersinia* chromosomes are peppered by ERIC repeats. By looking only at elements carrying both TIRs (Fig. 1), we found 247 and 167 ERICs in the genomes of the *Y. enterocolitica* 8081 and *Y. pestis* CO92 strains, respectively (Table 3). About 90% of the elements are scattered throughout the chromosome of either species as single-copy insertions. The remaining 10% is made up by clusters in which two to five elements are organized in head-to-tail configuration. On the whole, 235 and 151 ERIC-positive sites were identified in Y. *enterocolitica* 8081 and Y. *pestis* CO92 strains, respectively (Table 3). In contrast, the *Y. pestis* CO92 strain contains several moderately abundant families of insertion sequences (ISs) (65 copies of IS*1541*, 44 copies of IS*100*, 8 copies of IS*1661*, and 21 copies of IS*285*; see reference 7), while we found only 3 copies of IS*1541* in the Y. *enterocolitica* 8081 genome by BLAST analyses (not shown).

Unit-sized ERICs are 127 bp in length (Fig. 1). Shorter elements measure ~70 bp, and all lack a 50-bp-long internal segment. Larger elements are interrupted at specific sites by three different types of DNA insertions (Fig. 1). Type 1 and type 2 insertions have been found also in some *E. coli* ERICs (37), while type 3 insertions seem to be present only in yersiniae.

Y. pestis and *Y. enterocolitica* genomes both measure ~4.6 Mb. However, extensive genetic remodeling makes *Y. pestis* a species evolutionarily distant from other yersiniae (44). *Y. pestis* ERICs are fewer and exhibit more size heterogeneity than *Y. enterocolitica* elements (Fig. 1A). The *Y. pestis* CO92 and the *Y. enterocolitica* 8081 chromosomes share only 37 syntenic regions carrying ERIC repeats. Elements have the same size only in one-third of the cases. In the other instances, unit-length elements found in *Y. pestis* are replaced by either shorter or insertion-tagged ERICs in *Y. enterocolitica*, and vice versa (not shown), plausibly as a result of recombination events between ERIC family members.

The insertion of ERIC induces a 2-bp target site duplication. Several syntenic regions identified in *Y. enterocolitica* and *Y. pestis* carry an ERIC element in the former species only. ERICs terminate at either side with the dinucleotide TA. At many *Y. pestis* empty sites, ERIC is replaced by one copy of the dinucleotide (Fig. 2). The duplication of the dinucleotide TA is a hallmark of eukaryotic MITEs and is a feature shared by known eubacterial MITEs (24, 29). TA empty sites have been identified both in *Y. enterocolitica* and in *Y. pestis* (Table 3). This indicates that the mobilization of ERICs still occurred after the speciation of yersiniae into *Y. enterocolitica* and *Y. pseudotuberculosis*, from which *Y. pestis* eventually derived (1, 44).

Differences in the distributions of ERICs between the *Y.*

TABLE 2. Oligonucleotides used for PCR and RT-PCR analyses

Gene	Primer[a]			
	Direct		Reverse	
	Sequence	T_m (°C)	Sequence	T_m (°C)
cheW	CGAAACGGTAGGACAAGAATTCCTG	59.1	GGAACAATAACGCCGCGTAAGTTAG	59.2
cheA	TCATTTTACCATTGAACGCCGTAAT	57.7	CAGAAATACCTGGAACCTTGCGATA	57.5
cheA	AGGCATTGTTGTGATTCTACAAAGC	55.7	ACGACTCCATCATCAGCCGTAACAC	60.2
trpC	TTGAACGCTATGTTTTGGATAATGG	56.3	CAATCTTTTGGGGATCTTTAATGCC	58.1
trpB	ATCCCTATTTTGGCGAGTTCGGGGG	56.2	CAGAGCGGTTGGACGCCCAGCATAG	58.0
phoU	CATATTTCCGGCCAGTTCAATGCAG	62.3	CTTTTTATCACCCTCGATGACGCGC	63.5
phoT	GTTATAGCTTGTCGGGTGGGCAGCA	63.9	TTCTGCTGTGGTGCGGTAAACAGGG	64.8
phoT	AGGTATCGCCATTCGTCCAGATGTG	62.0		
cheY	TGGTAGACGATTTTTCGACCATGCG	62.8	TCGGCATGTTCCAGTCAGAAACCAC	63.0
cheB	AACTATCAGGTGCGTATTCATGATG	64.2	GCTTCGTTTTGTGCAATGGTATAAG	66.2
cheB	AATTACCGTGAAGGAAGCAGAAGAC	66.2		
glgA	CGTATGTTCTGAGCTATTCCCGTTG	58.1	AGCAAAGGTATCAATCTCCCTGACC	57.9
glgA	TACGGCATGGAGGGGTTGCTACAAG	60.5	AAGCCATACAACTGTGTTAGCCCAC	57.5
glgC	GTAGCCACGGTATGACCATGAACTC	57.7	GCAGCAGTAATGTGGAATCAATGGT	58.2
glgC			GATAAAAACGCTTGCTGTCTTCTTC	55.0
manX	GCAATATGCCAAAGACCGGGTCATG	64.0	TTCGGACTTCCAGTTCAATACCGCG	63.9
manY	ATTTCATCGTTGCCTGTATCGCCGG	63.7	ACCAAAGTACAGGCGACGAGGGGAC	64.1
manY			CAGCCGAGCGCGATCATTTCTAAGG	65.5
panB	GCTAACCAGCTATTGAAAGATGCTC	54.9	GAATATACAGCTTAATGGCAGCACG	56.3
panC	ATTGAAACTTTGCCACTGTTACGCC	59.5	ATACTCACGACGACAACATCGGCAC	61.0
pstS	TGCGGCAAAAGGTGTAGACTGGAGC	64.2	AATCTGAGTTTTCCAGGCAGCACGG	63.3
pstC	ATAAGCCGACAATCAAAGCACCGAG	61.4	CCACAGGAAAGCCCAACCAAATTTC	62.5
argB	TGGCAGTCTGATGAACGTAAATGCC	61.1	CATTCACTTTAACCACCATGCCGTC	60.4
argH	TCAGGCGGCAGATCAGCGTTTTAAG	64.2	GTTGCTGTTCGGCCGTCGTTAAAAC	63.4

[a] In some instances, the level of cistron-specific transcripts was monitored by using novel primers.

pestis CO92 and *Y. enterocolitica* 8081 genomes reflect species-specific, rather than strain-specific, variations. The *Y. pestis* strains CO92 (30) and KIM (12) belong to different biovars, have been responsible for the spreads of different plague epidemics, and show a remarkable amount of genome rearrangement (12). However, 155/157 ERIC-positive sites found in the CO92 strain are perfectly conserved, as revealed by BLAST analyses, in the KIM strain, and the remaining two sites vary only in terms of the number of tandemly inserted elements.

ERICs are cotranscribed with flanking genes. Genome-wide surveys revealed that 137 ERICs are inserted at close distance (≤50 bp) from either the start or the stop codons of *Y. enterocolitica* 8081 ORFs (not shown). This suggests that most elements are cotranscribed with flanking genes into mRNAs.

To investigate the issue, we first checked that ERIC-positive regions found in the 8081 strain were conserved in the *Y. enterocolitica* Ye161 strain. Subsequently, the corresponding ERIC-positive mRNAs synthesized in this strain were monitored by primer extension analyses (Fig. 3). The major products of extension of both *lpdA* and *uncE* transcripts extended beyond ERIC (Fig. 3). In contrast, extension products of both *cheW* and *trpB* transcripts were found to terminate at multiple sites within ERIC sequences (Fig. 3). The same pattern was obtained with different RNA preparations and reverse transcriptase batches. The multiple extension products detected may denote cleavage of *cheW* and *trpB* mRNAs at ERIC sequences.

ERICs flanking the *lpdA* and *uncE* genes are both inserted in the same orientation, which from here on we will arbitrarily refer to as the A orientation. In contrast, elements flanking the *cheW* and *trpB* genes are inserted in the alternative B orientation. About 110 ERICs are found, within a −50- to +45-bp distance range, downstream from the stop codons of annotated ORFs in the 8081 strain (Fig. 4). Curiously, most ERICs which either overlap or are inserted next to (0- to +6-bp distance range) stop codons are B-oriented elements. In contrast, ERICs located at larger distances from ORFs are predominantly A-oriented elements (Fig. 4). The orientation dependence rule works for elements located between unidirectionally transcribed ORFs as for elements separating convergently transcribed ORFs. A privileged orientation relative to the distance from translational stop codons was similarly displayed by ERICs found in the *Y. pestis* CO92 strain (not shown).

To investigate the functional significance of these observations, several pairs of *Y. enterocolitica* genes transcribed in the same direction, but separated by either A- or B-oriented ERICs, were selected for comparative RNA quantitations. Elements analyzed measured all 127 bp and exhibited 94% sequence similarity. Total RNA from Ye161 cells was reverse transcribed into cDNA, and the latter was subsequently am-

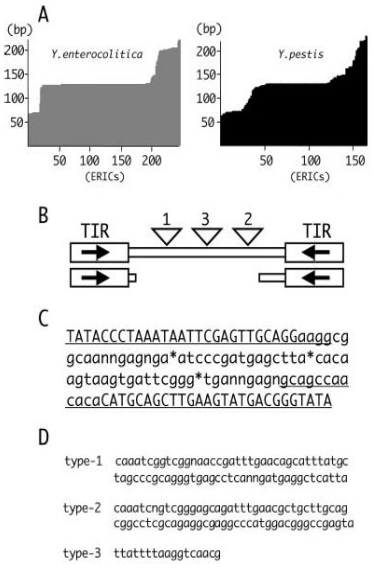


FIG. 1. ERIC elements in yersiniae. (A) ERIC-sized classes in *Y. enterocolitica* 8081 and *Y. pestis* CO92 strains. The number of elements carrying both TIRs found and their sizes in base pairs are indicated. (B) Structural organization of ERICs. Boxed arrows mark the TIRs. Triangles mark type 1-to-type 3 insertions interrupting ERIC sequences. (C) The consensus sequence of the 127-bp unit-length ERICs is shown in the A orientation. TIR residues are in capital letters. Underlined residues mark sequences conserved in the internally rearranged 70-bp-long ERICs. The integration sites of type 1-to-type 3 insertions are denoted by asterisks. (D) The consensus sequences of the three types of intervening sequences found to interrupt ERICs are shown.

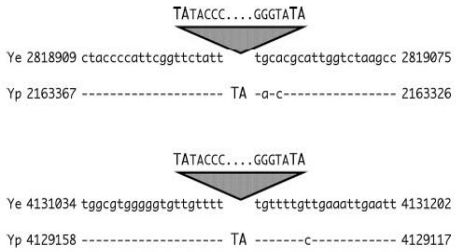


FIG. 2. Filled and empty ERIC sites. Homologous DNA regions from the *Y. enterocolitica* 8081 (Ye) and *Y. pestis* CO92 (Yp) strains are aligned. Numbers refer to genome residues; dashes denote sequence identities. The duplication of the TA target site at ERIC termini is highlighted.

plified by using different sets of primers. As evidenced by the detection of large mRNA segments, elements selected are cotranscribed with flanking genes (Fig. 5A). To monitor the relative abundances of RNA species corresponding to upstream and downstream cistrons, the cDNA was coamplified with pairs of cistron-specific oligomers (Table 2) under nonsaturating cycling conditions (Fig. 5B). Radiolabeled amplimers were separated electrophoretically, and their amounts were quantitated by phosphorimagery. Data obtained with alternative sets of primers were fairly comparable, ruling out technical artifacts. By looking at transcriptional units spanning B-oriented ERICs, we found that downstream gene transcripts accumulated ~4-fold more abundantly than upstream gene transcripts. In contrast, except for the *glgC-glgA* barrier, the levels of gene transcripts flanking A-oriented ERICs were comparable (Fig. 5C). Differences in the downstream/upstream gene transcript ratios measured at intercistronic barriers carrying A-oriented (*panB-panC*) and B-oriented (*cheY-cheB*) ERICs were confirmed by RNase protection experiments and magnified when de novo RNA synthesis was blocked by treating *Yersinia* cells with rifampin (Fig. 6). Both *panB* and *panC* transcripts, which are quite abundant in steady-state RNAs, were no longer detected after exposure of *Y. enterocolitica* cells to rifampin (Fig. 6B, compare lanes 20 and 21). By contrast, the difference in the steady-state levels of *cheY* and *cheB* transcripts made it still possible to detect *cheY* RNA sequences in rifampin-treated cells (Fig. 6A, compare lanes 9 and 10).

TABLE 3. ERIC elements in wholly sequenced *Y. enterocolitica* 8081 and *Y. pestis* CO92 strains

ERIC element or copy no.	No. in:	
	Y. enterocolitica 8081	*Y. pestis* CO92
Copy no.	247	167
ERIC$^+$ sites	235	151
Single-copy inserts	225	148
Type I insertions	13	11
Type II insertions	12	2
Type III insertions	5	35
TA empty sites	23	3

Data signal that the segmental stabilities of RNAs spanning A-oriented and B-oriented elements were substantially different.

Heterogeneity of ERIC-positive loci among yersiniae. The conservation of ERIC sequences in *Y. enterocolitica* was monitored by PCR-driven surveys. The Ye161 and Ye24 strains and the sequenced 8081 strain all belong to the O8 serogroup. It is therefore not surprising that 24/24 ERIC-positive loci analyzed (including those shown in Fig. 5) were conserved in the three strains (data not shown). In contrast, genetic variations at specific loci spanning ERIC sequences found in the 8081 strain were identified in Ye25, a serogroup O9 *Y. enterocolitica* strain, as well as in the YkSS47 strain of the apathogenic *Yersinia kristensenii* species and exploited for comparative RNA analyses. In Ye161, *cheA* and *cheW* genes are separated by a B-oriented ERIC, and *cheW* transcripts are ~5-fold more abundant than *cheA* transcripts. The difference is abolished in YkSS47 cells (Fig. 7). Sequence analysis showed that the YkSS47 *cheA-cheW* region did not experience the insertion of ERIC DNA. In Ye161, *argB* and *argH* genes are separated by a B-oriented ERIC inserted immediately downstream from the *argB* stop codon. In Ye25, in contrast, the two genes are separated by an A-oriented ERIC inserted 10 bp downstream from the *argB* stop codon. Changes in the position and the orientation of ERIC are associated with significant differences

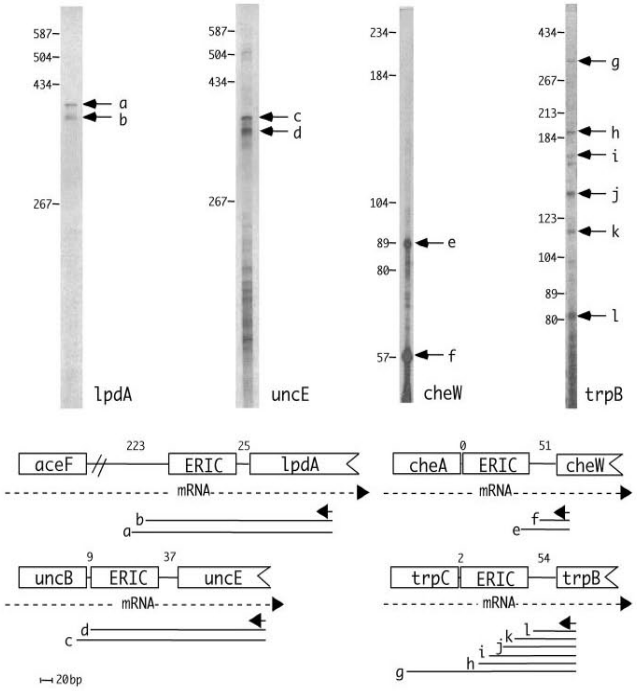


FIG. 3. Primer extension analyses of ERIC-positive transcripts. Primers that had been ^{32}P labeled at the 5′ end and were complementary to *lpdA*, *uncE*, *cheW*, and *trpB* transcripts were hybridized to total RNA (5 μg) derived from the *Y. enterocolitica* Ye161 strain. Annealed primer moieties were extended in the presence of nucleoside triphosphates by avian myeloblastosis virus reverse transcriptase. Reaction products were electrophoresed on 6% polyacrylamide-8 M urea gels. Major reaction products labeled "a" to "l" are marked by arrows. Numbers to the left of each autoradiogram refer to the size in nucleotides of coelectrophoresed DNA molecular size markers. In the diagrams at the bottom are sketched the organizations of the ERIC-positive regions analyzed. The direction of transcription of the genes analyzed is indicated by dotted arrows. Primers are denoted by arrows; lines labeled "a" to "l" denote the extended products. Numbers indicate the distances in base pairs separating ERICs from either the stop or the start codons of neighboring ORFs.

in the *argH-argB* transcript ratios (Fig. 7). Finally, the ERIC which separates *panB* and *panC* genes in Ye161 is missing in the YkSS47 strain. This correlates with a threefold decrease in the level of the *panB* transcripts (Fig. 7).

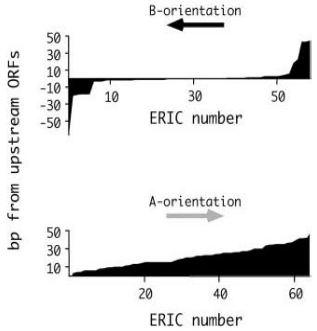


FIG. 4. Asymmetry in the orientations of ERICs. The distances in base pairs separating B-oriented and A-oriented ERICs from flanking upstream ORFs in the *Y. enterocolitica* 8081 strain are indicated.

Translating ribosomes and RNA folding. Data shown in Fig. 5 to 7 support the notion that the relative abundance of the mRNA segments flanking ERIC sequences may be influenced by the orientation of ERICs. The high downstream/upstream transcript ratio measured at intercistronic barriers spanning B-oriented elements may correlate with the activity of promoter sequences directing the synthesis of transcripts toward downstream genes. In A-oriented ERICs, the hypothetical promoter would also direct the synthesis of transcripts toward upstream genes, causing transcriptional collisions and allowing for the formation of antisense RNA. It is difficult to envisage how this may be advantageous to the organism. Moreover, it is left unexplained why B-oriented elements tend to be inserted so close to stop codons. We rather believe that transcribed ERICs may act as modulators of RNA decay and that A- and B-oriented elements may function in different ways. According to this hypothesis, the high downstream/upstream gene transcript ratios measured at intercistronic barriers carrying B-oriented ERICs may be the result of processing events promoted by ERIC repeats which enhance upstream RNA degradation.

The orientation-dependent mode of action suggests that a

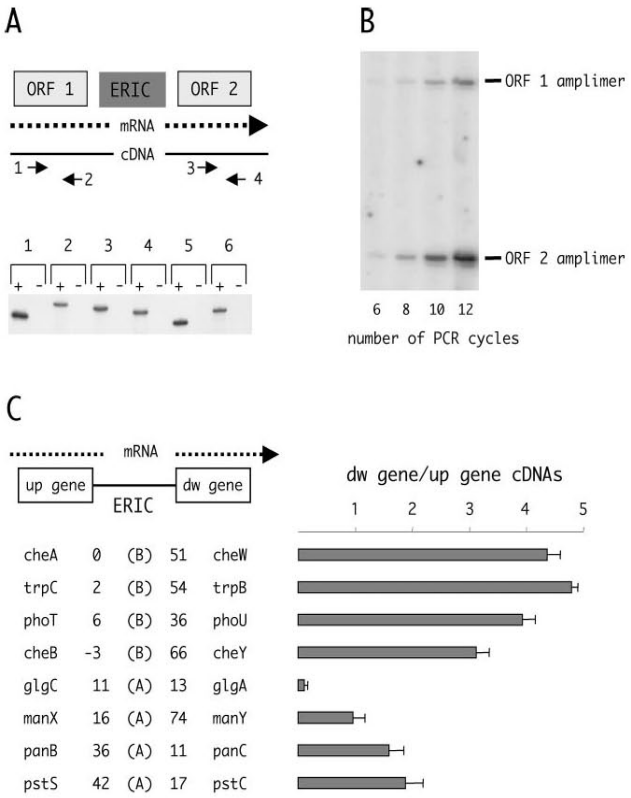


FIG. 5. RT-PCR analyses of ERIC-positive transcripts. Total RNA (200 nanograms) derived from the Ye161 strain had been reverse transcribed by using a mixture of random hexamers as primers. The cDNA obtained had been amplified by PCR with cistron-specific oligomers. One oligonucleotide of each pair of primers was ^{32}P end labeled to allow amplimer detection by autoradiography. Reaction products were run on 6% polyacrylamide-8 M urea gels. (A) Transcripts spanning ERIC sequences and *cheA-cheW* (lane 1), *cheB-cheY* (lane 2), *manX-manY* (lane 3), *panB-panC* (lane 4), *trpC-trpB* (lane 5), and *pstS-pstC* (lane 6) genes were detected by using external primers 1 and 4 under standard PCR cycling conditions (20 to 22 cycles). Amplimers were detected only when RNA samples were incubated with reverse transcriptase (+ lanes) prior to PCR. (B) Total cDNA from the Ye161 strain had been amplified by using pairs of ORF-specific primers for a limited number of PCR cycles (6 to 12). Amplimers were quantitated by phosphorimaging. In the example reported, amplimers 1 and 2 correspond to the *cheA* and *cheW* genes, respectively (C) The listed genes flanking ERIC repeats have been analyzed as described above. Distances in base pairs separating ERIC termini from stop and start codons of flanking ORFs are indicated. The orientation of the ERIC (A or B) is given in parenthesis. The number of transcripts corresponding to downstream (dw) and upstream (up) genes for each pair is expressed as a ratio. RT-PCR analyses were routinely repeated three to four times on two independent RNA preparations. Standard deviations are indicated. For each ORF analyzed (with the YE number assigned by the Sanger Centre shown in parentheses), the hypothesized function, system, and/or product(s) are as follows: for *cheA* (YE2577), chemotaxis protein CheA; for *cheW* (YE2576), chemotaxis protein CheW; for *trpC* (YE2212), tryptophan biosynthesis bifunctional protein; for *trpB* (YE2213), tryptophan synthase subunit B; for *phoT* (YE4198), high-affinity P-specific transport and cytoplasmic ATP-binding protein; for *phoU* (YE4196), P uptake, high-affinity P-specific transport system, and regulatory gene; for *cheB* (YE2571), glutamate methylesterase; for *cheY* (YE2570), chemotaxis protein CheY; for *glgC* (YE4011), glucose-1-phosphate adenylyltransferase; for *glgA* (YE4010), glycogen synthase; for *manX* (YE1777), mannose phosphotransferase system and EIIAB component; for *manY* (YE1776), mannose phosphotransferase system and EIIC component; for *panB* (YE0720), ketopantoate hydroxymethyltransferase; for *panC* (YE0719), pantoate-beta-alamine ligase; for *pstS* (YE4201), phosphate-binding periplasmic protein; and for *pstC* (YE4200), phosphate transport system permease.

sequence must be crucial for upstream RNA instability. RNAs corresponding to A-oriented and B-oriented ERICs may fold into secondary structures which have similar shapes and comparable calculated free energies (Fig. 8A; see references 17 and 38). The formation of RNA hairpins is preserved in the majority of elements by compensatory mutations and is unaffected in shorter as well as larger ERICs, because both type 1 and type 2 insertions feature self-complementary regions (Fig. 1D; see also reference 37). However, the left-hand TIRs of ERICs, which are inserted close to stop codons, are covered by terminating ribosomes, a translating ribosome protecting at least 30 residues of the mRNA (40). It is noteworthy that an AU-rich sequence (AAUUAUUUA; Fig. 8A) would not be base paired in B-oriented elements because of steric hindrance

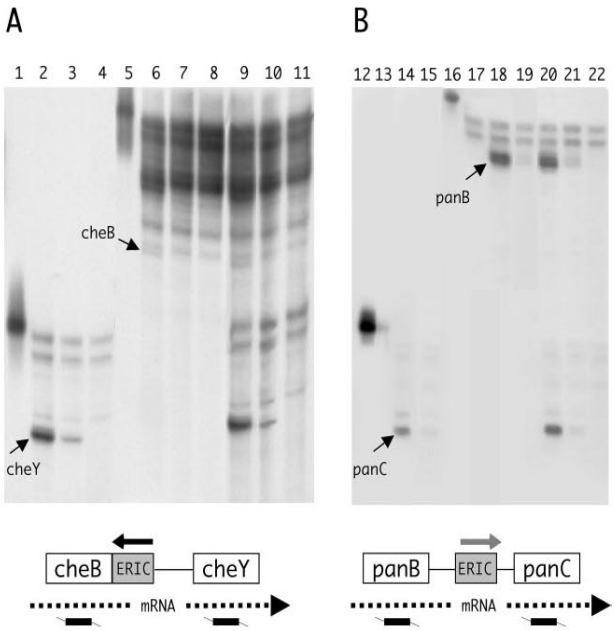


FIG. 6. RNase protection of ERIC-positive transcripts. Uniformly ^{32}P-labeled antisense RNA probes, complementary to the coding regions of the *Y. enterocolitica cheB-cheY* and *panB-panC* genes, were transcribed in vitro by the T7 RNA polymerase. In the diagram, RNA probes are sketched (not to scale) below the gene depictions. Thicker segments mark complementarity to mRNA. Probes were hybridized to 20 μg of total RNA from Ye161 cells untreated or exposed to rifampin (final concentration, 200 μg/ml) for 12 min before cell harvesting. RNase T_1-resistant RNA hybrids were electrophoresed on 6% polyacrylamide-8 M urea gels. Reaction products corresponding to *cheY*, *cheB*, *panC*, and *panB* RNAs are marked by arrows. (A) Analysis of *cheB-cheY* RNAs. Lanes: unreacted input probes (1, *cheY*; 5, *cheB*); probes hybridized separately to total RNA from Ye161 cells untreated (2, *cheY*; 6, *cheB*) or exposed for 12 min to rifampin (3, *cheY*; 7, *cheB*) or hybridized to *E. coli* tRNA (4, *cheY*; 8, *cheB*). Probes were hybridized together to total RNA from Ye161 cells untreated (9) or exposed for 12 min to rifampin (10) or hybridized to *E. coli* tRNA (11). (B) Analysis of *panB-panC* RNAs. Lanes: unreacted input probes (12, *panC*; 16, *panB*); probes hybridized separately to total RNA from Ye161 cells untreated (14, *panC*; 18, *panB*) or exposed for 12 min to rifampin (15, *panC*; 19, *panB*) or hybridized to *E. coli* tRNA (13, *panC*; 17, *panB*). Probes were hybridized together to total RNA from Ye161 cells untreated (20) or exposed for 12 min to rifampin (21) or hybridized to *E. coli* tRNA (22).

caused by ribosomes. Unfolded AU-rich sequences represent preferred cleavage sites for RNase E (13, 19, 21, 26). The enzyme, which is conserved both in *Y. enterocolitica* and *Y. pestis* (ORFs YE1627 and YPO1590, respectively), is the major endoribonuclease responsible for the mRNA decay in bacteria (8) and is associated in *E. coli* with the 3′-5′ exoribonucleases polynucleotide phosphorylase and RNase II in the molecular machine known as degradosome (8, 32). The mRNA degradation by 3′-5′ exonucleases subsequent to RNase E-mediated cleavage may explain the high downstream/upstream transcript ratios measured at specific ERIC-positive intercistronic barriers (Fig. 5 to 7).

Experimental support to this hypothesis is provided by data shown in Fig. 8B. Cleavage of ERIC-positive mRNAs should be favored by the occupancy of the left-hand ERIC TIR by translating ribosomes. Moreover, uncoupling transcription and translation should alter the downstream/upstream gene transcript ratio in ERIC-positive mRNAs spanning B-oriented ERICs only. ERICs located downstream from the *cheA* and *panB* genes are inserted in the B and A orientations, respectively (Fig. 5). Treatment of Ye161 cells with chloramphenicol significantly altered the *cheW-cheA* transcript ratio, as we measured a fourfold increase in the amount of *cheW* RNA but no effect on the *panC-panB* transcript ratio (Fig. 8B; see references 23 and 39).

It is noteworthy that the predominant extension products corresponding to the "e" and "I" bands in Fig. 3 nicely match in size RNA species generated by cleavage of *cheW* and *trpB* transcripts, respectively, at the AU-rich site within the upstream B-oriented ERICs.

A-oriented ERICs are found far from stop codons and therefore can fold into RNA hairpins. These elements may therefore act in a way opposite from that of to B-oriented ERICs and function as upstream RNA stabilizers (see Discussion).

DISCUSSION

Origin and evolution of ERIC sequences. ERIC repeats are present in several bacterial species as low-copy-number families, and PCR fingerprinting using ERIC primers is widely used for diagnostic purposes (43). In contrast, ERICs are a major genome component in pathogenic yersiniae, accounting for ~0.7% and ~0.45% of the total DNA contents of *Y. enterocolitica* and *Y. pestis*, respectively. In either species, elements

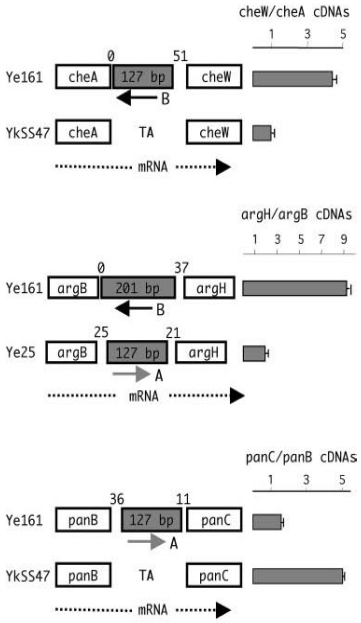


FIG. 7. Comparison of loci carrying or missing ERIC sequences in different *Yersinia* strains. ERIC elements are depicted as gray boxes, and numbers within refer to element sizes. Numbers above boxes signal the distances in base pairs separating ERIC from flanking ORFs. Total RNAs derived from *Y. enterocolitica* strains Ye161 and Ye25 and from *Y. kristensenii* strain YkSS47 were analyzed by RT-PCR as described for Fig. 5. At the empty genomic sites identified in the genome of the YkSS47 strain, ERIC sequences are replaced by the dinucleotide TA.

are scattered throughout the chromosome mostly as single-copy insertions. The genomic spread of ERICs occurred most probably by transposition. As unambiguously set by the comparison of empty and filled chromosomal sites, ERICs specifically duplicate the dinucleotide TA upon genomic insertion (Fig. 2). This is a hallmark of miniature transposable elements originating from members of the IS*630*-Tc1-mariner TE superfamily known as MITEs. The mobilization of ERICs, initially fostered by large codogenic progenitor ISs, also might have been eventually mediated, as has been previously suggested for eukaryotic MITEs (15, 18, 31, 36), by ISs whose transposases were able to recognize ERIC termini. ERICs are plausibly no longer mobile in yersiniae, as we could identify in silico neither bona fide ERIC progenitors nor potential cross-mobilizing TEs either in the sequenced *Y. enterocolitica* and *Y. pestis* strains or in the genome of the *Y. pseudotuberculosis* IP32593 strain, whose sequence has been recently determined (7). Data reported in this work support the notion that yersiniae learned during evolution to exploit the genomic spread of ERICs for functional purposes.

ERICs as modulators of RNA decay. In yersiniae, ERICs are frequently inserted next to codogenic regions, and most are plausibly transcribed into mRNAs. The ability of ERIC RNA to fold into relatively robust, low-free-energy RNA hairpins (Fig. 8A) is a feature previously noted (17, 38).

Whole in silico surveys surprisingly revealed a privileged orientation of ERIC sequences relative to their position in the mRNA. In the *Y. enterocolitica* 8081 strain, 56/60 elements which either overlap or are located 6 bp or less from the stop codon of annotated ORFs are inserted in the same orientation (B-oriented ERICs). By contrast, 45/47 elements located more distantly from stop codons (distance range, +7 to +35) are inserted in the opposite orientation (A-oriented ERICs). This peculiar organization must convey a selective advantage in evolution for functional purposes.

The preferential location next to stop codons implies that RNA hairpins formed by B-oriented ERICs are remodelled by terminating ribosomes (Fig. 8C). We hypothesize that inhibiting secondary structure formation unmasks a potential target site for RNase E, which is located in the right-hand TIR of these elements. In turn, the endonucleolytic cleavage activates the degradation of upstream RNA segments by polynucleotide phosphorylase and RNase II, the two 3′-5′ exoribonucleases associated with the RNase E in the degradosome (8, 32).

Translation should not interfere with the formation of RNA secondary structures in A-oriented ERICs. By folding into stable RNA hairpins, these repeats should be able to slow down the degradation of upstream RNA segments by impeding the passage of 3′-5′ exonucleases. These repeats may thus work analogously to the shorter intergenic sequences known as REPs, which are found in *E. coli* (16). The element found at the *glgC-glgA* intercistronic barrier seems to work this way (Fig. 5). A similar conclusion can be reached for the element found between *panB* and *panC* cistrons (Fig. 7). However, in other transcriptional units spanning A-oriented elements, upstream and downstream transcripts accumulated at similar levels (Fig. 5). We do not have an explanation for such discrepancies. Plausibly, several A-oriented ERICs cannot function as upstream RNA stabilizers because they are overridden by dominant instability determinants located in the mRNA. Such a phenomenon has been documented for different *E. coli* REPs (25, 27, 28). Similarly, the degradation of 5′ flanking RNA prompted by B-oriented ERICs may be impaired by mRNA stability determinants. The efficacy by which ERICs modulate RNA decay may vary not only because of the intrinsic stabilities of neighboring mRNA segments but also because of sequence heterogeneity among ERICs. Thus, conclusions on the abilities of members of the ERIC family to function as RNA control elements can be drawn in many instances only by integrating sequence data with functional RNA analyses.

In spite of the smaller size of their family, *Y. pestis* ERICs also can be largely sorted into A-oriented and B-oriented elements according to their distances from upstream ORFs. Whether the ERIC-dependent modulation of RNA decay works in this species, which rapidly evolved as an arthropod-adapted pathogen, remains to be established.

In the *Y. enterocolitica* 8081 strain, 30 elements are inserted relatively far from ORF stop codons but close (≤50-bp distance) to ORF start codons. These repeats may either stabilize downstream RNA sequences (*lpdA* and *uncE* transcripts in Fig. 3) or interfere with mRNA translation. Some ERICs, alternatively, could function as DNA, rather than as RNA, elements. However, deleting an ERIC from the promoter re-

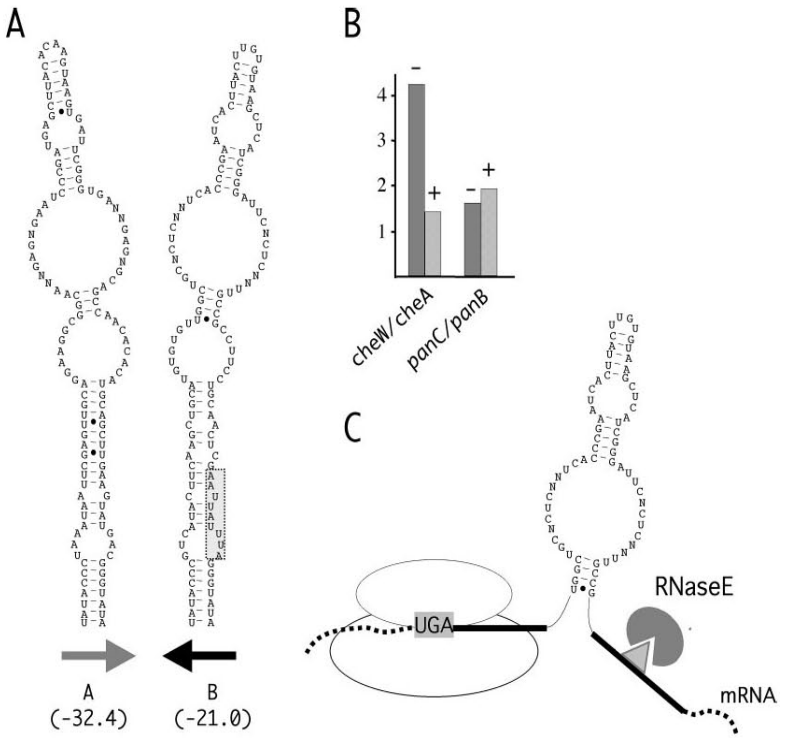


FIG. 8. (A) Predicted RNA foldings and relative calculated free energies at 37°C of unit-length ERIC consensus sequences inserted in A and B orientations. Non-Watson-Crick base pairings are highlighted by dots. The hypothesized cleavage site for RNase E, present in B-oriented ERICs, is boxed. (B) Translation-dependent processing of ERIC-positive RNAs. Total RNA derived from exponentially growing Ye161 cells untreated (−) or exposed for 30 min to chloramphenicol (+) (final concentration, 50 μg/ml) was analyzed as described for Fig. 5. (C) Ribosomes interfere with folding of ERIC-positive RNA. In mRNA-spanning B-oriented elements that are inserted close to the translational stop codon, the translating ribosome covers most of the ERIC left-hand TIR, unmasking the RNase E site (sketched as a triangle) located in the ERIC right-hand TIR.

gion of the *Y. enterocolitica cpdB* gene had no effect on *cpdB* expression (42). By contrast, the ERIC found in the promoter of the *Y. enterocolitica ybtA* yersiniabactin regulator may modulate yersiniabactin activity, as putative binding sites for the YbtA transcriptional regulator and the TATACCC motif found in ERIC TIRs coincide (33).

The numbers, the structural organizations, and the chromosomal distributions of ERICs and neisserial NEMIS sequences are similar. It is curious to note that members of these two MITE families, spread in evolutionarily distant gram-negative bacteria, independently evolved into substrates for the major cellular endoribonucleases. We would not be surprised to learn that bacterial MITEs yet to be discovered may have similarly evolved into *cis*-acting sequences regulating mRNA metabolism. Whether MITE-like repeats found in eukaryotes may similarly work as RNA regulatory elements remains to be established.

ACKNOWLEDGMENTS

We are indebted to Ida Luzzi and Francesca Berlutti for providing us with *Yersinia* strains and to Bruno Bruni for critical revision of the manuscript.

This work has been funded by a grant assigned to Pier Paolo Di Nocera by the PRIN 2004 agency of the Italian Ministry of the University and Scientific Research.

REFERENCES

1. **Achtman, M., K. Zurth, G. Morelli, G. Torrea, A. Guiyoule, and E. Carniel.** 1999. Yersinia pestis, the cause of plague, is a recently emerged clone of Yersinia pseudotuberculosis. Proc. Natl. Acad. Sci. USA **96:**14043–14048.
2. **Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman.** 1990. Basic local alignment search tool. J. Mol. Biol. **215:**403–410.
3. **Bachellier, S., J. M. Clement, and M. Hofnung.** 1999. Short palindromic repetitive DNA elements in enterobacteria: a survey. Res. Microbiol. **150:**627–639.
4. **Bottone, E. J.** 1999. Yersinia enterocolitica: overview and epidemiologic correlates. Microbes Infect. **1:**323–333.
5. **Brugger, K., P. Redder, Q. She, F. Confalonieri, Y. Zivanovic, and R. A. Garrett.** 2002. Mobile elements in archaeal genomes. FEMS Microbiol. Lett. **206:**131–141.
6. **Buisine, N., C. M. Tang, and R. Chalmers.** 2002. Transposon-like Correia elements: structure, distribution and genetic exchange between pathogenic Neisseria species. FEBS Lett. **522:**52–58.
7. **Chain, P. S., E. Carniel, F. W. Larimer, J. Lamerdin, P. O. Stoutland, W. M. Regala, A. M. Georgescu, L. M. Vergez, M. L. Land, V. L. Motin, R. R. Brubaker, J. Fowler, J. Hinnebusch, M. Marceau, C. Medigue, M. Simonet, V. Chenal-Francisque, B. Souza, D. Dacheux, J. M. Elliott, A. Derbise, L. J. Hauser, and E. Garcia.** 2004. Insights into the evolution of Yersinia pestis through whole-genome comparison with Yersinia pseudotuberculosis. Proc. Natl. Acad. Sci. USA **101:**13826–13831.

8. **Coburn, G. A., and G. A. Mackie.** 1999. Degradation of mRNA in *Escherichia coli*: an old problem with some new twists. Prog. Nucleic Acid Res. Mol. Biol. **62:**55–108.
9. **De Gregorio, E., C. Abrescia, M. S. Carlomagno, and P. P. Di Nocera.** 2002. The abundant class of *nemis* repeats provides RNA substrates for ribonuclease III in Neisseriae. Biochim. Biophys. Acta **1576:**39–44.
10. **De Gregorio, E., C. Abrescia, M. S. Carlomagno, and P. P. Di Nocera.** 2003. Asymmetrical distribution of *Neisseria* miniature insertion sequence DNA repeats among pathogenic and nonpathogenic *Neisseria* strains. Infect. Immun. **71:**4217–4221.
11. **De Gregorio, E., C. Abrescia, M. S. Carlomagno, and P. P. Di Nocera.** 2003. Ribonuclease III-mediated processing of specific Neisseria meningitidis mRNAs. Biochem. J. **374:**799–805.
12. **Deng, W., V. Burland, G. Plunkett III, A. Boutin, G. F. Mayhew, P. Liss, N. T. Perna, D. J. Rose, B. Mau, S. Zhou, D. C. Schwartz, J. D. Fetherston, L. E. Lindler, R. R. Brubaker, G. V. Plano, S. C. Straley, K. A. McDonough, M. L. Nilles, J. S. Matson, F. R. Blattner, and R. D. Perry.** 2002. Genome sequence of *Yersinia pestis* KIM. J. Bacteriol. **184:**4601–4611.
13. **Ehretsmann, C. P., A. J. Carpousis, and H. M. Krisch.** 1992. Specificity of Escherichia coli endoribonuclease RNase E: *in vivo* and *in vitro* analysis of mutants in a bacteriophage T4 mRNA processing site. Genes Dev. **6:**149–159.
14. **Feschotte, C., N. Jiang, and S. R. Wessler.** 2002. Plant transposable elements: where genetics meets genomics. Nat. Rev. Genet. **3:**329–341.
15. **Feschotte, C., L. Swamy, and S. R. Wessler.** 2003. Genome-wide analysis of mariner-like transposable elements in rice reveals complex relationships with stowaway miniature inverted repeat transposable elements (MITEs). Genetics **163:**747–758.
16. **Higgins, C. F., R. S. McLaren, and S. F. Newbury.** 1988. Repetitive extragenic palindromic sequences, mRNA stability and gene expression: evolution by gene conversion? A review. Gene **72:**3–14.
17. **Hulton, C. S. J., C. F. Higgins, and P. M. Sharp.** 1991. ERIC sequences, a novel family of repetitive elements in the genomes of *Escherichia coli*, *Salmonella typhimurium* and other enterobacteria. Mol. Microbiol. **5:**825–834.
18. **Jiang, N., C. Feschotte, X. Zhang, and S. R. Wessler.** 2004. Using rice to understand the origin and amplification of miniature inverted repeat transposable elements (MITEs). Curr. Opin. Plant Biol. **7:**115–119.
19. **Kaberdin, V. R.** 2003. Probing the substrate specificity of Escherichia coli RNase E using a novel oligonucleotide-based assay. Nucleic Acids Res. **31:**4710–4716.
20. **Kidwell, M. G., and D. R. Lisch.** 2000. Transposable elements and host genome evolution. Trends Ecol. Evol. **15:**95–99.
21. **Lin-Chao, S., T. T. Wong, K. J. McDowall, and S. N. Cohen.** 1994. Effects of nucleotide sequence on the specificity of *rne*-dependent and RNase E-mediated cleavages of RNA I encoded by the pBR322 plasmid. J. Biol. Chem. **269:**10797–10803.
22. **Liu, S. V., N. J. Saunders, A. Jeffries, and R. F. Rest.** 2002. Genome analysis and strain comparison of Correia repeats and Correia repeat-enclosed elements in pathogenic *Neisseria*. J. Bacteriol. **184:**6163–6173.
23. **Lopez, P. J., I. Marchand, O. Yarchuk, and M. Dreyfus.** 1998. Translation inhibitors stabilize Escherichia coli mRNAs independently of ribosome protection. Proc. Natl. Acad. Sci. USA **95:**6067–6072.
24. **Mazzone, M., E. De Gregorio, A. Lavitola, C. Pagliarulo, P. Alifano, and P. P. Di Nocera.** 2001. Whole-genome organization and functional properties of miniature DNA insertion sequences conserved in pathogenic Neisseriae. Gene **278:**211–222.
25. **McCarthy, J. E., B. Gerstel, B. Surin, U. Wiedemann, and P. Ziemke.** 1991. Differential gene expression from the Escherichia coli atp operon mediated by segmental differences in mRNA stability. Mol. Microbiol. **10:**2447–2458.
26. **McDowall, K. J., S. Lin-Chao, and S. N. Cohen.** 1994. A+U content rather than a particular nucleotide order determines the specificity of RNase E cleavage. J. Biol. Chem. **269:**10790–10796.
27. **Meyer, B. J., and J. L. Schottel.** 1992. Characterization of cat messenger RNA decay suggests that turnover occurs by endonucleolytic cleavage in a 3′ to 5′ direction. Mol. Microbiol. **9:**1095–1104.
28. **Meyer, B. J., A. E. Bartman, and J. L. Schottel.** 1996. Isolation of a mRNA instability sequence that is cis-dominant to the ompA stability determinant in Escherichia coli. Gene **179:**263–270.
29. **Oggioni, M. R., and J. P. Claverys.** 1999. Repeated extragenic sequences in prokaryotic genomes: a proposal for the origin and dynamics of the RUP element in Streptococcus pneumoniae. Microbiology **145:**2647–2653.
30. **Parkhill, J., B. W. Wren, N. R. Thomson, R. W. Titball, M. T. Holden, M. B. Prentice, M. Sebaihia, K. D. James, C. Churcher, K. L. Mungall, S. Baker, D. Basham, S. D. Bentley, K. Brooks, A. M. Cerdeno-Tarraga, T. Chillingworth, A. Cronin, R. M. Davies, P. Davis, G. Dougan, T. Feltwell, N. Hamlin, S. Holroyd, K. Jagels, A. V. Karlyshev, S. Leather, S. Moule, P. C. Oyston, M. Quail, K. Rutherford, M. Simmonds, J. Skelton, K. Stevens, S. Whitehead, and B. G. Barrell.** 2001. Genome sequence of Yersinia pestis, the causative agent of plague. Nature **413:**523–527.
31. **Plasterk, R. H., Z. Izsvak, and Z. Ivics.** 1999. Resident aliens: the Tc1/mariner superfamily of transposable elements. Trends Genet. **15:**326–332.
32. **Py, B., C. F. Higgins, H. M. Krisch, and A. J. Carpousis.** 1996. A DEAD-box RNA helicase in the Escherichia coli RNA degradosome. Nature **381:**169–172.
33. **Rakin, A., C. Noelting, S. Schubert, and J. Heesemann.** 1999. Common and specific characteristics of the high-pathogenicity island of *Yersinia enterocolitica*. Infect. Immun. **67:**5265–5274.
34. **Redder, P., Q. She, and R. A. Garrett.** 2001. Non-autonomous mobile elements in the crenarchaeon Sulfolobus solfataricus. J. Mol. Biol. **306:**1–6.
35. **Rychlik, W., and R. E. Rhoads.** 1989. A computer program for choosing optimal oligonucleotides for filter hybridization, sequencing and in vitro amplification of DNA. Nucleic Acids Res. **17:**8543–8551.
36. **Shao, H., and Z. Tu.** 2001. Expanding the diversity of the IS630-Tc1-mariner superfamily: discovery of a unique DD37E transposon and reclassification of the DD37D and DD39D transposons. Genetics **159:**1103–1115.
37. **Sharp, P. M.** 1997. Insertions within ERIC sequences. Mol. Microbiol. **24:**1314–1315.
38. **Sharples, G. J., and R. G. Lloyd.** 1990. A novel repeated DNA sequence located in the intergenic regions of bacterial chromosome. Nucleic Acids Res. **18:**6503–6508.
39. **Sousa, S., I. Marchand, and M. Dreyfus.** 2001. Autoregulation allows Escherichia coli RNase E to adjust continuously its synthesis to that of its substrates. Mol. Microbiol. **42:**867–878.
40. **Steitz, J. A.** 1969. Polypeptide chain initiation: nucleotide sequences of the three ribosomal binding sites in bacteriophage R17 RNA. Nature **224:**957–964.
41. **Thompson, J. D., D. G. Higgins, and T. J. Gibson.** 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:**4673–4680.
42. **Trulzsch, K., A. Roggenkamp, C. Pelludat, A. Rakin, C. Jacobi, and J. Heesemann.** 2001. Cloning and characterization of the gene encoding periplasmic 2′,3′-cyclic phosphodiesterase of *Yersinia enterocolitica* O:8. Microbiology **147:**203–213.
43. **Versalovic, J., T. Koeuth, and J. R. Lupski.** 1991. Distribution of repetitive DNA sequences in eubacteria and application to fingerprinting of bacterial genomes. Nucleic Acids Res. **19:**6823–6831.
44. **Wren, B. W.** 2003. The yersiniae—a model genus to study the rapid evolution of bacterial pathogens. Nat. Rev. Microbiol. **1:**55–64.
45. **Zuker, M.** 1989. On finding all suboptimal foldings of an RNA molecule. Science **244:**48–52.

BMC Genomics

BioMed Central

Research article

Open Access

Stem-loop structures in prokaryotic genomes

Mauro Petrillo[1], Giustina Silvestro[2], Pier Paolo Di Nocera[2], Angelo Boccia[1] and Giovanni Paolella*[1,3,4]

Address: [1]CEINGE Biotecnologie Avanzate scarl Via Comunale Margherita 482, 80145 Napoli, Italy, [2]Dipartimento di Biologia e Patologia Cellulare e Molecolare, Università Federico II Via S. Pansini 5, 80131 Napoli, Italy, [3]Dipartimento SAVA Università del Molise Via De Sanctis, 86100 Campobasso, Italy and [4]Dipartimento di Biochimica e Biotecnologie Mediche, Università Federico II Via S. Pansini 5, 80131 Napoli, Italy

Email: Mauro Petrillo - petrillo@ceinge.unina.it; Giustina Silvestro - gsilvest@unina.it; Pier Paolo Di Nocera - dinocera@unina.it; Angelo Boccia - boccia@ceinge.unina.it; Giovanni Paolella* - paolella@dbbm.unina.it

* Corresponding author

Published: 04 July 2006

BMC Genomics 2006, **7**:170 doi:10.1186/1471-2164-7-170

Received: 15 February 2006

Accepted: 04 July 2006

This article is available from: http://www.biomedcentral.com/1471-2164/7/170

Abstract

Background: Prediction of secondary structures in the expressed sequences of bacterial genomes allows to investigate spontaneous folding of the corresponding RNA. This is particularly relevant in untranslated mRNA regions, where base pairing is less affected by interactions with the translation machinery. Relatively large stem-loops significantly contribute to the formation of more complex secondary structures, often important for the activity of sequence elements controlling gene expression.

Results: Systematic analysis of the distribution of stem-loop structures (SLSs) in 40 wholly-sequenced bacterial genomes is presented. SLSs were searched as stems measuring at least 12 bp, bordering loops 5 to 100 nt in length. G-U pairing in the stems was allowed. SLSs found in natural genomes are constantly more numerous and stable than those expected to randomly form in sequences of comparable size and composition. The large majority of SLSs fall within protein-coding regions but enrichment of specific, non random, SLS sub-populations of higher stability was observed within the intergenic regions of the chromosomes of several species. In low-GC firmicutes, most higher stability intergenic SLSs resemble canonical rho-independent transcriptional terminators, but very frequently feature at the 5'-end an additional A-rich stretch complementary to the 3' uridines. In all species, a clearly biased SLS distribution was observed within the intergenic space, with most concentrating at the 3'-end side of flanking CDSs. Some intergenic SLS regions are members of novel repeated sequence families.

Conclusion: In depth analysis of SLS features and distribution in 40 different bacterial genomes showed the presence of non random populations of such structures in all species. Many of these structures are plausibly transcribed, and might be involved in the control of transcription termination, or might serve as RNA elements which can enhance either the stability or the turnover of cotranscribed mRNAs. Three previously undescribed families of repeated sequences were found in *Yersiniae*, *Bordetellae* and *Enterococci*.

Background

The tremendous flow of information generated by large scale genome-sequencing provided, as far as the prokaryotic world is concerned, the complete DNA sequence of over 200 bacterial strains, and more are becoming available every month. Most annotation work has been directed to the assessment of the protein repertoire encoded by a given microbe, aiming to the genome-scale reconstruction of bacterial metabolism [1], the identification of gene sets unique to pathogenic microorganisms [2,3] or the development of new vaccines [4]. The availability of massive amount of sequence data also stimulated in depth evaluation of the organization of the bacterial chromosome [5-9]. The basic organization of the genetic material (DNA curvature and stacking energy, base and oligo skews, etc.; see ref. [10]), and the presence of simple or more complex sequence repeats [11,12] have also been analyzed for most sequenced bacterial genomes.

Information associated to the folding of specific, single stranded sequence regions into secondary structures is relatively ill-defined in prokaryotes. Prediction of RNA secondary structures may show different and even contrasting results, depending on the methodologies and the genomic regions evaluated [13-15].

In bacteria, protein coding sequences may be regarded as able to be transcribed and to form predictable secondary structures, although in many instances the spontaneous folding of the corresponding mRNA may be affected by interactions with the translation machinery. Stem-loop-structures (SLSs) in RNA may in turn control transcription, as in the attenuation mechanism [16], or influence translation, as SECIS elements do for the insertion of selenocysteine at stop codons [17]. Secondary structure prediction is very effective for relatively small RNA with defined ends, especially when corroborated by phylogenetic data, but it is more ambiguous in larger RNAs, where SLSs, especially those containing short stems, are easily formed, or lost, when a sliding window is used to tentatively delimit the boundaries of a folding domain.

Longer stems significantly contribute to the formation of complex secondary structures where they affect RNA stability and functionality. Many non coding RNA structures are known to fold around a stem which delimits either a small, simple, single-strand loop or a larger, highly structured sequence. Examples are found in self-splicing introns [18], riboswitches [19], transcribed intergenic repeats such as *E.coli* BIME, *Yersinia* ERIC and *Neisseria* NEMIS sequences. In these cases the stem is often essential to the attainment of the correct secondary structure and may be directly recognized by ribonucleases [20-23]. Some predicted SLSs might also form in DNA and affect its conformation: base pairing of single stranded DNA is known to play a role in recombination, replication and transcription [24-26].

Here we present a systematic analysis of SLS distribution in prokaryotic genomes. Sequences able to fold into stem-loop structures featuring relatively large (12 or more bp) stems have been searched and analyzed in 40 wholly-sequenced bacterial chromosomes. SLSs found in searched bacterial genomes are more numerous and more stable than those randomly expected to form in sequences of comparable size and composition. The enrichment of specific SLS sub-populations may be observed within selected intergenic regions (IGRs).

Results

Identification of stem-loop structures (SLSs)

A relatively large number of completely sequenced bacterial genomes is currently available, from different species of medical, industrial or purely scientific interest. While for some species only one or two strains have been sequenced, for others, such as *E. coli* and *Salmonellae*, multiple variant strains have been sequenced, leading to over-representation of these sequences in available databases. For the present study, we selected a set of 40 genomes from different bacterial species (Table 1), constituting a representative sample of the prokaryotic world in terms of evolutionary distance, genome complexity and GC content.

Genomes of bacteria listed in Table 1 were analyzed to identify all the single-strand sequence regions able to fold into SLSs featuring double-stranded stem regions measuring at least 12 bp, and loops 5 to 100 nt long. GU base-pairs within the stem were allowed. Similarly allowed was a single mismatch or bulge located at least two matches away from the ends of the stem. These settings were chosen in order to identify both short 'canonical' stem-loop structures (i.e. with simple loops) and larger ones containing 'highly structured loops'. The number of SLSs found in each bacterial genome, grouped according to the SLS position, relatively to the boundaries of known and predicted genes annotated in the TIGR database, is reported in Fig. 1. SLSs are classified according to the following categories: a) *coding*, entirely contained within a coding sequence, located either on the sense or the anti-sense strand; b) *intergenic*, entirely located between coding sequences; c) *end-spanning*, spanning one of the ends of a coding sequence. The number of SLSs ranges from the slightly more than 20.000, in *Mycoplasmae* and other small genomes, to about 200.000, in large genomes as those featured by *Bordetellae* and *Pseudomonaceae*.

The large majority of SLSs falls within, or spans the ends of, genic regions; only about 10% of SLSs were found in IGRs. This distribution is not surprising as it reflects the

Table 1: Bacterial species analyzed in this study are numbered 1 to 40. The strains used for *in silico* analyses, the size of their genomes in base pairs and their relative GC content are shown. Representative species chosen for comparative analyses are labeled a through to v.

Division	Species	Strain			Genome size	GC%
low-GC Firmicutes	*Bacillus anthracis*	Ames	1	a	5227293	35.4
	Bacillus halodurans	C-125	2		4202353	43.6
	Bacillus subtilis	168	3		4214810	43.5
	Clostridium perfringens	13	4		3031430	28.5
	Clostridium tetani	E88	5	b	2799250	28.7
	Enterococcus faecalis	V583	6	c	3218031	37.0
	Lactobacillus johnsonii	NCC533	7		1992676	34.6
	Listeria innocua	CLIP11262	8		3011208	37.3
	Listeria monocytogenes	EGD-e	9	d	2944528	37.9
	Staphylococcus aureus	MW2	10		2820462	32.7
	Streptococcus pneumoniae	TIGR4	11	e	2160837	39.6
	Streptococcus pyogenes	SF370	12		1852442	38.4
Mollicutes	*Mycoplasma genitalium*	G-37	13	f	580074	31.6
	Mycoplasma pneumoniae	M129	14	g	816394	39.9
	Ureaplasma urealyticum	serovar 3	15		751719	25.4
high-GC Firmicutes	*Corynebacterium diphtheriae*	NCTC13129	16		2488635	53.5
	Mycobacterium leprae	TN	17		3268203	57.7
	Mycobacterium tuberculosis	H37Rv	18	h	4411529	65.5
Chlamydiae	*Chlamydia pneumoniae*	AR39	19	i	1229853	40.5
	Chlamydia trachomatis	serovar D	20		1042519	41.2
Spirochaetae	*Treponema pallidum*	Nichols	21	j	1138012	52.7
α-Proteobacteria	*Brucella melitensis*	16 M	22	k	3294931	57.1
	Rickettsia conorii	Malish 7	23		1268755	32.4
	Rickettsia prowazekii	Madrid E	24	l	1111523	28.9
β-Proteobacteria	*Bordetella bronchiseptica*	RB50	25	m	5339179	68.1
	Bordetella parapertussis	12822	26		4773551	68.1
	Bordetella pertussis	Tohama I	27		4086189	67.7
	Neisseria meningitidis	MC58	28	n	2272351	51.4
γ-Proteobacteria	Buchnera	APS	29		640681	26.2
	Escherichia coli K12	MG1655	30	o	4639221	49.8
	Haemophilus influenzae	KW20 Rd	31	p	1830138	38.0
	Pasteurella multocida	PM70	32		2257487	40.3
	Pseudomonas aeruginosa	PA01	33	q	6264403	66.4
	Pseudomonas putida	KT2440	34		6181863	61.5
	Salmonella typhi	CT-18	35	r	4809037	52.0
	Salmonella typhimurium LT2	SGSC141	36	s	4857432	52.2
	Vibrio cholerae	N16961	37		4033464	47.2
	Yersinia pestis	CO92	38	t	4653728	47.5
ε-Proteobacteria	*Campylobacter jejuni*	NCTC11168	39	u	1641481	30.5
	Helicobacter pylori	26695	40	v	1667867	38.8

high fraction (87–90%) of sequences annotated as coding in most tested genomes. In some species, however, the number of SLSs found in the IGRs was noticeably higher. In *B. anthracis*, *C. perfringens* and *N. meningitidis*, the fraction of SLSs found in non-coding sequences exceeds 20%. A slightly lower number of intergenic SLSs was found in the *P. putida* genome.

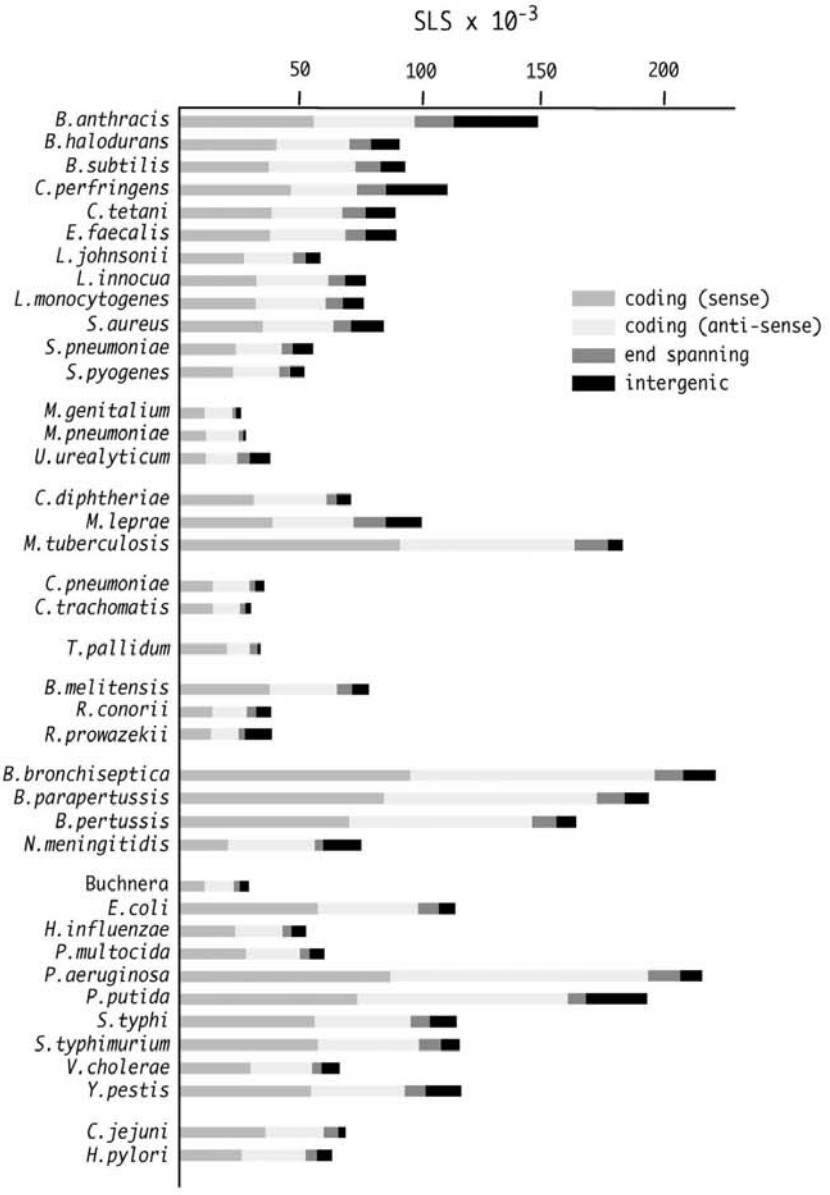


Figure 1
SLSs in bacterial genomes. The number of SLSs found in the 40 bacterial genomes listed in Table 1 is reported. SLS located completely within intergenic or coding regions are labeled as such ("sense" and "anti-sense" indicates the coding and non-coding strand, respectively); those spanning the border of coding/non-coding regions are marked as "end spanning".

SLSs in naturally occurring and reshuffled genomes

Fig. 2A reports the number of SLSs found, as a function of genome size, in the subset of 22 genomes labeled a to v in Table 1. As also shown in Fig. 1, larger genomes contain more SLSs than smaller ones, and a rough linear correspondence between genome size and SLS abundance may be observed (Fig. 2A). The same search, done on random sequences of the same length and GC content as the original genomes, produces a lower number of SLSs, linearly correlated with genome size, with the only exception of *Clostridium tetani* and *Rickettsia* genomes (Fig. 2A). As expected, for a given random genome, the number of SLSs perfectly correlates with the sequence length, when smaller fragments are tested (data not shown).

The attitude of a sequence to randomly give rise to stem-loop structures is expected to depend on a number of features, such as base composition and word frequencies. Moving away from the equally-split 25% frequency of each base, or 50% GC content, sequence complexity is reduced, and this facilitates the formation of complementary structures. This is easily seen in Fig. 2B, where SLSs found in naturally occurring genomes are plotted against GC content, after sequence length normalization. The dotted line represents SLSs found in random sequences of different GC content, all 1 Mbase long, produced by ten runs of the reshuffle tool. Variations are always within a very small (about 1%) range. As expected, random sequences stochastically give rise to a number of SLSs, which regularly grows from a minimum, for a 50% GC content, to larger numbers as GC content either decreases or increases (Fig. 2B).

Naturally occurring genomes feature a larger number of SLSs. With the only exception of *C. tetani*, the number of predicted SLSs is always higher, by several standard deviations, than in random sequences of comparable GC content, which is statistically significant for $P < 0.0001$ or lower. This indicates that some non-random component of the natural genome sequence is responsible for the larger number of SLSs, which cannot be reproduced in the shuffled sequences. Reduced complexity of the sequence may be due to consistently repeating patterns in the natural sequence, such as the tendency to prefer the use of specific di- or tri-nucleotides, or higher order words, which are not conserved in the shuffled genomes, or constraints imposed by the presence of coding regions. To test these effects, SLS found in natural genomes were compared to those found in randomized genomes, produced by shuffling while keeping constant the frequency of words of size ranging between 2 and 13 nucleotides. Shuffling genomes by only preserving word frequencies does not take into account the constraints imposed by the presence of coding regions. For this reason an alternative method (DS, double shuffle) for shuffling the natural genomes was devised, where for non-coding regions, dinucleotide frequencies are preserved, whereas for coding regions dinucleotide frequencies, encoded protein, and codon usage are all preserved, as described in reference [15]. Each randomization was repeated ten times, with very small changes in the number of SLSs found (typically <1%). The results, for three genomes of different GC content, are reported in Fig. 3, where SLSs are classified according to their stability (dG) and loop size. Progressively larger SLSs numbers are obtained by keeping the frequency of 2- to 4-nt words constant. For larger word sizes the trend appears to slow down, and subsequent increases only give rise to marginally higher numbers. Natural genomes also contain more SLSs than sequences produced according to the second, more complex, model, which distinguishes between coding and non-coding genomic regions. The differences, also in this case, are typically above four standard deviations, significant for $P < 0.0001$ or lower. SLS numbers obtained with this method are similar to those from genomes randomized by preserving 4- or 5-nt words. Interestingly, preserving larger k-lets ranging from 6 to 13, produces even higher numbers than the random genomes obtained by preserving codon usage.

Specific SLS subsets are selectively enriched in the natural genomes. The largest differences are observed with higher stability structures where the random component is expected to be lower. SLSs including the smallest loops (shorter than 20 bases) also appear to be more frequent in natural genomes, possibly including specific classes of RNA structures (Fig. 3).

Identification of specific SLS groups

From the previous data, it emerges that in most species the pool of predicted SLSs shows a bias towards energy levels and genome localization, which is highly indicative of the inclusion of non-random SLS sub-populations. As folding of SLSs containing larger 'loops' might produce alternative structures, possibly excluding the expected stem, minimum energy structures were predicted both freely and by imposing a constraint for SLS formation (see Methods). Most higher stability (dG < -10 KCal/mol) SLS-containing regions, when minimum energy structures are predicted by imposing no constraint for SLS formation, produce results within 5 KCal/mol of the SLS based structure (60 to 80% in practically all species, not shown). This indicates that, for these higher stability regions, the SLS containing structure is expected to be either the optimal or a close suboptimal structure. The relative frequency of these regions, within coding and non-coding genomic areas, was determined in the 40 bacterial genomes listed in Table 1. The results, reported in Fig. 4, were normalized to genome length and total SLS genomic frequency. Only SLSs entirely located within coding and intergenic regions

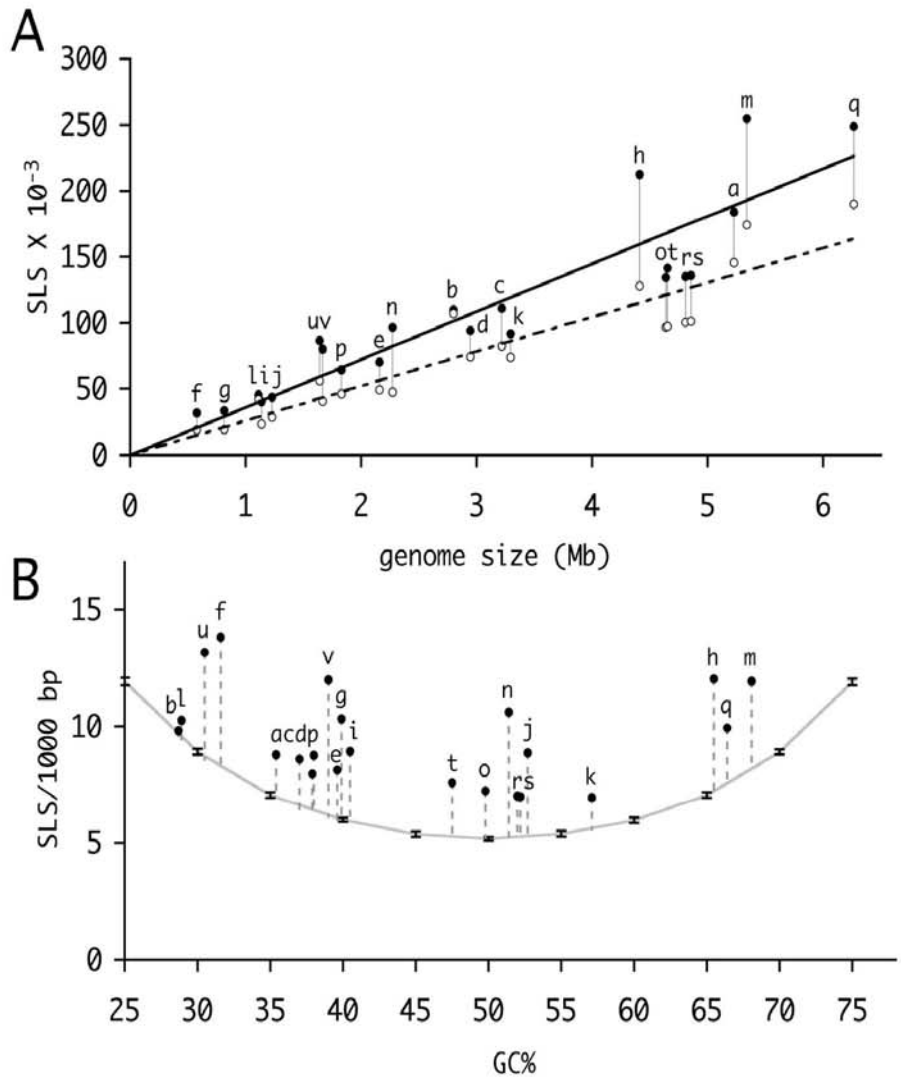


Figure 2
SLSs in natural genomes and comparable random sequences. SLSs found in natural bacterial genomes (filled circles) and random sequences of the same base composition (empty circles) are plotted against genome size in panel **A.** Filled and empty circles are connected for clarity. The same SLSs, normalized according to genome size, are plotted as a function of the GC content in panel **B**. The grey curve was obtained by determining the number of SLSs in a 1 Mbase random sequence of the indicated GC%. Each point is the average of 10 sequences, bars indicate standard deviations. Pre-filtering of SLSs was not performed. In both panels letters indicate bacterial species as listed in Table 1.

were counted. An evident enrichment in intergenic SLSs can be observed in the genomes of all the low-GC firmicutes (a bracket in Fig. 4) and in a few proto-bacteria (b-e). In both H. influenzae and P. multocida (d and e, respectively, in Fig. 4), the SLS enrichment reflects the genomic over-representation of the decameric sequence, known as DUS (for DNA Uptake Sequence), which plays a role in transformation [27]. Most of the >≈1000 DUS repeats found in either species are localized in intergenic spaces, and several are located next to each other in inverted orientation [27,28]. We assessed that this fraction of DUS repeats accounts for the formation of higher stability SLSs (not shown). The abundance of intergenic SLSs in the R. conorii and N. meningitidis genomes (b and c, respectively, in Fig. 4), correlates to the presence of species-specific palindromic repeats [29,30]. In contrast, the enrichment in intergenic SLSs in low-GC firmicutes cannot be explained by the presence of large repeated DNA families. In these genomes higher stability SLSs range in size from 30 to 50 nt, and show heterogeneity in both stem and loop lengths (not shown).

AT-rich terminator-like sequences in low-GC firmicutes

The analysis of higher stability, intergenic SLSs found in low-GC firmicutes revealed that these elements are mostly AT-rich, and frequently found close to the stop codon of genes located upstream. Typical rho-independent transcriptional terminators are relatively short SLSs, in which GC-rich stems made by 6 to 8 bp pairs are flanked on the 3' side by a stretch of 4 or more Ts [31-33]. To test the potential for SLSs from low-GC firmicutes to act as terminators, the distribution of As and Us at their termini was analyzed. Most (65 to 75%) of the higher stability SLSs located immediately downstream from annotated CDSs feature four Ts at their 3' border (Fig. 5, panel A, grey bars). The number of SLSs exhibiting the same features drops to 20%, or less, in other bacteria (Fig. 5, panels A and B). Interestingly, in low-GC firmicutes more than 50% of the SLSs featuring four Ts at the 3' border carry also four As at the 5' border (Fig. 5, panel A, black bars). Again, SLSs with identical features are 5- to 10- times less abundant in other bacteria (Fig. 5, panels A and B). The concomitant presence in low-GC firmicutes of 4As and 4Ts respectively at the 5' and 3' SLS termini is not merely due to the high AT content of their genome, but rather appears to be the result of some functional selection. In fact, very low numbers of SLSs with the inverted organization, namely carrying 5' 4Ts and 3' 4As, were found (see Fig. 5).

Distribution of intergenic SLSs

The relative positions of higher stability SLSs within the IGRs were analyzed in all the species listed in Table 1. Based on the orientation of flanking CDSs, IGRs were combined (Fig. 6) to form three intergenic spaces (IGS): a) uni-IGS, between CDSs transcribed unidirectionally, i.e. along the same orientation; b) conv-IGS, between convergently transcribed CDSs; c) div-IGS, between divergently transcribed CDSs. SLSs falling within each intergenic space are accordingly referred to as uni-, conv- and div-SLSs. In all species uni-SLSs are the largest (around 60%) SLS fraction, but no enrichment is observed, as their number reflects the length of the uni-IGS. In contrast conv-SLSs, which represent 20 to 30% of total intergenic SLSs, are concentrated in a much smaller space, as the corresponding conv-IGS covers 8 to 12% of the overall intergenic space in practically all tested species. Conversely, div-IGS, which covers 25–35% of the intergenic space, only hosts about 10% of SLSs. A corollary of this distribution is that SLSs tend to favour, as a preferential location, the 3'- over the 5'- end of flanking CDSs. To test this hypothesis also on the uni-SLSs, a representative set of these regions were further sub-divided into three sub-regions corresponding to the two 50 base spans named *left* and *right*, respectively close to 3'- and 5'-ends of the flanking CDSs, and the remaining, variable length, intermediate subregion, named *center*. Short IGRs, which could not be split into appropriate subregions, were not included in the analysis. Similarly a small number of extremely long regions, which might derive from inaccurate genome annotation, were not used. The number of SLSs found in the described subregions (Fig. 7) shows that also the uni-SLSs clearly favour the 3'-end location: in the vast majority of species SLSs found within left subregions outnumber by 2- to 4-fold those found in the equally long right subregions.

SLSs spanning repetitive DNA elements

Some intergenic SLSs may coincide with, or be a modular component of repeated DNA families. The clustering of intergenic SLS at the 3'-end of genes opens the possibility that this relative enrichment may be related to a functional role, not necessarily connected to termination. A search for DNA repeats, known from literature to cluster at the 3'-end of coding regions, revealed that REPs (repetitive extragenic palindromic sequences) found in *E. coli* [34] and *P. putida* [35] are a component of the selected population of 3'-end clustered SLS. By using SLSs as BLAST query sequences, we could identify repeated, previously undescribed DNA elements in various species (Fig. 8A). The *Bor* repeats are short SLSs ranging in size from 26 to 30 bp over-represented in *Bordetellae*. The 30-mer is found in numbers ranging from 42 to 75 in different *Bordetella* species, whereas the smaller 26 bp core is much more abundant (Fig. 8B). *Bor* are found close to coding regions, and share some similarity with the *E. coli* REPs (Fig. 8A). Novel DNA sequence elements, larger than REPs were found in *Y. pestis* and *E. faecalis* (*Ype* and *Efa* elements, respectively: Fig. 8A). Members of both DNA fam-

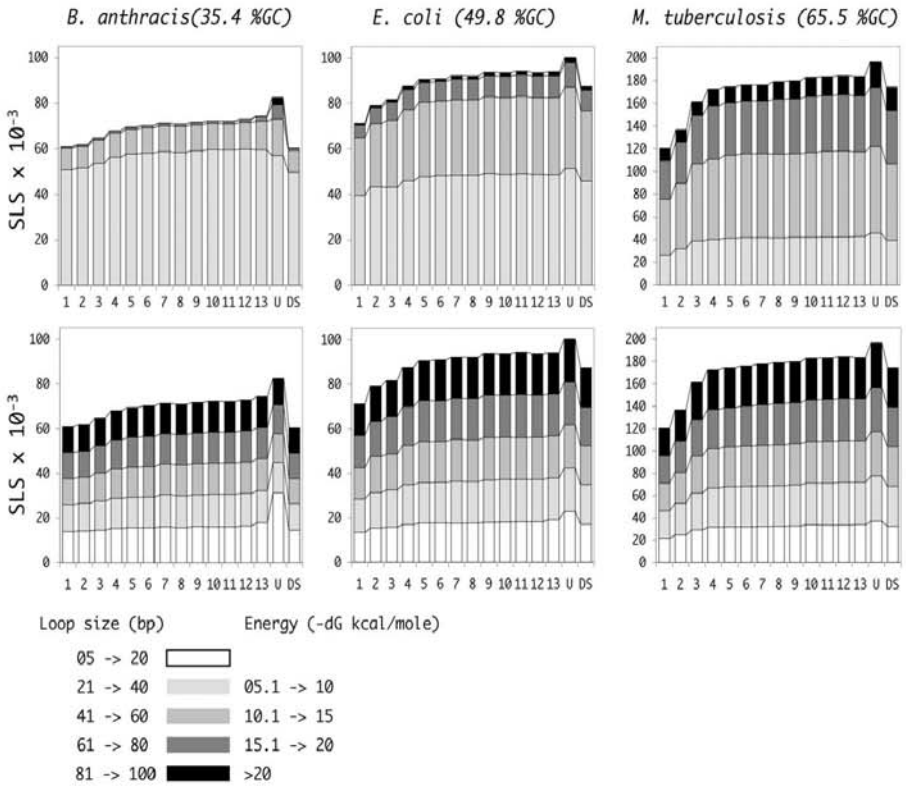


Figure 3

SLSs in *B. anthracis*, *E. coli*, and *M. tuberculosis*. In the six panels, bars represent SLSs found in *B. anthracis*, *E. coli*, and *M. tuberculosis* genomes and in randomized genomes produced from them. Bars 1 to 13 refer to randomly shuffled genomes preserving the frequency of 1 to 13 nucleotide words, respectively. Bar U (Unshuffled) refers to the natural genome, while bar DS (Double Shuffle) refers to the genome shuffled preserving information about coding regions (see Results and Methods). In each column, stacked bars are used to separate SLSs of different *dG* (top panels) or loop size (bottom panels), as indicated. Only SLSs with *dG* < -5 KCal/mole were selected. Pre-filtering of SLSs was not performed. Standard deviations are always lower than 1% and are not reported in the figure.

ilies tend also to be preferentially inserted close to the 3'-end of annotated CDSs.

Discussion

The ability of a DNA or an RNA segment to fold into a stem-loop structure derives from the presence of complementary bases, and such segments stochastically occur in every large sequence, no matter the origin, even randomly generated, provided that some level of balanced distribution of nucleotides within single strand is guaranteed. This is certainly true in bacterial genome sequences, where oligonucleotide distribution reveals compositional symmetries in a variety of complete genomes [36,37]. The problem of evaluating the relevance of a particular motif in terms of the likelihood of generating it by chance in a given sequence has been extensively faced (see for example the work by Robin and coworkers [38] for the probability of finding a motif composed of two 'boxes' separated by a variable distance). Here we chose an 'experimental' approach, based on randomized genomes produced by reshuffling the natural one, with two types of constraints: preservation of a variety of k-let frequencies and a more complex model where genic and intergenic regions are separately shuffled with conservation of ami-

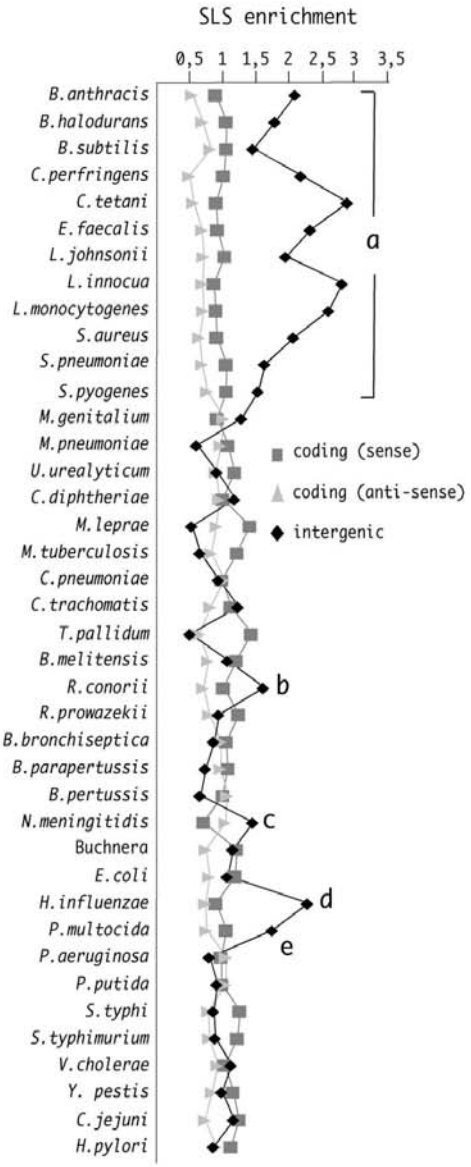


Figure 4

Species-specific enrichment in intergenic SLSs. The regional enrichment of higher stability SLSs (dG < -10 KCal/mole) completely located within coding (sense and anti-sense) and intergenic regions has been monitored in the 40 bacterial genomes listed in Table 1. Enrichment is expressed as the ratio of the SLS frequencies within each region to average SLS frequencies in the total genome. Letters **a** to **e** signal enrichments in intergenic SLSs observed in specific genomes (see Results).

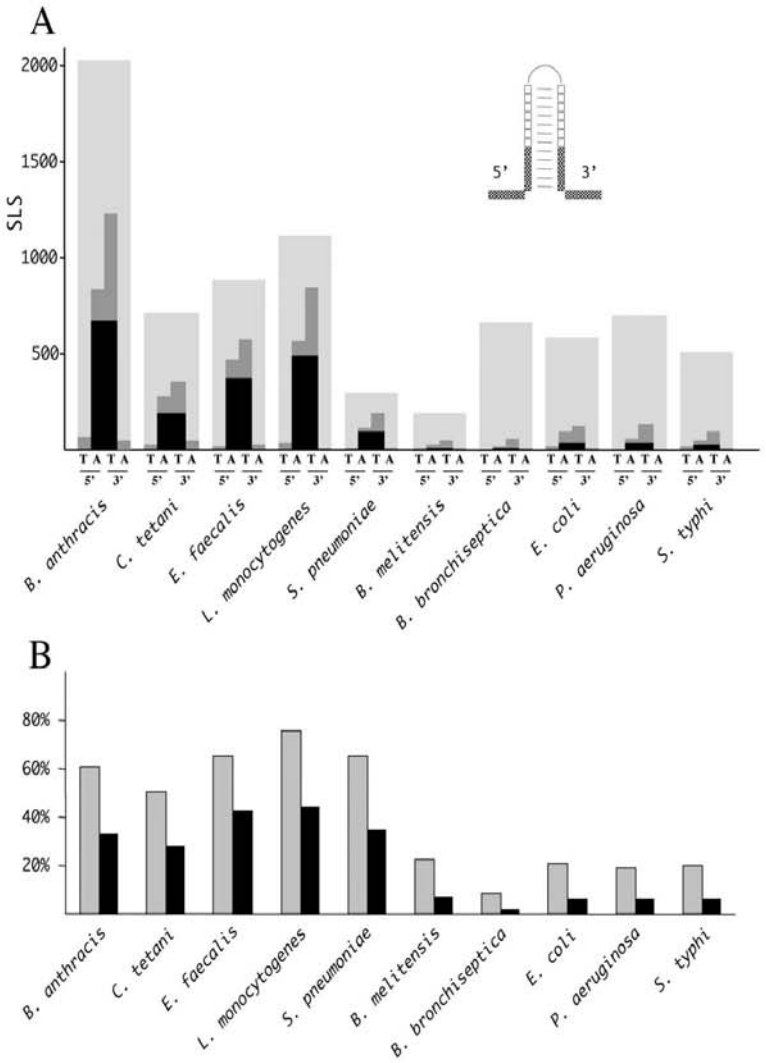


Figure 5

Distribution of A- and T-runs at SLS *termini* . Intergenic SLSs from the indicated genomes with a *dG* < -10 KCal/mole and a loop length < 30 nt, located 20 bp or less from the stop codons of CDSs (large grey bars) were screened for the presence of either A- or T-tetramers or both tetramer types at the 5' and 3' borders. Border is defined here as the 10 nt long region including, in each SLS, the first 5 nt of the stem and the 5 nt located immediately outside of the stem, either at the 5' or the 3' side (see inset in panel A). **A)** For each analyzed species, the four bars respectively indicate the number of SLSs containing 5'T-, 5'A-,3'T-and 3A'-runs of at least four identical residues. The black portions of the bars indicate the contemporary presence of both 5'As and 3'Ts. The height of the light grey regions in the background represents the total number of SLSs in the analyzed pools. **B**) The fraction of SLSs carrying 3'T-runs (grey bars) or both 5'A- and 3'T-runs (black bars) in the analyzed genomes is shown.

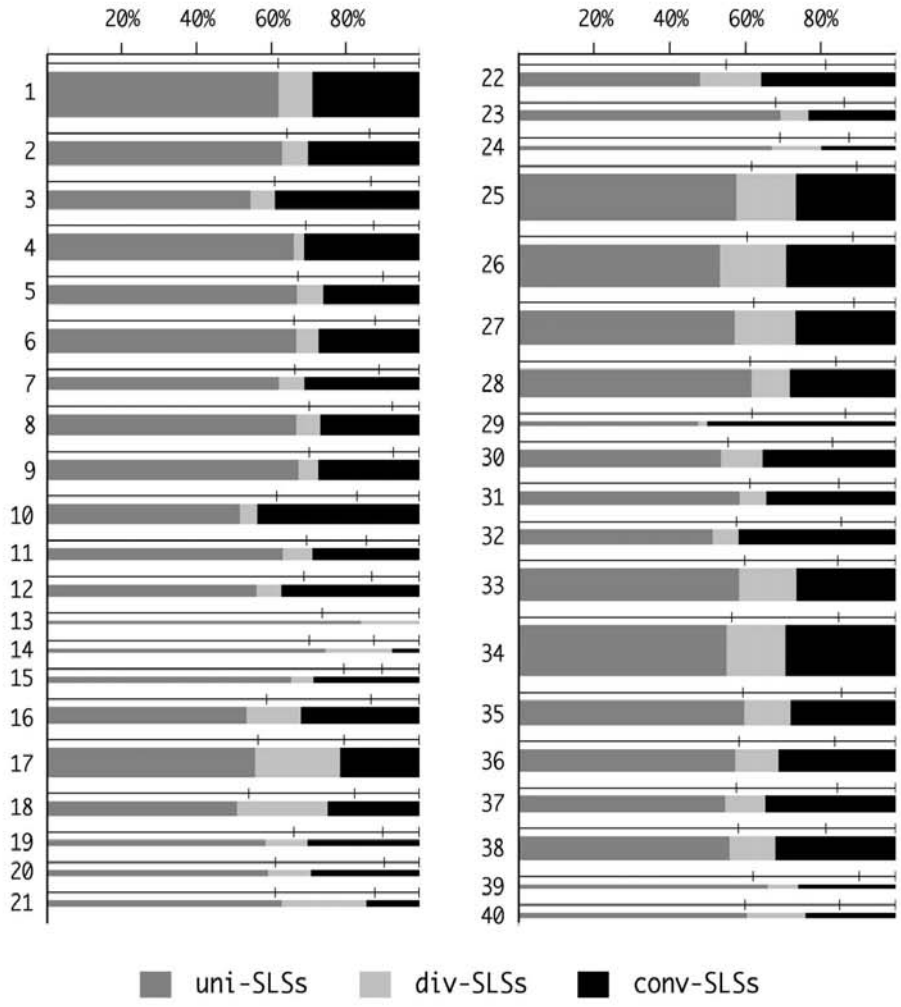


Figure 6

Classes of intergenic SLSs. Based on the orientation of flanking CDSs, higher stability intergenic SLSs ($dG < -10$ KCal/mole) have been sorted into three categories, as indicated at the bottom (see Results). The width of each stacked bar denotes the fraction of SLSs belonging to the three categories. The thickness of the bars is proportional to the cumulative sizes of IGRs (lengths below 25000 bp are not to scale, but are represented by a minimal bar width). Lines above bars represent the intergenic space, split by vertical dashes in three segments respectively corresponding, left to right, to the cumulative lengths of IGRs flanked by unidirectionally, divergently and convergently transcribed CDSs. According to the parameters adopted, no conv-IGS was found in the genome *of M. genitalium* (see row 13). Only IGRs ranging from 29 to 500 bp were taken into account, since smaller regions can not contain the shortest detectable SLSs, and bigger ones might derive from inaccurate genome annotation. Bacterial genomes are numbered 1 through to 40 as in Table 1.

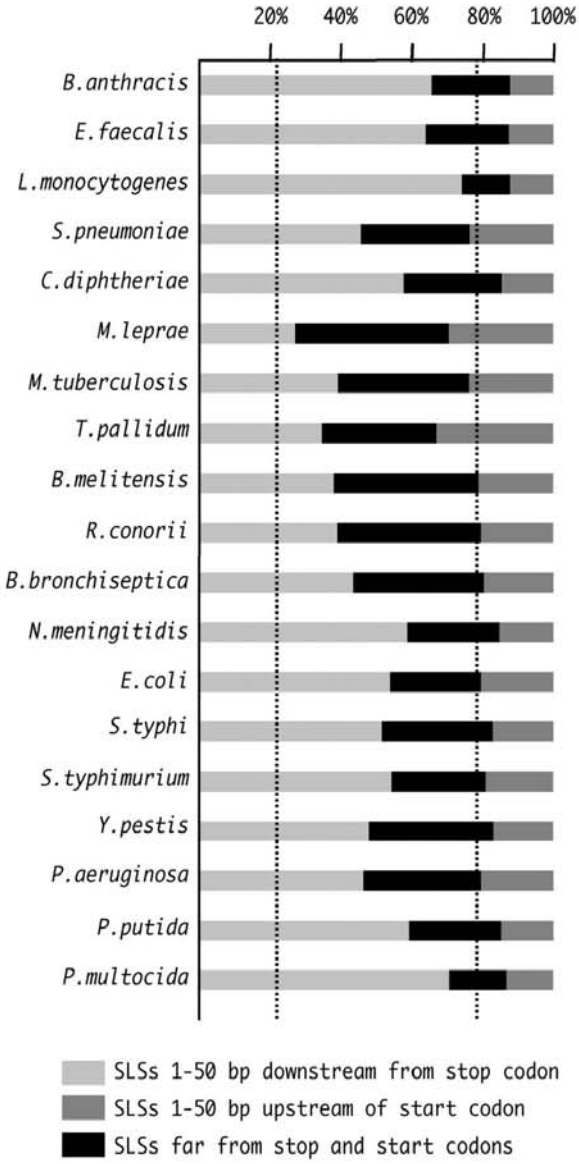


Figure 7
Subsets of uni-SLSs. SLSs located between unidirectionally transcribed CDSs have been subdivided into three categories, relatively to their distance from flanking CDSs as indicated at the bottom. The position of the dotted lines across bars denotes the averaged amount of intergenic space occupied by the three categories of SLSs analyzed. The uni-IGS minimum size was raised to 129 bp, since shorter IGSs cannot be assigned to any of the three categories. SLSs were selected as in Fig. 6.

A

species	repeat	size (bp)	abundance	described in
E.coli	REP	35-40	hundreds	reference 34
P.putida	REP	35-40	hundreds	reference 35
B.bronchiseptica	*Bor*	26-30	hundreds	this paper
E.faecalis	*Efa*	165-180	tens	this paper
Y.pestis	*Ype*	130-160	tens	this paper

B

Bor (26 bp) GTGCCTGTCCCCGCNGGGACAGGCAC
Bor (30 bp) GTGTGCCTGTCCCCGCNGGGACAGGCACAC

REP ATTGCCTGATG-CGCTACGCTTATCAGGCCTACR

	30 mer	26 mer
B.bronchiseptica	75	225
B.parapertussis	50	198
B.pertussis	42	131

Figure 8
SLSs and sequence repeats. A) Repeated DNA families spanning SLSs identified in different prokaryotic species are listed. **B)** The consensus sequence of the *Bordetella* Bor repeats is shown at the top. Sequence identities with the *E.coli* REP z1 subtype are highlighted. N, any nucleotide; R, purine. The relative abundance of the 26- and 30-bp long members of the *Bor* family in sequenced *Bordetella* genomes is reported.

noacid sequence and codon usage. SLSs found in naturally occurring genomes clearly outnumber those expected from the result of similar analysis in their randomized counterparts (Figs. 2 and 3). It appears that natural genomes somehow tend to favour the formation of specific sets of stem-loop structures, typically the more stable ones. These sets significantly contribute to the higher SLS numbers observed in naturally occurring genomes, compared to their random counterparts. The phenomenon has been observed in bacterial genomes which widely differ in terms of size, GC content, evolutionary relatedness. Data are in agreement with literature reports, showing that, in large-scale analyses of prokaryotic mRNA populations, coding regions had a significant bias toward more local secondary structure potential than expected [15].

The evolutionary pressure promoting the potential formation of stem-loop structures at genome-wide level may serve different functional purposes. At the DNA level, stem-loop structures may play a role in replication, transcription, and recombination. However, as the vast majority of prokaryotic genomes is composed of expressed, protein-coding, regions, the contribution to mRNA secondary structure formation should be taken into account for most SLSs, especially those including G-U pairs. Most SLSs fall within coding regions (Fig. 1), in agreement with their size, which typically exceeds those of non-coding regions by a factor of ten. Still, when evaluating their significance, ribosome coverage and formation of secondary structures within protein-coding regions should be regarded as alternative, ribosomes being expected to prevent the formation of most low stability mRNA structures. Higher stability structures may however result in translational pausing, possibly used in regulatory mechanisms such as attenuation [16]. In specific instances, coding SLSs correspond either to remnants of transposon-like sequences [30], or to regions encoding repetitive protein domains, such as those found in the mycobacterial PE genes or in anchored cell-wall proteins conserved in several microorganisms (not shown).

Although less numerous, SLSs tend to be more frequent within the much smaller IGRs, where a typical bias towards energy levels and genome localization may be observed, highly indicative of specific, non-random, SLS subpopulations. All the analyzed low-GC firmicutes feature a marked enrichment in higher stability intergenic SLSs. Both structure and genomic location suggest that most of these sequences may function as rho-independent transcriptional terminators. The finding is not surprising *per se*, since the transcriptional factor *rho* is not essential in *Bacillus subtilis* and *Staphylococcus aureus*, and other Gram-positive bacteria with a low GC-content lack a *rho* homolog [39,40]. However, SLSs found in low-GC firmicutes are atypical as transcriptional terminators, as most of them carry, in addition to the canonical 3' U-rich tract, a complementary A-rich tract at the 5'-end (Fig. 5). This arrangement is known not to impair termination as, for example, in the *E. coli thr* operon attenuator, the terminator features a GC-rich stem-loop flanked by 9 Us at the 3'-, and 6 As at the 5'-end, and site-directed mutagenesis has shown that upstream adenines are neither essential, nor detrimental to transcription termination [41]. The 4A/4T containing SLSs found in low-GC firmicutes, when located at a short distance from convergently transcribed genes, may function as bi-directional terminators [42]. Alternatively, these AU-rich SLSs may serve additional functions, such as mRNA stabilization, as point mutations in transcription terminators are known to affect the stability of upstream RNA segments [43,44].

Bacteria other than low-GC firmicutes do not feature similar AT-rich terminator-like structures, still the distribution of SLSs within IGRs is clearly non random. When the frequency of SLSs is analyzed according to the type of IGR, all bacteria show a strong preference for SLSs within convergent, i.e. flanking the 3'-end of CDSs, rather than divergent IGRs (Fig. 6). Furthermore within unidirectional IGRs, higher stability intergenic SLSs are also preferentially found within the 50 bp tract immediately following the stop codon of the neighbouring CDS (Fig. 7). This distribution strongly favours the notion that most higher stability intergenic SLSs are transcribed, and may therefore function at the RNA level. Although termination is the expected role for a large fraction of them, especially in bacteria where *rho* dependent termination is not relevant, their number and the observed sequence features leave open the possibility of additional roles, such as RNA stabilization, translational regulation by riboswitches and attenuators [19,16]. Alternatively these SLS may be targeted by specific nucleases and rapidly degraded, thus functioning as RNA instability determinants. Finally, it must be recalled that some intergenic SLSs may be transcribed independently of the flanking genes. In recent years several groups provided support to the notion that prokaryotic intergenic sequences encode a variety of small, non-coding (nc) RNAs fulfilling diverse functions [reviewed [45]]. It will be of interest to assess whether selected intergenic SLSs may lead to the identification of novel nc-RNAs in RNA populations.

Some SLSs show strong similarity with each other, and may be grouped into families of repetitive sequences. Here we describe *Bor* sequences (Fig. 8), a set of palindromic elements, over-represented in all *Bordetellae*, which recall in length and sequence the *E. coli* REP sequences. *Bor* containing RNA may fold into hairpins similar to REP RNA, and possibly play an analogous role, i.e. the stabilization of the cotranscribed mRNA [34]. The larger *Ype* and

Efa elements (Fig. 8) are members of less numerous DNA families spread in the genomes of *Y. pestis* and *E. faecalis*, respectively. These sequences are similar in size and abundance to other intergenic repeats, such as NEMIS in *N. meningitidis* [22] and ERIC in *Yersiniae* [23], which are cotranscribed with flanking genes and may fold into similarly organized RNA hairpins. Preliminary data indicate that both *Ype* and *Efa* RNA elements may indeed enhance the stability of cotranscribed mRNA sequences [De Gregorio E, Silvestro G and Di Nocera PP, unpublished results]. Quantitatively, members of these families only account for a small fraction of intergenic SLSs. As revealed by a preliminary BLAST analyses (not shown), further substantial similarities may be detected within the identified SLSs. Each of these families may therefore be extended, by including more elements sharing sequence similarity, but not initially found because of the presence of defective, or less pronounced, secondary structures. Further work will be necessary to eventually obtain a systematic classification of bacterial DNA families spanning, or coinciding with SLSs.

Conclusion

An in-depth analysis of SLS features and distribution was carried out in 40 different bacterial species. Data suggest that an evolutionary pressure preserved specific non random populations of higher stability SLSs in most of the analyzed genomes. Many of these sequences are plausibly transcribed, and may be involved in transcriptional and/or post-transcriptional control. Specific SLS containing sequences are members of three previously undescribed families of repeated sequences found in *Yersiniae*, *Bordetellae* and *Enterococci*.

Methods

Genomic sequence data

Complete genomic sequences and their annotations about CDS, rRNA and tRNA were downloaded from the online repository made available at The Institute for Genomic Research (TIGR). Automatic annotations have been stored into a SQL database (SLS-DB), for further analysis. PostgreSQL has been used as the SQL Database Management System [46], according to techniques previously described [47,48].

SLS identification

SLS identification was performed by using the program *rnamotif* of the package RNAMOTIF, version 2.1.2 [49] according to the following rules:

-GU pairing in the stem was allowed

-the minimal stem length was 12 bp

-loop length could vary from 5 to 100 nt

-1 bulged or 1 mispaired base, at least two matches away from the ends of the stem, was allowed.

As a consequence of the constraints imposed, the smallest SLS that could be found is 29 bp. Due to the allowance for GU pairing, *mamotif* had to be run on both strands of the input sequence. Completely overlapping SLSs were discarded by 6 runs of the *rnprune* tool, also from the RNAMOTIF package.

The Gibbs free energy (*dG*) of each SLS containing region was calculated by calling the built-in function *efn2* of *rnamotif*. The minimum free energy with no constraint for SLS formation was obtained by running the program *mfold* developed by Zuker and coworkers [50] on the SLS sequences.

SLS pre-filtering

When two or more SLSs where found overlapping, only the most stable one was counted.

Intergenic regions

Intergenic regions (IGRs) were derived, stored and annotated into the SLS-DB, according to the ORF collection provided by TIGR. For some tests, IGRs of size ranging from 29 to 500 bp were selected (see also legend to Figure 6).

Shuffled genomes

The program *Shufflet* [51] was used to generate random sequences and to shuffle bacterial genomes by preserving k-lets of different lengths. In order to shuffle sequences with k-let higher than 6, *Shufflet* was compiled by setting the variable MAXORDER to 15. An alternative shuffling method (referred to as DS in Fig. 3), was used to take into account the information about protein coding sequences. Basically coding regions were shuffled by using the program *Dicodonshuffle* [15], while *Shufflet* set to k-let = 2 was used for non coding regions.

Abbreviations

bp, base pair

nt, nucleotide

Mb, megabase

Authors' contributions

MP created the pipeline for identification and automatic annotation of SLSs within bacterial sequences and contributed to sequence and statistical analysis. GS retrieved the sequences and other informations and provided manual annotation and analysis. AB contributed to the development of the pipeline. PPDN and GP conceived and

coordinated the study. All authors read and approved the final manuscript.

Acknowledgements

We wish to thank Luca Cozzuto for suggestions and useful discussions, Tommaso Russo, Carmelo Bruno Bruni, Concetta Pietropaolo and Maria Stella Carlomagno for critically reading the manuscript. Informatic support by Gianluca Busiello is also acknowledged.

This work has been supported by a grant of the agency PRIN 2004 to PPDN, by a grant of the agency PRIN 2005 to GP, by a MIUR grant to CEINGE (12/2000) and a MIUR FIRB (LITBIO).

References

1. Borodina I, Krabben P, Nielsen J: **Genome-scale analysis of Streptomyces coelicolor A3(2) metabolism.** *Genome Res* 2005, **15**:820-829.
2. Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, Yokoyama K, Han CG, Ohtsubo E, Nakayama K, Murata T, Tanaka M, Tobe T, Iida T, Takami H, Honda T, Sasakawa C, Ogasawara N, Yasunaga T, Kuhara S, Shiba T, Hattori M, Shinagawa H: **Complete genome sequence of enterohemorrhagic Escherichia coli O157:H7 and genomic comparison with a laboratory strain K-12.** *DNA Res* 2001, **8**:11-22.
3. Dobrindt U, Agerer F, Michaelis K, Janka A, Buchrieser C, Samuelson M, Svanborg C, Gottschalk G, Karch H, Hacker J: **Analysis of genome plasticity in pathogenic and commensal Escherichia coli isolates by use of DNA arrays.** *J Bacteriol* 2003, **185**:1831-1840.
4. Grifantini R, Bartolini E, Muzzi A, Draghi M, Frigimelica E, Berger J, Ratti G, Petracca R, Galli G, Agnusdei M, Giuliani MM, Santini L, Brunelli B, Tettelin H, Rappuoli R, Randazzo F, Grandi G: **Previously unrecognized vaccine candidates against group B meningococcus identified by DNA microarrays.** *Nat Biotechnol* 2002, **20**:914-921.
5. Frank AC, Amiri H, Andersson SG: **Genome deterioration, loss of repeated sequences and accumulation of junk DNA.** *Genetica* 2002, **115**:1-12.
6. Achaz G, Coissac E, Netter P, Rocha EP: **Associations between inverted repeats and the structural evolution of bacterial genomes.** *Genetics* 2003, **164**:1279-1289.
7. Audit B, Ouzounis CA: **From genes to genomes, universal scale-invariant properties of microbial chromosome organisation.** *J Mol Biol* 2003, **332**:617-633.
8. Rocha EP, Danchin A: **Gene essentiality determines chromosome organisation in bacteria.** *Nucleic Acids Res* 2003, **31**:6570-6577.
9. Chain PS, Carniel E, Larimer FW, Lamerdin J, Stoutland PO, Regala WM, Georgescu AM, Vergez LM, Land ML, Motin VL, Brubaker RR, Fowler J, Hinnebusch J, Marceau M, Medigue C, Simonet M, Chenal-Francisque V, Souza B, Dacheux D, Elliott JM, Derbise A, Hauser LJ, Garcia E: **Insights into the evolution of Yersinia pestis through whole-genome comparison with Yersinia pseudotuberculosis.** *Proc Natl Acad Sci U S A* 2004, **101**:13826-13831.
10. Hallin PF, Ussery DW: **CBS Genome Atlas Database: a dynamic storage for bioinformatic results and sequence data.** *Bioinformatics* 2004, **20**:3682-3686.
11. van Belkum A, van Leeuwen W, Scherer S, Verbrugh H: **Occurrence and structure-function relationship of pentameric short sequence repeats in microbial genomes.** *Res Microbiol* 1999, **150**:617-626.
12. Salaun L, Linz B, Suerbaum S, Saunders NJ: **The diversity within an expanded and redefined repertoire of phase-variable genes in Helicobacter pylori.** *Microbiology* 2004, **150**:817-830.
13. Seffens W, Digby D: **RNAs have greater negative folding free energies than shuffled or codon choice randomized sequences.** *Nucleic Acids Res* 1999, **27**:1578-1584.
14. Workman C, Krogh A: **No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution.** *Nucleic Acids Res* 1999, **27**:4816-4822.
15. Katz L, Burge CB: **Widespread selection for local RNA secondary structure in coding regions of bacterial genes.** *Genome Res* 2003, **13**:2042-2051.
16. Henkin TM, Yanofsky C: **Regulation by transcription attenuation in bacteria, how RNA provides instructions for transcription termination/antitermination decisions.** *Bioessays* 2002, **8**:700-707.
17. Berg BL, Baron C, Stewart V: **Nitrate-inducible formate dehydrogenase in E. coli K12. II. Evidence that a mRNA stem-loop structure is essential for decoding opal (UGA) as selenocysteine.** *J Biol Chem* 1991, **33**:22386-22391.
18. Martínez-Abarca F, Toro N: **Group II introns in the bacterial world.** *Mol Microbiol* 2000, **38**:917-926.
19. Nudler E, Mironov AS: **The riboswitch control of bacterial metabolism.** *Trends Biochem Sci* 2004, **29**:11-17.
20. Coburn GA, Mackie GA: **Degradation of mRNA in Escherichia coli: an old problem with some new twists.** *Prog Nucleic Acid Res Mol Biol* 1999, **62**:55-108.
21. Gilson E, Saurin W, Perrin D, Bachellier S, Hofnung M: **The BIME family of bacterial highly repetitive sequences.** *Res Microbiol* 1991, **142**:217-222.
22. De Gregorio E, Abrescia C, Carlomagno MS, Di Nocera PP: **Ribonuclease III-mediated processing of specific Neisseria meningitidis mRNAs.** *Biochem J* 2003, **374**:799-805.
23. De Gregorio E, Silvestro G, Petrillo M, Carlomagno MS, Di Nocera PP: **Genomic organization and functional properties of ERIC DNA repeats in Yersiniae.** *J Bact* 2005, **187**:7945-7954.
24. Krasilnikov AS, Podtelezhnikov A, Vologodskii A, Mirkin SM: **Large-scale effects of transcriptional DNA supercoiling in vivo.** *J Mol Biol* 1999, **292**:1149-1160.
25. Jin R, Novick RP: **Role of the double-strand origin cruciform in pT181 replication.** *Plasmid* 2001, **46**:95-105.
26. Yamada K, Ariyoshi M, Morikawa K: **Three-dimensional structural views of branch migration and resolution in DNA homologous recombination.** *Curr Opin Struct Biol* 2004, **14**:130-137.
27. Smith HO, Gwinn ML, Salzberg SL: **DNA uptake signal sequences in naturally transformable bacteria.** *Res Microbiol* 1999, **150**:603-616.
28. Davidsen T, Rodland EA, Lagesen K, Seeberg E, Rognes T, Tonjum T: **Biased distribution of DNA uptake sequences towards genome maintenance genes.** *Nucleic Acids Res* 2004, **32**:1050-1058.
29. Parkhill J, Achtman M, James KD, Bentley SD, Churcher C, Klee SR, Morelli G, Basham D, Brown D, Chillingworth T, Davies RM, Davis P, Devlin K, Feltwell T, Hamlin N, Holroyd S, Jagels K, Leather S, Moule S, Mungall K, Quail MA, Rajandream MA, Rutherford KM, Simmonds M, Skelton J, Whitehead S, Spratt BG, Barrell BG: **Complete DNA sequence of a serogroup A strain of Neisseria meningitidis Z2491.** *Nature* 2000, **404**:502-506.
30. Claverie JM, Ogata H: **The insertion of palindromic repeats in the evolution of proteins.** *Trends Biochem Sci* 2003, **28**:75-80.
31. Ermolaeva MD, Khalak HG, White O, Smith HO, Salzberg SL: **Prediction of transcription terminators in bacterial genomes.** *J Mol Biol* 2000, **301**:27-33.
32. Lesnik EA, Sampath R, Levene HB, Henderson TJ, McNeil JA, Ecker DJ: **Prediction of rho-independent transcriptional terminators in Escherichia coli.** *Nucleic Acids Res* 2001, **29**:3583-3594.
33. Unniraman S, Prakash R, Nagaraja V: **Conserved economics of transcription termination in eubacteria.** *Nucleic Acids Res* 2002, **30**:675-684.
34. Bachellier S, Clement JM, Hofnung M: **Short palindromic repetitive DNA elements in enterobacteria: a survey.** *Res Microbiol* 1999, **150**:627-639.
35. Aranda-Olmedo I, Tobes R, Manzanera M, Ramos JL, Marques S: **Species-specific repetitive extragenic palindromic (REP) sequences in Pseudomonas putida.** *Nucleic Acids Res* 2002, **30**:1826-1833.
36. Qi D, Cuticchia AJ: **Compositional symmetries in complete genomes.** *Bioinformatics* 2001, **17**:557-559.
37. Baisnee PF, Hampson S, Baldi P: **Why are complementary DNA strands symmetric?** *Bioinformatics* 2002, **18**:1021-1033.
38. Robin S, Daudin JJ, Richard H, Sagot MF, Schbath S: **Occurrence probability of structured motifs in random sequences.** *J Comput Biol* 2002, **9**:761-773.
39. Ingham CJ, Tennis J, Fumeaux PA: **Autogenous regulation of transcription termination factor Rho and the requirement for Nus factors in Bacillus subtilis.** *Mol Microbiol* 1999, **31**:651-663.

40. Washburn RS, Marra A, Bryant AP, Rosenberg M, Gentry DR: **Rho is not essential for viability or virulence in Staphylococcus aureus.** *Antimicrob Agents Chemother* 2001, **45:**1099-1103.
41. Yang MT, Scott HB 2nd, Gardner JF: **Transcription termination at the thr attenuator Evidence that the adenine residues upstream of the stem and loop structure are not required for termination.** *J Biol Chem* 1995, **270:**23330-23336.
42. Carlomagno MS, Riccio A, Bruni CB: **Convergently functional, Rho-independent terminator in Salmonella typhimurium.** *J Bacteriol* 1985, **163:**362-368.
43. Abe H, Aiba H: **Differential contributions of two elements of rho-independent terminator to transcription termination and mRNA stabilization.** *Biochimie* 1996, **78:**1035-1042.
44. Cisneros B, Court D, Sanchez A, Montanez C: **Point mutations in a transcription terminator, lambda tI, that affect both transcription termination and RNA stability.** *Gene* 1996, **181:**127-133.
45. Storz G, Altuvia S, Wassarman KM: **An abundance of RNA regulators.** *Annu Rev Biochem* 2005, **74:**199-217.
46. **PostgreSQL Project** 2005 [http://www.postgresql.com/index.html].
47. Boccia A, Petrillo M, di Bernardo D, Guffanti A, Mignone F, Confalonieri S, Luzi L, Pesole G, Paolella G, Ballabio A, Banfi S: **DG-CST (Disease Gene Conserved Sequence Tags), a database of human-mouse conserved elements associated to disease genes.** *Nucleic Acids Res* 2005:D505-10.
48. Milanesi L, Petrillo M, Sepe L, Boccia A, D'Agostino N, Passamano M, Di Nardo S, Tasco G, Casadio R, Paolella G: **Systematic analysis of human kinase genes: a large number of genes and alternative splicing events result in functional and structural diversity.** *BMC Bioinformatics* 2005, **6(Suppl 4):**S20.
49. Macke TJ, Ecker DJ, Gutell RR, Gautheret D, Case DA, Sampath R: **RNAMotif, an RNA secondary structure definition and search algorithm.** *Nucleic Acids Res* 2001, **29:**4724-4735.
50. Mathews DH, Sabina J, Zuker M, Turner DH: **Expanded sequence dependence of thermodynamic parameters provides robust prediction of rna secondary structure.** *J Mol Biol* 1999, **288:**910-940.
51. Coward E: **Shufflet, shuffling sequences while conserving the k-let counts.** *Bioinformatics* 1999, **12:**1058-1059.

JOURNAL OF BACTERIOLOGY, Nov. 2006, p. 7876–7884 Vol. 188, No. 22
0021-9193/06/$08.00+0 doi:10.1128/JB.00942-06

Structural Organization and Functional Properties of Miniature DNA Insertion Sequences in Yersiniae[▽]

Eliana De Gregorio, Giustina Silvestro, Rossella Venditti, Maria Stella Carlomagno, and Pier Paolo Di Nocera*

Dipartimento di Biologia e Patologia Cellulare e Molecolare, Facoltà di Medicina, Università Federico II, Via S. Pansini 5, 80131 Napoli, Italy

Received 29 June 2006/Accepted 14 August 2006

YPALs (*Yersinia pal*indromic sequences) are miniature DNA insertions scattered along the chromosomes of yersiniae. The spread of these intergenic repeats likely occurred via transposition, as suggested by the presence of target site duplications at their termini and the identification of syntenic chromosomal regions which differ in the presence/absence of YPAL DNA among *Yersinia* strains. YPALs tend to be inserted closely downstream from the stop codon of flanking genes, and many YPAL targets overlap rho-independent transcriptional terminator-like sequences. This peculiar pattern of insertion supports the hypothesis that most of these repeats are cotranscribed with upstream sequences into mRNAs. YPAL RNAs fold into stable hairpins which may modulate mRNA decay. Accordingly, we found that YPAL-positive transcripts accumulate in *Yersinia enterocolitica* cells at significantly higher levels than homologous transcripts lacking YPAL sequences in their 3′ untranslated region.

Bacterial insertion sequences (ISs) are mobile genetic elements ranging in size from 800 to 2,500 bp which are widely distributed among bacteria (21, 5). Typically, ISs encode a transposase which mediates their movement and feature terminal inverted repeats (TIRs) 10 to 50 bp long, which serve as recognition sites for the transposase. Most ISs generate short direct repeats (target site duplications [TSDs]) at the point of insertion. For each element, the length of the TSD is fixed and ranges from 2 to 13 bp. Differences in the structural organization, coding capacity, and transposition properties make it possible to sort ISs into approximately 20 major subfamilies (21).

In recent years, it has emerged that small IS-like sequences called MITEs (for *mi*niature *t*ransposable elements) may be a relevant genome component in several eukaryotic species (16, 20). These elements characteristically measure 150 to 400 bp, carry long TIRs, and are flanked by TSDs of variable lengths. MITEs likely represent deletion derivatives of longer, autonomous ISs, which have been mobilized by tranposases encoded by partly related mobile donor elements (20). Several MITE families have also been identified in archaeabacteria (5). Three families of MITEs have been described for eubacteria. RUP (*r*epeat *u*nit of *p*neumococcus) elements are spread in ~100 copies in the genome of the *Streptococcus pneumoniae* 4 strain, and it has been proposed that their mobilization is mediated by the IS*630*-Spn1 element (24). NEMIS (*Ne*isseria *m*iniature *i*nsertion *s*equences) are 108- to 158-bp-long repeats which make up ~2% of the *Neisseria meningitidis* genome. In contrast to RUP elements, which are mostly interspersed with other repeated DNA sequences, NEMIS sequences are frequently located next to *Neisseria* genes and are transcribed into mRNAs. Hairpins formed by the pairing of NEMIS TIRs are targeted by RNase III, and this interaction regulates sets of *N. meningitidis* genes at the posttranscriptional level (12, 13).

A different type of eubacterial MITE is represented by ERIC (*e*nterobacterial *r*epetitive *i*ntergenic *c*onsensus) sequences. These elements, which measure 69 to 127 bp, are moderately abundant (20 to 25 copies) in the genomes of several enterobacteria (19) but are overrepresented in the genomes of yersiniae (15). In *Yersinia* species, most ERIC sequences are inserted immediately downstream from open reading frames (ORFs) and hence are cotranscribed into mRNAs with upstream genes. ERIC RNA may fold into robust RNA hairpins, and changes in their relative position and orientation within the mRNA molecule have been shown to differently influence the processing rate of neighboring RNA segments (15).

In this paper we report on the genomic and functional organization of a novel family of repeated DNA sequences present in yersiniae. The YPAL (*Yersinia pal*indromic elements) sequences share properties with both ERIC and NEMIS sequences and represent a novel example of MITEs which can play a role as RNA elements in posttranscriptional control.

MATERIALS AND METHODS

Bacterial strains and growth conditions. *Yersinia enterocolitica* strains used in this work and growth conditions have been described previously (15). The *Escherichia coli* strains W3110 and HT115 (W3110 *mc-14*::ΔTn*10*) are described in the work of Takiff et al. (28). The strains SK5006 (*thr leu* pDK39 Cm^r *mb-500*), SK5003 (*thr leu pnp-7 rnb-500*), and SK5695 (*thr leu rne-1*) are described in reference 2.

RNA analyses. Total bacterial RNA was purified on RNeasy columns (QIAGEN). YPAL-positive transcripts were monitored by Northern analyses by using as probes radiolabeled DNA segments 300 to 400 bp in length resulting from the amplification of *Y. enterocolitica* strain Ye161 DNA with pairs of gene-specific oligonucleotides. Processed YPAL RNA species were detected by high-resolution Northern analyses as described previously (14). Reverse transcription (RT)-PCR analyses were carried out by reverse transcribing 200 nanograms of total *Y.*

* Corresponding author. Mailing address: Dipartimento di Biologia e Patologia Cellulare e Molecolare, Facoltà di Medicina, Università Federico II, Via S. Pansini 5, 80131 Naples, Italy. Phone: 0039-81-7462059. E-mail: dinocera@unina.it.

▽ Published ahead of print on 8 September 2006.

TABLE 1. Primers used in this work

Primer function	Name	Sequence (5′ to 3′)
RT-PCR	rRNA for	GCGCTTAACGTGGGAACTGC
	rRNA rev	TCAGTCTTTGTCCAGGGGGC
	1047 for	TATGGTGGCGGTCTTCCTGG
	1047 rev	TTTCGGATCGTTAGGCACGC
	686-A for	ACTTATTCTGTCGGGCGCTG
	686-A rev	TAACCGTCTTACCTTCCGCC
Northern probe	408-A for	CATTTGCCCGTTTCCAGATC
	408-A rev	ACGCGGATCCGATCAAATAC
	686-B for	TGTGGTGACTGGTGAGATGG
	686-B rev	CGGTTTTCTTGCTCTGCTAC
	ypal for	TGAGGTTAATGACAAAGTGCCCGTA
	ypal rev	TCAGACTGCTGACAAACCTCAAGGA
	Ypal45 for	TAATACGACTCACTATAGGGAGAACCCGCGAAATGCGGGTTGAGG
	Ypal45 rev	ATTTAGGTGACACTATAGAATACACCCGCATTTCGCGGGTTCAGA
YPAL antisense RNA template	408-B for	ATTTAGGTGACACTATAGAATACACCTTGTCATTAGATGGGGACCC
	408-B rev	TAATACGACTCACTATAGGGAGAAAAATGATAAGCCGCAACGCTAG
	686-C for	ATTTAGGTGACACTATAGAATACGCGAAGGCCGAGTTGTTGAAGAG
	686-C rev	TAATACGACTCACTATAGGGAGATTAGAGGAGGGCTATCCGGTGGG
YPAL sense RNA template	408-C for	TAATACGACTCACTATAGGGAGACAGTGCCTGGATATACCTTTACC
	408-C rev	ATTTAGGTGACACTATAGAATACCACAGCAAACTGACCAAAGAATA
	408-D for	TAATACGACTCACTATAGGGAGACAGTGCCTGGATATACCTTTACC
	408-D rev	ATTTAGGTGACACTATAGAATACGCCGAAATCCTGTCCTCTTTCGA

enterocolitica RNA by random priming. The resulting cDNAs were amplified by using pairs of gene-specific oligonucleotides. One oligonucleotide within each pair had been previously ^{32}P end labeled at the 5′ terminus with polynucleotide kinase. To adequately monitor gene-specific RNA levels by RT-PCR, cDNAs were amplified under nonsaturating cycling conditions, and low-cycle (14 to 18 cycles) PCR analyses were performed for each set of amplified genes. As an internal control, a 140-nucleotide (nt)-long amplified segment of 16S rRNA from *Y. enterocolitica* was used. rRNA primers were added to the PCRs two to four cycles before the amplification of the ORF-specific mRNAs was completed. Amplified DNA fragments were separated using 6% polyacrylamide-8 M urea gels and quantitated by phosphorimagery.

The uniformly ^{32}P-labeled RNAs used as probes in the RNase protection assays were obtained by transcribing in vitro linear DNA templates as described previously (12). DNA templates were obtained by PCR amplification of DNA derived from the *Y. enterocolitica* strain Ye161. One of the two primers included in its 5′ end region the sequence of the T7 RNA polymerase promoter (see underlined residues in Table 1). Twenty micrograms of total RNA derived from *Y. enterocolitica* cells was mixed with ^{32}P-labeled antisense RNA probes in 30 μl of hybridization buffer (75% formamide, 20 mM Tris, pH 7.5, 1 mM EDTA, 0.4 M NaCl, 0.1% sodium dodecyl sulfate). Samples were incubated at 95°C for 5 min, cooled down slowly, and kept at 45°C for 16 h. After a 60-min incubation at 33°C with RNase T1 (2 μg/ml), samples were treated with proteinase K (50 μg/ml) for 15 min at 37°C, extracted once with phenol, precipitated with ethanol, resuspended in 80% formamide, and loaded onto 6% polyacrylamide-8 M urea gels.

To obtain the YPAL-positive RNA control electrophoresed in lane 3 of Fig. 6B, the YPAL 45 element and flanking sequences were amplified by using a primer including the T7 RNA polymerase promoter.

In vitro RNA cleavage assays. Uniformly ^{32}P-labeled RNAs were obtained by transcribing in vitro linear DNA templates with T7 RNA polymerase in the presence of radiolabeled [^{32}P]UTP. Templates were obtained by PCR amplification of *Yersinia* DNA with pairs of 50-mers, one of which included the T7 RNA polymerase promoter in its 5′ end region. The T7 promoter sequence allowed the in vitro synthesis of RNA substrates for processing assays. Degradation assays with whole bacterial cell extracts were carried out essentially as described previously (11).

Oligonucleotides. Sequences of all PCR primers used in this study are reported in Table 1.

Computer analyses. A YPAL element found next to ERIC sequences in the sequenced 8081 strain of *Y. enterocolitica* (www.sanger.ac.uk/Projects/Y_enterocolitica) was used as a query in BLAST searches to fetch homologous DNA segments from the genomes of the same strain and the *Yersinia pestis* CO92 strain (26).

Consensus sequences from multiple alignments of YPAL family members were established with the program CONS of the EMBOSS package. Secondary structure modeling was done using the Mulfold program (www.bioinfo.rpi.edu/applications/mfold), which predicts RNA secondary structure by free energy minimization (30).

RESULTS

Organization of YPAL repeats. During analyses aimed at the characterization of the family of ERIC repeats in *Y. enterocolitica*, we fortuitously encountered a palindromic intergenic sequence located next to an ERIC element. As revealed by BLAST searches, the latter was found to be a member of an abundant DNA family spread throughout the chromosomes of both *Y. enterocolitica* and *Y. pestis*. A member of this family had been earlier called YPAL (3), and we will keep the acronym for clarity. The palindromic nature of YPAL repeats is highlighted in Fig. 1. As RNA sequences, i.e., allowing for the formation of GU base pairs, YPALs can fold, regardless of their orientation, into stable, low-free-energy branched hairpins (Fig. 1). YPALs are present in ~100 copies in sequenced *Yersinia* genomes (Fig. 2) and feature a modular organization. Complete repeats include the same external modules (A and H modules in Fig. 2) but a different set of internal modules. Eight different YPAL subfamilies could be distinguished by sequence alignments. The highest numbers of repeats belong to subfamilies 2 and 8. The size of subfamily 2, which includes elements measuring 167 bp, is similar in *Y. enterocolitica* and *Y. pestis*. In contrast, the size of subfamily 8, made by 130-bp-long repeats, is much reduced in *Y. pestis*. The 98-bp-long module I, defining the members of the subfamily 8, comes in two sequence variants (Ia and Ib), neither of which bears significant

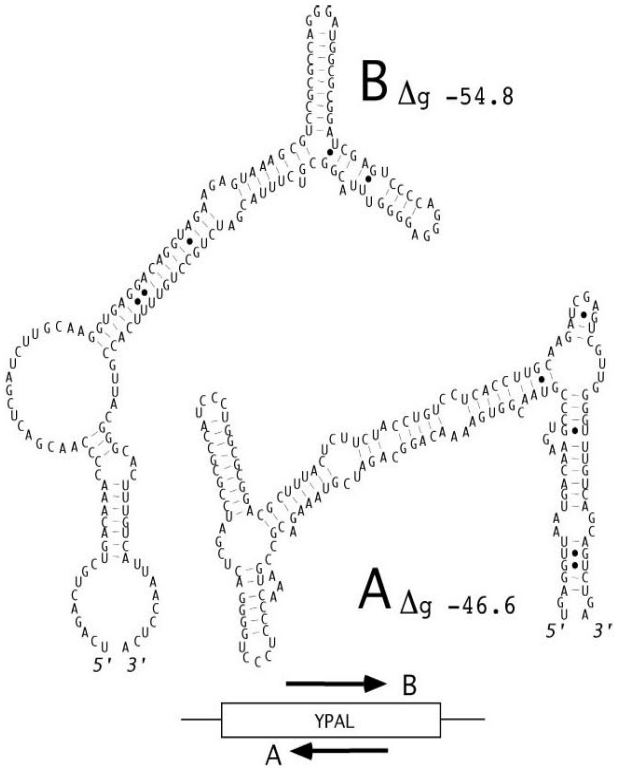


FIG. 1. Secondary structure of YPAL repeats. Hairpin structures formed by a representative YPAL element inserted in the mRNA in the two possible A and B orientations are shown. GU pairing is highlighted by dots. The Gibbs free energy of each hairpin is indicated. Secondary structure modeling was done using the Mulfold program. The element analyzed is the 167-bp-long YPAL 45 element (see the text).

sequence homology to the bodies of other YPALs (Fig. 2). Interestingly, modules Ia and Ib were found in 2/3 and 1/3 of the elements identified in the *Y. enterocolitica* 8081 strain, respectively. In contrast, members of YPAL subfamily 8 identified in the *Y. pestis* CO92 strain all feature module Ia.

In the chromosome of *Y. enterocolitica* strain 8081, six members of subfamily 2 are interrupted at the same position between modules E and F (Fig. 2) by the insertion of 1,376 bp of foreign DNA. Homology searches run at the IS finder site (http://www-is.biotoul.fr) enabled us to establish that the intervening DNA sequence is ISYen*1*, a low-copy-number IS found exclusively in the *Y. enterocolitica* species (27). The two additional members of the ISYen*1* family resident in the chromosome of *Y. enterocolitica* strain 8081 are not associated with YPAL DNA. The insertion of ISYen*1* is not accompanied by the duplication of target site sequences.

YPALs are miniature mobile DNA elements. YPAL repeats show the typical structure of miniature DNA insertion sequences. The hypothesis that their genomic spread occurred via transposition was supported by the presence of TSDs at the termini of most elements (Fig. 3). In contrast to the majority of ISs, whose insertion is accompanied by the generation of TSDs of fixed length, most (85%) YPALs are flanked by TSDs of variable size (Fig. 3). The length of the duplicated segment ranged from 3 to 26 bp, with the most frequently found TSDs measuring 18 or 20 bp (Fig. 3B). YPAL targets do not exhibit sequence homologies. However, it is remarkable that they span, or are located next to, regions of dyad symmetry. Intriguingly, YPAL target sites often correspond to the DNA counterpart of rho-independent transcription terminators, the RNA structures responsible for the detachment of the transcribing RNA polymerase from the DNA template. This conclusion was supported both by the presence of runs of thymidines immediately next to palindromes targeted by YPAL (Fig. 3A) and by the finding that most targets were found located 20 to 50 bp downstream from the stop codon of annotated ORFs in sequenced *Yersinia* genomes.

The notion that YPALs are mobile DNA elements was directly supported by in silico analyses. The chromosomal distribution of YPALs varies significantly between *Y. pestis* and *Y. enterocolitica*, and only a few syntenic YPAL-positive regions are shared by the two species. This is not surprising, since *Y. pestis* and *Y. enterocolitica* are evolutionarily distant. In contrast, *Y. pestis* and *Yersinia pseudotuberculosis* are evolutionarily closer, *Y. pestis* being regarded as a clone that evolved from *Y. pseudotuberculosis* only 1,500 to 20,000 years ago (1). Com-

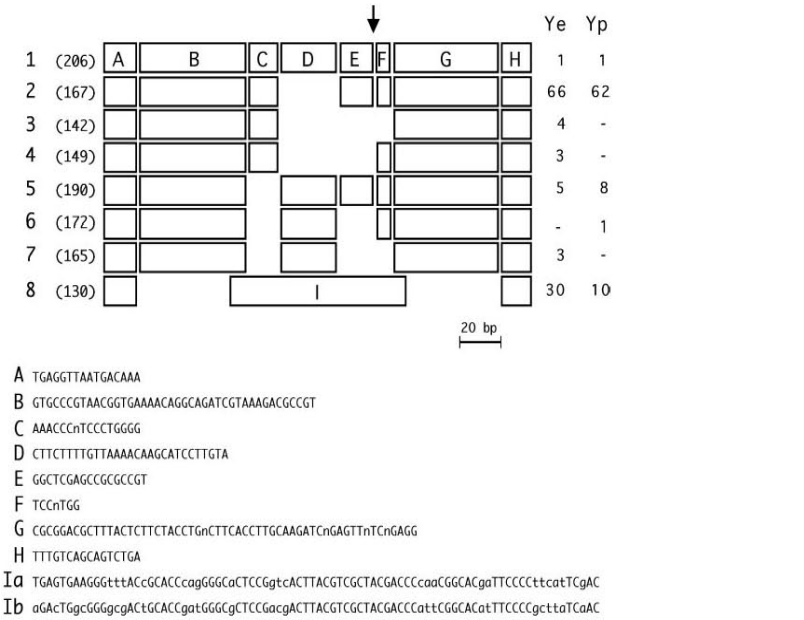


FIG. 2. Structural organization of YPAL elements. The nine modules labeled A to I found in YPAL elements are shown. Numbers to the left (1 to 8) denote YPAL subfamilies. The length in bp of the elements of each subfamily is given in parenthesis. The number of elements within each subfamily found in the *Y. enterocolitica* 8081 (Ye) and the *Y. pestis* CO92 (Yp) strains is reported. The consensus sequence content of the nine modules derived from the comparison of *Y. enterocolitica* YPAL repeats is shown at the bottom. Two versions of the module I (Ia and Ib) are reported. Uppercase residues denote sequence identity. An arrow marks the site of insertion of ISYen*1*.

parisons of the genomes of the *Y. pestis* CO92 and *Y. pseudotuberculosis* IP32953 strains led to the identification of homologous chromosomal regions which carry YPAL DNA only in one of the two species (Fig. 4). The finding of a single copy of the TSD at the empty chromosomal sites (Fig. 4) corroborated the notion that target site duplication is induced upon insertion, ruling out the formal possibility that YPALs could preferentially integrate between preexisting tandem duplications. Empty and filled YPAL chromosomal sites also were subsequently identified among different laboratory strains of *Y. enterocolitica* by DNA sequence analyses (data not shown).

Changes in the distribution of YPAL sequences among *Yersinia* strains might also reflect the loss of YPAL DNA from specific sites, prompted either by specific recombination between the flanking TSDs or by replicational slippage between the TSDs.

YPAL elements transcribed into RNA. All the YPAL sequences found in the *Y. enterocolitica* chromosome are located within intergenic regions. Remarkably, most elements (70 repeats) are inserted close to the stop codons of flanking ORFs. This suggests that YPALs may be cotranscribed along with upstream coding sequences and that their ability to fold into RNA hairpins may have functional significance.

To address this issue, total RNA from *Y. enterocolitica* strain 161 was analyzed by Northern blotting. We monitored two members of YPAL subfamily 2 (repeats 45 and 9) located 89 bp and 56 bp downstream from the stop codon of the IMP dehydrogenase/GMP reductase gene encoded by ORF YE686 and the pyrophosphatase gene encoded by ORF YE408, respectively (Fig. 5A). The ORF YE686 probe detected a major mRNA species measuring ~1,380 nt, which plausibly spans both ORF YE686 and the flanking YPAL 45 element. Two transcripts corresponding to ORF YE408, measuring ~650 and ~850 nt, were detected by the ORF YE408 probe. The sizes of the bands are in accord with the hypothesis that the 850-nt-long transcript spans both ORF YE408 and the flanking YPAL 9 element and that the 650-nt-long RNA species may originate with the removal of YPAL sequences from the transcript (Fig. 5A).

Northern data were complemented by RNase protection experiments (Fig. 5B). When YE686 transcripts were monitored, a predominant RNA species, corresponding to the accumulation in *Y. enterocolitica* cells of transcripts encompassing both ORF YE686 and the flanking YPAL 45 element, was protected (Fig. 5B, lane 3). In contrast, when YE408 transcripts were monitored, two bands of protection were detected (Fig. 5B, lane 6). The size of the protected species is in accord with the hypothesis that two predominant mRNA segments spanning ORF YE408 accumulate within *Y. enterocolitica* cells, one of which encompasses the flanking YPAL 9 element. From the relative intensities of the bands, it could be inferred that the latter RNA species is two to three times more abundant than the 400-nt-long RNA species lacking YPAL sequences.

Processing of YPAL-positive transcripts. Both the Northern and RNase protection experiments suggest that YPAL sequences may be cleaved off from YPAL-positive transcripts in

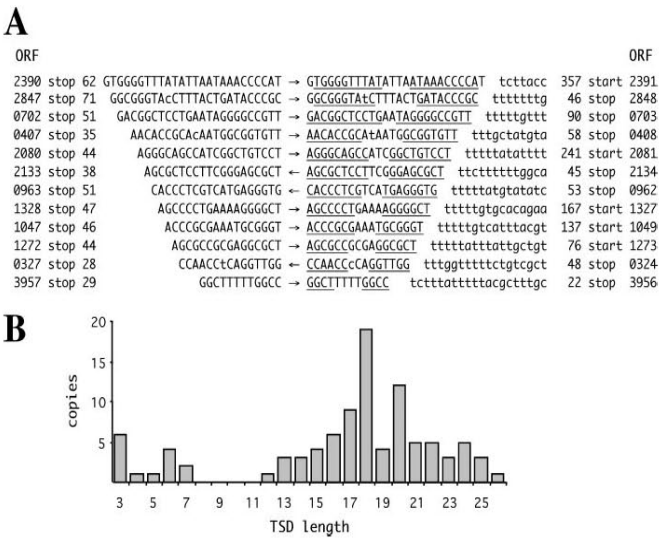


FIG. 3. YPAL target sites. (A) Some of the TSDs induced by the insertion of YPALs in the chromosome of the *Y. enterocolitica* 8081 strain, as well as flanking T-rich segments, are shown. Base changes at TSDs are in lowercase letters. Arrows mark YPAL sequences and their orientations. Regions of dyad symmetry are underlined. ORFs flanking YPAL sequences are indicated. The distance in bp separating YPAL termini from either the start or the stop codon of each ORF is shown. (B) Length variation of TSDs flanking YPAL elements in the *Y. enterocolitica* 8081 strain.

vivo. This hypothesis was confirmed by in vitro RNA degradation assays. A DNA fragment spanning the 3′ end region of ORF YE408 and the flanking YPAL 9, the 167-nt-long element belonging to the abundant subfamily 2 already analyzed in Fig. 5, was used as a template to synthesize in vitro-radiolabeled YPAL-positive RNAs of known length. By using the Mulfold program, we checked that foreign sequences present in the RNA synthesized in vitro did not interfere with the formation of the YPAL RNA hairpins shown in Fig. 1. Challenging the RNA obtained (408-A RNA) with *Y. enterocolitica* whole-cell extracts resulted in the accumulation of three major RNA species (Fig. 6A, lane 2). Bands a to c had the size expected for RNA moieties generated by cleavages occurring at the boundaries of YPAL 9 sequences. Results were corroborated by experiments carried out with a substrate similar to the 408-A RNA but lacking 161 nt at the 3′ end (408-B RNA) (Fig. 6A, lane 4). By using this shortened RNA substrate, we were able to detect the same a and b bands derived from the processing of the 408-A RNA. The faintness of the RNA species labeled d, easily detectable only upon prolonged exposure of the autoradiogram, plausibly reflects its intrinsic instability as an RNA segment. The same cleavage pattern was obtained by challenging the 408-B RNA probe with whole-cell extracts derived from the wild-type *E. coli* W3110 strain (Fig. 6A, lane 5). The 408-B RNA was therefore challenged with whole cellular extracts derived from *E. coli* strains harboring mutant alleles for different ribonucleases.

The *E. coli* W3110 (Fig. 6A, lane 5) and RNase III-negative HT115 (lane 6) strains are isogenic. The strain SK5006 (lane 7), which carries a temperature-sensitive allele of the *rnb* gene, encoding RNase II, was grown at 32°C in order to serve as a control for the isogenic SK5003 strain (lanes 8 and 9), which in addition carries the inactive *pnp*-7 allele of the polynucleotide phosphorylase gene, and the SK5695 strain (lanes 10 and 11),

FIG. 4. Filled and empty YPAL sites. Homologous DNA regions from the *Y. pseudotuberculosis* IP32953 (Ypt) (9) and *Y. pestis* CO92 (Yp) (14) strains are aligned. Numbers refer to genome residues. Capital letters denote TSDs induced by the insertion of YPAL DNA. Y2 and Y8 are members of YPAL subfamilies 2 and 8, respectively.

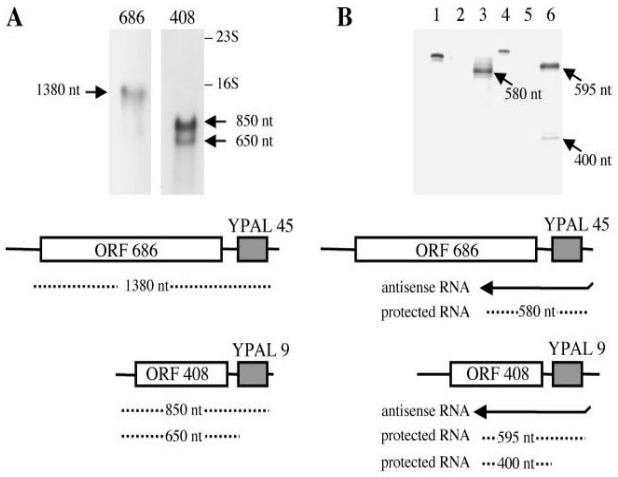


FIG. 5. Analysis of YPAL-positive transcriptional units. (A) Ten micrograms of total RNA from Ye161 cells was separated on a 1% agarose gel, blotted onto nitrocellulose, and probed with amplimers spanning the IMP dehydrogenase/GMP reductase (ORF YE686) and the pyrophosphatase (ORF YE408) genes, respectively. The positions of 16S and 23S rRNA are indicated. Major hybridization bands are indicated by arrows. The mRNA species detected and the corresponding chromosomal regions are drawn at the bottom. (B) RNase protection of YPAL-positive transcripts. In the diagram shown, the RNA probes and protected RNA species are drawn as thick and dotted lines, respectively. T1 RNase-resistant RNA hybrids were electrophoresed on 6% polyacrylamide-8 M urea gels. Protected RNA species are marked by arrows. Lanes: ORF 686 and ORF 408 RNA probes unreacted (lanes 1 and 4) or hybridized to 50 μg of tRNA from *E. coli* (lanes 2 and 5) or to 20 μg of total RNA from the *Y. enterocolitica* Ye161 strain (lanes 3 and 6).

which carries the *me-1* temperature-sensitive allele of the RNase E gene. Significant alteration of the cleavage pattern was observed only with extracts derived from HT115 cells, suggesting that RNase III may play a role in the processing of YPAL-positive mRNAs.

Short YPAL-specific RNA moieties matching in size the b band in Fig. 6A were identified in vivo (Fig. 6B, lane 1) and may originate from cleavage of YPAL-positive transcripts. RNA species detected in vitro and in vivo measure ~200 nt, thus exceeding in length canonical YPAL repeats, such as YPAL 9, analyzed in Fig. 6A, which measure 167 bp. Highly structured RNA also can run abnormally in denaturing gels. To rule out technical artifacts, an RNA species of known length made by the 167 nt of the YPAL 45 element, the 34 nt of the two flanking TSDs, and an additional 30 nt was used as a control in Northern experiments (Fig. 6B, lanes 2 and 3).

Similarly, by using a 130-bp-long YPAL subfamily 8 repeat as a probe in Northern analyses, a ~160-nt RNA species was detected in vivo (data not shown). Discrepancies between the sizes of YPAL elements and the lengths of the hypothetical RNA cleavage products may suggest that YPAL sequences are not targeted per se and that additional base pairing provided by complementary sequences generated at the site of YPAL insertion is required for endonucleolytic cleavage by RNase III to occur. The hypothesis was reinforced by the finding that in vitro, the cleavage products shown in Fig. 6A were no longer detected when a shorter RNA substrate carrying the 167 residues of the YPAL 9 element but lacking flanking TSDs was used as a substrate (data not shown).

Stability of YPAL-positive RNAs in vivo. As shown by both Northern and RNase protection data, YPALs are transcribed into mRNAs. To investigate the role that these sequences may play as RNA elements in vivo, we first identified, by means of PCR analyses, *Yersinia* strains which lack YPAL sequences at specific loci (not shown). Subsequently, the levels of homologous mRNAs were measured in "filled" and "empty" strains by quantitative RT-PCR analyses. Total RNA from *Y. enterocolitica* cells was reverse transcribed into cDNA and amplified under nonsaturating cycling conditions to ensure that the yield of amplified products was proportional to the amount of cellular RNA targeted by PCR. *Y. enterocolitica* 16S rRNA sequences were used as an internal control. The 16S rRNA primers were added to the PCRs two to four cycles before the amplification of the ORF-specific mRNAs was completed (see Materials and Methods). In strain Ye161, ORFs YE686 and YE1047 are flanked 3′ by YPALs 45 and 41, respectively. In contrast, in strain Ye25, the corresponding intergenic sequences both lack YPAL DNA. The levels of transcripts corresponding to either ORF were found to be ~10-fold higher in strain Ye161 than in strain Ye25 (Fig. 7). As for the YE686 transcripts (Fig. 5B), RNase protection experiments indicated that most transcripts spanning the YE1047 ORF retained in their 3′ end sequences corresponding to YPAL 41 (data not shown).

These data support the hypothesis that the presence of YPAL sequences at the 3′ end may increase mRNA stability. The simplest interpretation is that YPAL RNAs may act as stabilizers by forming RNA hairpins able to counteract the progression of the 3′-5′ exonucleases present in the degradosome (8).

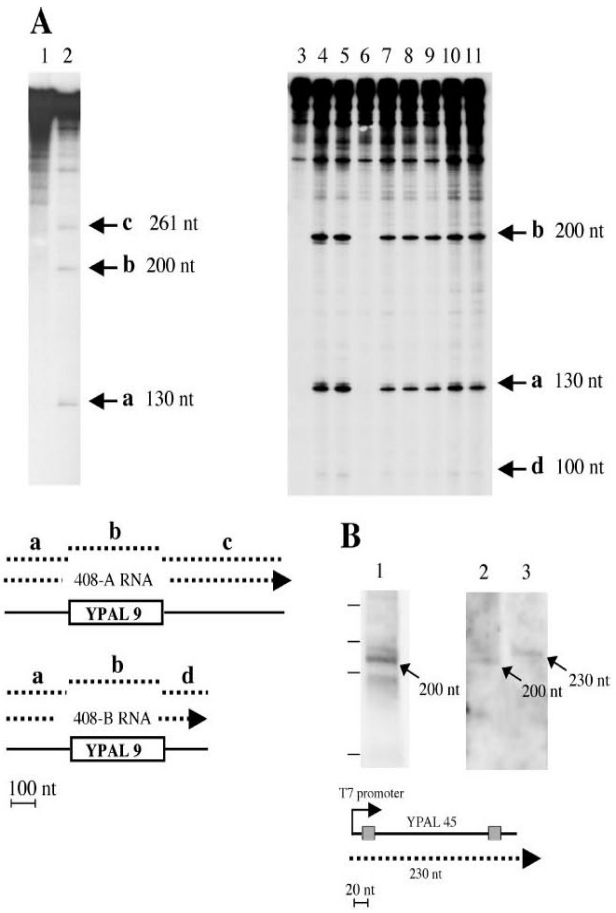


FIG. 6. (A) Processing of YPAL-positive RNA in vitro. The radiolabeled 408-A RNA spanning the element YPAL 9 was incubated for 5 min at 37°C either alone (lane 1) or with 0.5 μg of S100 cellular lysates from the *Y. enterocolitica* strain Ye161 (lane 2). The shorter 408-B RNA was incubated for 5 min at 37°C either alone (lane 3) or with 0.5 μg of S100 cellular lysates from strain Ye161 (lane 4) or *E. coli* strain W3110 (lane 5), HT115 (lane 6), SK5006 (lane 7), SK5003 (lanes 8, 9), or SK5695 (lanes 10, 11). All strains were grown at 32°C; extracts in lanes 9 and 11 were derived from SK5003 and SK5695 cells grown to early logarithmic phase at 32°C and shifted to 44°C for 45 min before harvesting to inactivate RNases II and E, encoded by the *rnb-500* and *rne-1* alleles, respectively. Reaction products were separated on 6% polyacrylamide-8 M urea gels. The RNA substrates and the corresponding processed RNA species a to d are sketched at the bottom. (B) Identification of processed YPAL RNA species in vivo. Total RNA (20 μg) from the Ye161 strain was separated on 6% polyacrylamide-8 M urea gels and transferred to nitrocellulose (lanes 1 and 2). A T7-driven transcript spanning YPAL 45 and flanking TSDs (filled boxes in the diagram at the bottom) and containing in addition 7 nt at the 5′ end and 23 nt at the 3′ end was electrophoresed in lane 3 as a control. The filters were hybridized to a DNA segment spanning the YPAL 9 element. The ~200-nt-long YPAL RNA identified in vivo, the size of the YPAL RNA synthesized in vitro, and the positions of the coelectrophoresed DNA molecular weight markers are shown.

DISCUSSION

We describe in this report the structural and functional characteristics of a relatively abundant set of sequences spread in *Yersinia* genomes. YPAL elements vary in size and sequence composition and can be sorted into several subfamilies. Subfamily 1 includes the largest repeat type, present in one copy in both *Y. enterocolitica* and *Y. pestis*. Members of subfamilies 2 to 7 differ from this prototype sequence in the lack of one or more body segments. Elements belonging to subfamily 8 share the terminal A and H repeats with other YPALs but feature a different body. Taking into account only complete elements, i.e., those including both A and H terminal modules, the YPAL family found in the sequenced 8081 strain of *Y. enterocolitica* is composed of >100 members. The family size is smaller in the sequenced CO92 strain of *Y. pestis*, and this correlates primarily with the reduced size of subfamily 8 in this species. The coexistence of subfamilies is typical of DNA repeats spread in a stepwise manner by bursts of transposition, and YPALs represent indeed a novel class of miniature ISs or MITEs. The mobile nature of YPAL sequences is directly

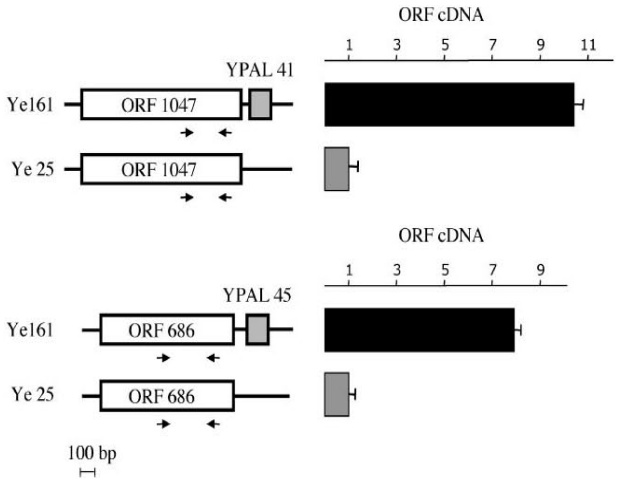


FIG. 7. RT-PCR analyses of YPAL-positive transcripts. Total RNA (200 nanograms) derived from *Y. enterocolitica* strains Ye161 and Ye25 was reverse transcribed, and the cDNA obtained was amplified by PCR with pairs of gene-specific oligonucleotides. Within each pair, one oligonucleotide was ^{32}P end labeled to allow the detection of the amplified segments by autoradiography. Reaction products were run on 6% polyacrylamide-8 M urea gels and quantitated by phosphorimaging. RNA levels were calculated relative to the amount of the coamplified 16S-specific rRNA fragment (see also Materials and Methods) and are expressed in arbitrary units. Black and gray bars denote the amounts of YE686- and YE1047-specific mRNA in the YPAL-positive strain Ye161 and the YPAL-negative strain Ye25, respectively. The ORF YE686 has been described in the legend to Fig. 5. The ORF YE1047 encodes a serine hydroxymethyltransferase. Arrows denote PCR primers.

supported by the finding of homologous chromosomal regions carrying or lacking YPAL DNA in different yersiniae. The identification of empty chromosomal sites in both *Y. pestis* and *Y. pseudotuberculosis* genomes allows us to conclude that transposition of YPAL still occurred after the speciation of *Y. pseudotuberculosis* into *Y. pestis*. As documented by the identification of empty and filled YPAL chromosomal sites among laboratory strains, the mobilization of YPAL sequences is still active in *Y. enterocolitica*.

All MITE families identified so far in eubacteria induce the duplication of the dinucleotide TA upon genomic insertion (15, 22, 24). In contrast, the TSDs which flank YPALs vary in both sequence and size (Fig. 3 and 4). Most YPAL targets overlap or coincide with palindromic sequences. This finding is not novel, since IS*30* (25), IS*1397* (29), IS*903* (18), and IS*621* (10) similarly tend to insert into regions of dyad symmetry. The notion that many YPAL targets may coincide with rho-independent transcriptional terminators also is not unprecedented, since *Y. pestis* IS*1541* (23) and *Mycoplasma fermentans* IS*1630* (6) have also been found inserted at rho-independent transcriptional terminators.

MITEs are mobilized by transposases encoded by ancestral ISs as well as by evolutionarily unrelated elements (5, 16, 20, 24). Distinct ISs may have played a role in the mobilization of YPALs. The conclusion stems primarily from the observation that target sequences were not always found duplicated at YPAL termini. The original targets may have been altered by mutation. However, two IS families are known to insert at transcriptional terminators. IS*1541* does not induce TSDs (23), whereas IS*1630* produces TSDs, ranging from 9 to 22 bp in length (6). It is possible that multiple endonucleases mobilize YPAL insertion and some produce TSDs while others do not.

YPALs are frequently located at the 3′ ends of *Yersinia* genes and often between genes transcribed in a convergent manner. Because some YPAL-positive transcripts accumulate in *Y. enterocolitica* at levels 8 to 10 times higher than those of homologous, YPAL-negative transcripts (Fig. 7), the element may impede the degradosome. In this respect, YPAL may function analogously to the intergenic REP sequences found in enterobacteriaceae (17). YPAL elements might also stimulate degradation of some mRNAs in which they are embedded. Because RNase III cleavage occurs near several YPAL hairpins (Fig. 6), stabilization or degradation of RNA may depend on context (4, 7, 12). Future work is needed to define the role of TSD sequences in RNA cleavage/stabilization.

YPALs are not restricted to *Yersinia*. A dozen 170-bp-long YPAL elements were found by BLAST analyses in the plant pathogen *Erwinia carotovora*. *Yersinia* and *Erwinia* elements are 70% homologous. YPALs may thus be an ancient component of the gamma-proteobacteria that were eliminated from most bacterial species.

Sequence repeats closely resembling YPALs have been identified by in silico surveys for several microorganisms (G. Silvestro and P. P. Di Nocera, unpublished data). Interestingly, many of these sequences may similarly function in posttranscriptional control as *cis*-acting elements according to the pattern of interspersion with coding segments.

ACKNOWLEDGMENTS

We are indebted to Ida Luzzi and Francesca Berlutti for providing us with *Yersinia* strains and Judah Lee Rosner for suggestions and critical reading of the manuscript.

This work was funded by a grant assigned to Pier Paolo Di Nocera by the PRIN 2004 agency of the Italian Ministry of University and Scientific Research.

REFERENCES

1. **Achtman, M., K. Zurth, G. Morelli, G. Torrea, A. Guiyoule, and E. Carniel.** 1999. *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. Proc. Natl. Acad. Sci. USA **96:**14043–14048.
2. **Arraiano, C. M., S. D. Yancey, and S. R. Kushner.** 1988. Stabilization of discrete mRNA breakdown products in *ams pnp rnb* multiple mutants of *Escherichia coli* K-12. J. Bacteriol. **170:**4625–4633.
3. **Bachellier, S., J. M. Clement, and M. Hofnung.** 1999. Short palindromic repetitive DNA elements in enterobacteria, a survey. Res. Microbiol. **150:** 627–639.
4. **Blaszczyk, J., J. Gan, J. E. Tropea, D. L. Court, D. S. Waugh, and X. Ji.** 2004. Noncatalytic assembly of ribonuclease III with double-stranded RNA. Structure **12:**457–466.
5. **Brugger, K., P. Redder, Q. She, F. Confalonieri, Y. Zivanovic, and R. A. Garrett.** 2002. Mobile elements in archaeal genomes. FEMS Microbiol. Lett. **206:**131–141.
6. **Calcutt, M. J., J. L. Lavrrar, and K. S. Wise.** 1999. IS*1630* of *Mycoplasma fermentans*, a novel IS*30*-type insertion element that targets and duplicates inverted repeats of variable length and sequence during insertion. J. Bacteriol. **181:**7597–7607.
7. **Calin-Jageman, I., and A. W. Nicholson.** 2003. RNA structure-dependent uncoupling of substrate recognition and cleavage by Escherichia coli ribonuclease III. Nucleic Acids Res. **31:**2381–2392.
8. **Carpousis, A. J.** 2002. The Escherichia coli RNA degradosome: structure, function and relationship in other ribonucleolytic multienzyme complexes. Biochem. Soc. Trans. **30:**150–155.
9. **Chain, P. S., E. Carniel, F. W. Larimer, J. Lamerdin, P. O. Stoutland, W. M. Regala, A. M. Georgescu, L. M. Vergez, M. L. Land, V. L. Motin, R. R. Brubaker, J. Fowler, J. Hinnebusch, M. Marceau, C. Medigue, M. Simonet, V. Chenal-Francisque, B. Souza, D. Dacheux, J. M. Elliott, A. Derbise, L. J. Hauser, and E. Garcia.** 2004. Insights into the evolution of *Yersinia pestis* through whole-genome comparison with *Yersinia pseudotuberculosis*. Proc. Natl. Acad. Sci. USA **101:**13826–13831.
10. **Choi, S., S. Ohta, and E. Ohtsubo.** 2003. A novel IS element, IS*621*, of the IS*110*/IS*492* family transposes to a specific site in repetitive extragenic palindromic sequences in *Escherichia coli*. J. Bacteriol. **185:**4891–4900.
11. **De Gregorio, E., C. Abrescia, M. S. Carlomagno, and P. P. Di Nocera.** 2002. The abundant class of nemis repeats provides RNA substrates for ribonuclease III in Neisseriae. Biochim. Biophys. Acta **1576:**39–44.
12. **De Gregorio, E., C. Abrescia, M. S. Carlomagno, and P. P. Di Nocera.** 2003. Ribonuclease III-mediated processing of specific *Neisseria meningitidis* mRNAs. Biochem. J. **374:**799–805.
13. **De Gregorio, E., C. Abrescia, M. S. Carlomagno, and P. P. Di Nocera.** 2003. Asymmetrical distribution of *Neisseria* miniature insertion sequence DNA repeats among pathogenic and nonpathogenic *Neisseria* strains. Infect. Immun. **71:**4217–4221.
14. **De Gregorio, E., L. Chiariotti, and P. P. Di Nocera.** 2001. The overlap of Inr and TATA elements sets the use of alternative transcriptional start sites in the mouse galectin-1 gene promoter. Gene **268:**215–223.
15. **De Gregorio, E., G. Silvestro, M. Petrillo, M. S. Carlomagno, and P. P. Di Nocera.** 2005. Enterobacterial repetitive intergenic consensus sequence repeats in yersiniae: genomic organization and functional properties. J. Bacteriol. **187:**7945–7954.
16. **Feschotte, C., N. Jiang, and S. R. Wessler.** 2002. Plant transposable elements: where genetics meets genomics. Nat. Rev. Genet. **3:**329–341.
17. **Higgins, C. F., R. S. McLaren, and S. F. Newbury.** 1988. Repetitive extragenic palindromic sequences, mRNA stability and gene expression, evolution by gene conversion? A review. Gene **72:**3–14.
18. **Hu, W. Y., W. Thompson, C. E. Lawrence, and K. M. Derbyshire.** 2001. Anatomy of a preferred target site for the bacterial insertion sequence IS*903*. J. Mol. Biol. **306:**403–416.
19. **Hulton, C. S. J., C. F. Higgins, and P. M. Sharp.** 1991. ERIC sequences, a novel family of repetitive elements in the genomes of *Escherichia coli*, *Salmonella typhimurium* and other enterobacteria. Mol. Microbiol. **5:**825–834.
20. **Jiang, N., C. Feschotte, X. Zhang, and S. R. Wessler.** 2004. Using rice to understand the origin and amplification of miniature inverted repeat transposable elements MITEs. Curr. Opin. Plant Biol. **7:**115–119.
21. **Mahillon, J., C. Leonard, and M. Chandler.** 1999. IS elements as constituents of bacterial genomes. Res. Microbiol. **150:**675–687.
22. **Mazzone, M., E. De Gregorio, A. Lavitola, C. Pagliarulo, P. Alifano, and P. P. Di Nocera.** 2001. Whole-genome organization and functional properties of miniature DNA insertion sequences conserved in pathogenic neisseriae. Gene **278:**211–222.
23. **Odaert, M., A. Devalckenaere, P. Trieu-Cuot, and M. Simonet.** 1998. Molecular characterization of IS*1541* insertions in the genome of *Yersinia pestis*. J. Bacteriol. **180:**78–181.
24. **Oggioni, M. R., and J. P. Claverys.** 1999. Repeated extragenic sequences in prokaryotic genomes, a proposal for the origin and dynamics of the RUP element in Streptococcus pneumoniae. Microbiology **145:**2647–2653.
25. **Olasz, F., J. Kiss, P. Konig, Z. Buzas, R. Stalder, and W. Arber.** 1998. Target specificity of insertion element IS30. Mol. Microbiol. **28:**691–704.
26. **Parkhill, J., B. W. Wren, N. R. Thomson, R. W. Titball, M. T. Holden, M. B. Prentice, M. Sebaihia, K. D. James, C. Churcher, K. L. Mungall, S. Baker, D. Basham, S. D. Bentley, K. Brooks, A. M. Cerdeno-Tarraga, T. Chillingworth, A. Cronin, R. M. Davies, P. Davis, G. Dougan, T. Feltwell, N. Hamlin, S. Holroyd, K. Jagels, A. V. Karlyshev, S. Leather, S. Moule, P. C. Oyston, M. Quail, K. Rutherford, M. Simmonds, J. Skelton, K. Stevens, S. Whitehead, and B. G. Barrell.** 2001. Genome sequence of Yersinia pestis, the causative agent of plague. Nature **413:**523–527.
27. **Partridge, S. R., and R. M. Hall.** 2003. The IS*1111* family members IS*4321* and IS*5075* have subterminal inverted repeats and target the terminal inverted repeats of Tn*21* family transposons. J. Bacteriol. **185:**6371–6384.
28. **Takiff, H. E., S. M. Chen, and D. L. Court.** 1989. Genetic analysis of the *nuo* operon of *Escherichia coli*. J. Bacteriol. **171:**2581–2590.
29. **Wilde, C., S. Bachellier, M. Hofnung, and J.-M. Clement.** 2001. Transposition of IS*1397* in the family *Enterobacteriaceae* and first characterization of ISKpn*1*, a new insertion sequence associated with *Klebsiella pneumoniae* palindromic units. J Bacteriol. **183:**4395–4404.
30. **Zuker, M.** 1989. On finding all suboptimal foldings of an RNA molecule. Science **244:**48–52.

A novel class of small repetitive DNA sequences in *Enterococcus faecalis*

Rossella Venditti[1], Eliana De Gregorio[1], Giustina Silvestro[1], Tullia Bertocco[1], Maria Francesca Salza[2], Raffaele Zarrilli[2,3] & Pier Paolo Di Nocera[1]

[1]Dipartimento di Biologia e Patologia Cellulare e Molecolare, Facoltà di Medicina, Università Federico II, Napoli, Italy; [2]Dipartimento di Scienze Mediche Preventive, Sezione di Igiene, Università Federico II, Napoli, Italy; and [3]CEINGE Biotecnologie Avanzate, Napoli, Italy

Correspondence: Pier Paolo Di Nocera, Dipartimento di Biologia e Patologia Cellulare e Molecolare, Facoltà di Medicina, Università Federico II, Via S. Pansini 5, 80131 Napoli, Italy. Tel.: +39 81 7462059; fax: +39 81 7703285; e-mail: dinocera@unina.it

Received 26 January 2007; revised 6 March 2007; accepted 6 March 2007.
First published online 10 April 2007.

DOI:10.1111/j.1574-6968.2007.00717.x

Editor: Marco Soria

Keywords
stem-loop structures; palindromic DNA; miniature insertion sequence; genome analysis; microbiological diagnostic.

Abstract

The structural organization of *Enterococcus faecalis* repeats (EFAR) is described, palindromic DNA sequences identified in the genome of the *Enterococcus faecalis* V583 strain by *in silico* analyses. EFAR are a novel type of miniature insertion sequences, which vary in size from 42 to 650 bp. Length heterogeneity results from the variable assembly of 16 different sequence types. Most elements measure 170 bp, and can fold into peculiar L-shaped structures resulting from the folding of two independent stem-loop structures (SLSs). Homologous chromosomal regions lacking or containing EFAR sequences were identified by PCR among 20 *E. faecalis* clinical isolates of different genotypes. Sequencing of a representative set of 'empty' sites revealed that 24–37 bp-long sequences, unrelated to each other but all able to fold into SLSs, functioned as targets for the integration of EFAR. In the process, most of the SLS had been deleted, but part of the targeted stems had been retained at EFAR termini.

Introduction

Enterococci are nonspore-forming Gram-positive microorganisms normally considered commensal of the gastrointestinal tracts of humans and animals, and are commonly found in soil, sewage, water and food, frequently through fecal contamination. In recent years, Enterococci have received growing attention as opportunistic pathogens of clinical significance because they are capable of causing serious diseases (Murray, 2000; Malani *et al.*, 2002). *Enterococcus faecalis* is responsible for severe sepsis and endocarditis, and is an important etiological agent of nosocomial infections. The intrinsic antibiotic resistance, the tolerance of adverse environmental conditions, the promiscuity in acquisition and dissemination of genetically mobile antibiotic resistance elements are all factors that present serious challenges to the treatment of enterococcal infections.

Multilocus sequence typing allowed recently to identify two clonal complexes of *E. faecalis*, CC2 and CC9, responsible for outbreaks and life-threatening infections, mostly in the hospital environment (Ruiz-Garbajosa *et al.*, 2006). In turn, CC2 includes the BVE (Bla^+, Van^r endocarditis; Nallapareddy *et al.*, 2005) complex to which belongs the wholly sequenced V583 strain (Paulsen *et al.*, 2003). BVE clones are rarely found among *E. faecalis* isolates. Progression towards hospital adaptation is likely a multi-step cumulative process where genetic exchange or mutation may lead to epidemic, rather than to clonal population structures (see Feil & Spratt, 2001). Not surprisingly, a prominent feature of the V583 genome is the extraordinary abundance (*c.* 25%) of probable mobile and/or foreign DNA including a plethora of insertion elements (IS), multiple transposons, integrated phage regions and plasmid genes (Paulsen *et al.*, 2003).

Herein is reported the organization of a family of repeated DNA sequences identified in the V583 genome by means of bioinformatic approaches (Petrillo *et al.*, 2006), which was called EFAR (for *Enterococcus faecalis* repeat). These elements partly resemble miniature insertion transposable elements or MITEs, small noncodogenic sequences, which also fold into secondary structures. MITEs feature long terminal inverted repeats (TIRs), and their mobilization is mediated by transposases encoded by ISs featuring similar TIRs (Oggioni & Claverys, 1999; Brugger *et al*, 2002; De Gregorio *et al.*, 2003a, b, 2005). EFAR lack TIRs, and seem to transpose by an unusual cut-and-paste process. It is demonstrated, by means of *in silico* and *in vivo* data, that EFAR have a highly modular structure. Changes in the

organization of the EFAR family among clinical isolates can be easily detectable, making EFAR repeats suitable probes to investigate the epidemiology and population structure of *E. faecalis*.

Materials and methods

Bacterial isolates and growth conditions

Twenty clinical isolates of *E. faecalis* collected from different patients in the Neapolitan area were included in the study (Zarrilli *et al.*, 2005). Relevant background and characteristics of the isolates are detailed in Table 1. All isolates were grown in blood-agar plates at 37 °C and stored at − 70 °C in brain heart infusion (BHI) broth plus 10% glycerol. Bacteria were identified by conventional methods (Gram stain, catalase test) and by biochemical tests using API 20 Strep (bioMérieux, France). All isolates were further identified as *E. faecalis* by amplification and sequence analysis of the 16S rRNA gene performed as previously described (Angeletti *et al.*, 2001).

PCR analyses

Genomic DNA was purified from cultures of *E. faecalis* grown at 37 °C in BHI broth by phenol–chloroform extractions as described (Sambrook *et al.*, 1989). EFAR sequences were amplified by standard protocols using 10 ng of genomic DNA and 160 ng of the *for* (5′- GAGCGTGGGA CAAAAATCAC-3′) and *rev* (5′-GAGGTCGGGACAGAA CCGTT-3′) primers. One oligonucleotide of the pair had been ^{32}P-end-labelled at the 5′ terminus with the polynucleotide kinase. Amplimers were electrophoresed on 6% acrylamide-urea gels and detected by autoradiography. Amplimers labeled as IX–XV in Fig. 2 were gel-purified and reamplified with the *for* and *rev* EFAR primers. Amplimers were then purified from 1.4% agarose gels, and their nucleotide sequence determined by the dye-terminator method. Specific chromosomal segments of the *E. faecalis* genome were amplified by PCR using pairs of oligonucleotides complementary to coding regions flanking EFAR elements in the V583 genome at the concentrations described above. The amplimers were electrophoresed on 1.4% agarose gels along with a commercial DNA ladder (Ladder 100 plus, MBI) as molecular weight marker. Sequences of the PCR primers used are available upon request.

Slot-blot analyses

0.25, 0.5 and 1 μg of DNA from specific *E. faecalis* isolates were loaded onto a Hybond filter and cross-linked by UV treatment as described (Carlomagno *et al.*, 1988). The filter was hybridized to a ^{32}P-radiolabeled PCR product spanning a unit length EFAR repeat. Radioactivity signals were quantitated by phosphorimagery. Signals resulting from hybridization to cold probe DNA loaded on the filter were used to estimate the relative abundance of EFAR DNA in each isolate.

Table 1. *Enterococcus faecalis* isolates by source of isolation, resistance phenotype, PFGE type*

			Resistance phenotype			
Isolate	Clinical source	Month/year of isolation	Bla	VR	HLAR	PFGE type
67	Endocarditis	12/1987	−	−	−	24
72	Bronchial aspirate	12/1987	−	−	−	25
75	Urine	12/1987	−	−	−	51
81	Urine	12/1987	−	−	−	19
93	Urine	12/1988	−	−	−	13
183	Bronchial aspirate	1/1990	−	−	+	20
308	Wound swab	12/1989	−	−	+	11
412	Urine	6/1991	−	−	−	17
413	Urine	6/1991	−	−	−	21
595	Wound swab	11/1992	−	−	−	28
617	Urine	12/1992	−	−	−	34
921	Wound swab	4/1995	−	−	+	10
1070	Endocarditis	2/1997	−	+	+	3
1146	Blood	5/1998	−	−	−	1
1185	Bronchial aspirate	1/1999	−	−	+	9d
1226	Endocarditis	7/1999	−	−	−	47
1319	Endocarditis	11/2000	−	−	+	7
1340	Urine	5/2001	−	−	+	39b
1342	Endocarditis	6/2001	+	−	+	8
1423	Urine	2/2003	−	−	+	23a

*As designated in Zarrilli *et al.* (2005).

Bla, resistant to betalactams; VR, vancomycin resistance; HLAR, high-level aminoglycoside resistance.

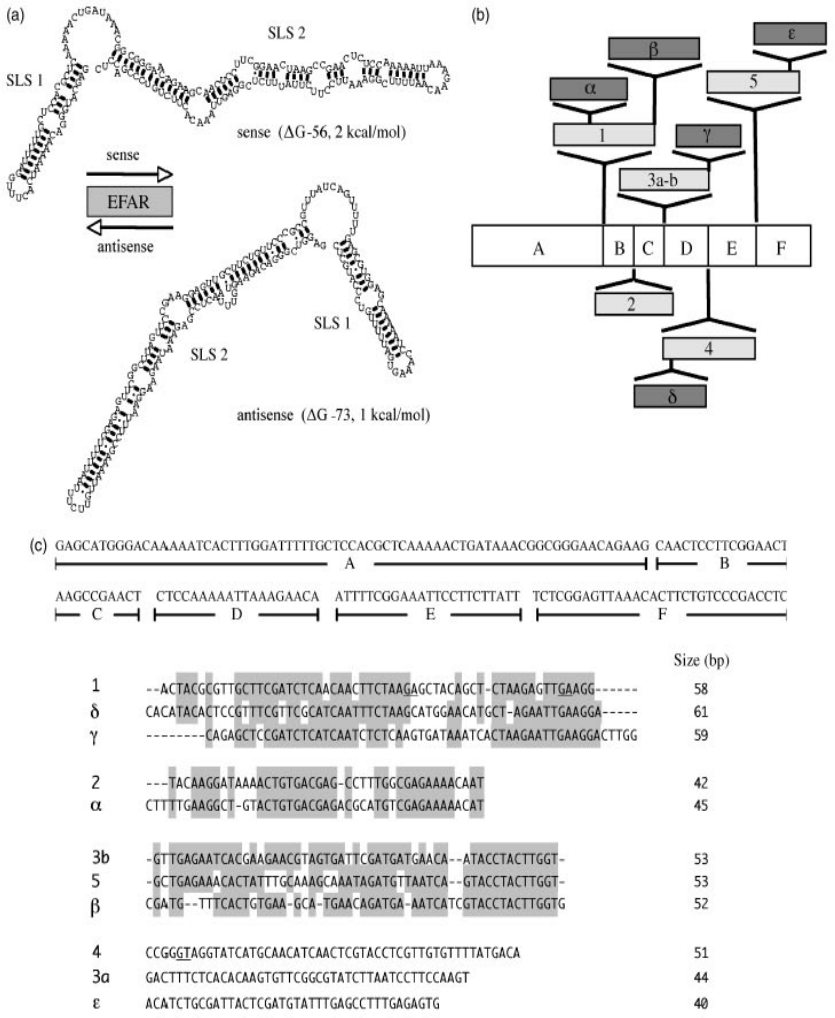


Fig. 1. Structure and organization of EFAR repeats. (a) Secondary structures formed by a consensus 170 bp-long EFAR sequence in the 'sense' and 'antisense' orientation. SLS1 and SLS2 are shown. The Gibbs free energies are indicated. GU pairing is marked by dots. (b) Modular organization of EFAR repeats. The A–F modules, the sites of integration of primary insertions 1–5, and secondary insertions α–ε are shown; 3a and 3b are inserted at the same site. (c) Consensus sequences of EFAR modules and insertions. Sequence relatedness is highlighted. Dashes have been introduced to maximize homologies. Underlined residues mark the sites of integration of secondary insertions. Numbers to the right denote the size in bp of each insertion.

Results

Structure and modular organization of EFAR repeats

The interest was in developing systematic searches for prokaryotic sequences able to fold into stem-loop structures (SLSs; see Petrillo *et al*, 2006). SLSs were searched as stems measuring at least 12 bp, bordering loops 5–100 nt in length. G-U pairing in the stems was allowed. *In silico* analyses of the chromosome of the *E. faecalis* V583 strain (Paulsen *et al.*, 2003) led to the identification of a novel class of repetitive sequences that was called EFAR in this study. The EFAR family includes 55 members, which vary in size from 42 to

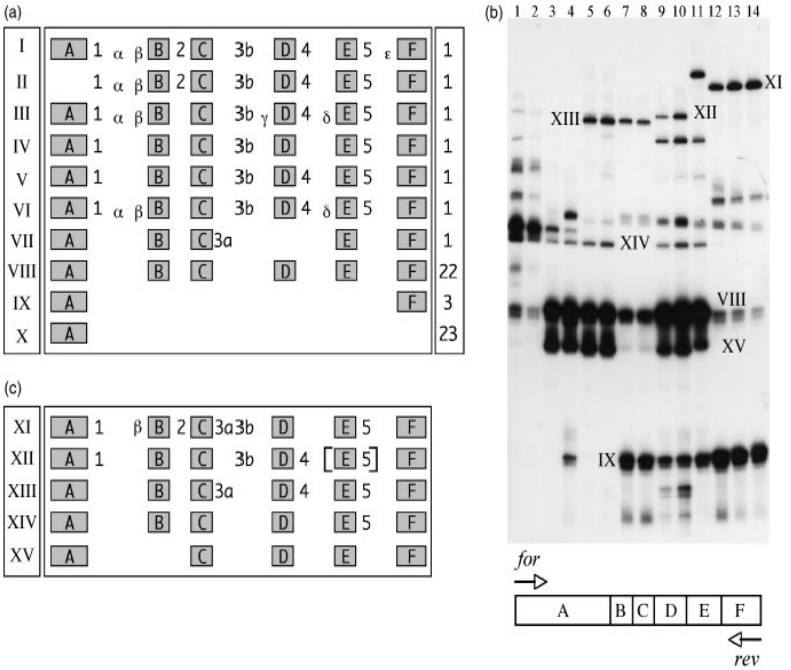


Fig. 2. Genomic organization of EFAR sequences. (a) Modules and insertions defining the 10 (I–X) EFAR subtypes identified in the V583 genome are shown. The number of elements within each subtype is given in the column to the right. (b) DNA derived from 14 isolates of *Enterococcus faecalis* was amplified with the oligomers *for* and *rev* complementary to the EFAR A and F modules. The PCR products were resolved on a 6% polyacrylamide-8 M urea gel, and detected by autoradiography. Amplimers labeled VIII–XV were excised from the gel and reamplified using the same primers. The cold PCR products obtained were purified and subjected to sequence analysis. Lanes: 1 (isol. 75), 2 (isol. 617), 3 (isol. 1146), 4 (isol. 595), 5 (isol. 413), 6 (isol. 1226), 7 (isol. 412), 8 (isol. 93), 9 (isol. 1185), 10 (isol. 921), 11 (isol. 67), 12 (isol. 1379), 13 (isol. 1340), 14 (isol. 1342). (c) Modules and insertions in the EFAR subtypes labeled in (b).

650 bp (see Table 2) and exhibit 90–95% sequence homology. The most abundant repeats measure 170 bp. These unit-length elements can fold into L-shaped structures of relatively low free energy, in which a short SLS1 and a long SLS2 are separated by a 20 nt single stranded region (Fig. 1a). The folding of EFAR is peculiar, since elements of comparable size found in other prokaryotes have been shown to potentially fold into single SLSs (see De Gregorio *et al.*, 2003a, b, 2005, 2006). Thirty-five of the EFAR family members listed in Table 2 are at a distance of 30 bp or less from the stop codon of adjacent ORFs. Thus, it is plausible that most EFAR are cotranscribed along with flanking coding sequences.

EFAR have a peculiar modular composition. The presence at specific sites of five DNA sequences (primary insertions 1–5) brought us to subdivide unit-length repeats into the six A–F modules shown in Fig. 1b. Primary insertions may in turn be interrupted by other DNA (secondary insertions α–ε in Fig. 1b). Interestingly, insertions 1, δ and γ are 68–70% homologous to each other. The same holds true for insertions 2 and α, as for insertions 3b, 5 and β (Fig. 1c). Taking into account the presence/absence of all modules and insertions, 10 different element subtypes (Fig. 2a) could be defined. One third of the repeats located in the V583 strain were found lacking one or more modules. Subtype IX repeats keep the terminal A and F modules, subtype X repeats just sequences spanning the A module. Of these, some are heterogeneous in size and clearly represent deletion derivatives of larger EFAR repeats, others measure 42–44 bp, and span just the segment that encodes SLS1 (see Fig. 1a). This supports the hypothesis that EFAR may have originated, as suggested from secondary structure data, from the fusion of independent DNA sequences.

EFAR families in the *E. faecalis* population

To validate knowledge on the structure of the EFAR family emerging from *in silico* analyses, PCR analyses on 14 clinical

Table 2. EFARs identified in the V583 strain and flanking ORFs*

ORF	EFAR		ORF		ORF		EFAR	ORF	
89	→	1	→	90	1943	→	29	←	1945
167	←	2	←	168	1980	→	30	→	1982
253	→	3	←	255	2052	←	31	←	2055
401	→	4	→	402	2148	←	32	←	2149
672	→	5	←	673	2150	→	33	→	2151
747	→	6	→	748	2374	←	34	←	2376
781	→	7	→	782	2378	←	35	←	2379
809	→	8	→	810	2480	→	36	←	2481
811	→	9	→	812	2495	←	37	→	2496
822	→	10	←	824	2504	←	38	←	2505
897	→	11	←	899	2583	→	39	→	2585
921	←	12	←	922	2595	←	40	←	2597
958	→	13	→	960	2664	←	41	←	2665
1031	→	14	←	1032	2698	→	42	→	2700
1104	←	15	←	1105	2706	→	43	←	2708
1116	→	16	←	1117	2918	←	44	←	2919
1123	→	17	←	1124	3081	←	45	←	3082
1197	→	18	→	1198	3090	→	46	→	3091
1313	→	19	→	1314	3133	←	47	←	3134
1391	→	20	→	1392	3144	→	48	←	3145
1400	→	21	←	1402	3206	→	49	←	3207
1709	→	22	←	1710	3213	→	50	→	3214
1728	→	23	←	1730	3277	←	51	←	3278
1790	→	24	←	1791	3281	←	52	←	3282
1809	←	25	←	1810	3283	←	53	←	3284
1811	←	26	←	1812	3292	→	54	←	3293
1904	→	27	←	1906	3301	←	55	←	3303
1923	→	28	←	1925					

*As designated by Paulsen *et al.* (2003).
Arrows denote ORF orientation.

isolates of *E. faecalis* were performed. Since most EFAR elements carry both A and F modules, oligomers complementary to either module were used as primers to monitor the distribution of EFAR repeats among the different *E. faecalis* isolates by PCR. Using unlabeled primers, in several isolates the 170 bp-long EFAR sequences were the predominant amplimers detected. When PCR experiments were performed with 32-P-end labelled primers, a more complex scenario was obtained. A representative electrophoretic profile obtained by this kind of experiment is shown in Fig. 2b. While major patterns of amplification could be distinguished, most clones exhibited a unique PCR pattern. To validate data, several amplimers were gel-purified and reamplified with the same oligonucleotides. The reaction products were electrophoresed onto 1.4% agarose gels, purified and their sequence determined (Fig. 2). Some amplimers were identical to the EFAR subtypes VIII and IX present also in the V583 genome. In contrast, because of changes in the organization of EFAR modules and insertions, other PCR products resulted to be sequence variants not found in the V583 genome. The degree of variability is illustrated by the comparison of subtypes XII and XIII. While similar in size, the two novel subtypes differ for the presence/absence of insertion 1, and for the alternative presence of insertions 3a and 3b. Moreover, the E module and insertion 5 are partly duplicated in subtype XII (see brackets in Fig. 2c). Several isolates were found to contain unit-length EFARs either selectively decorated by insertion 5 (subtype XIV), or specifically devoided of the B module (subtype XV). Subtype XV resulted in as abundant as subtype VIII in several isolates (lanes 3–6, 9–11). EFARs carrying just the terminal A and F modules (subtype IX repeats) were detected in several isolates, and in some resulted in apparently more abundant than 170 bp-long repeats (see lanes 12–14).

Genomic conservation of EFAR$^+$ loci

Next, the extent of conservation of EFAR$^+$ loci in the population was assessed. To this end, 12/22 chromosomal regions marked in the V583 strain by the presence of 170 bp-long EFAR sequences were monitored by PCR in 20 *E. faecalis* isolates of different genotypes (Table 1). Genomic DNAs were amplified using oligomers complementary to DNA segments flanking EFAR repeats, located 300–700 bp in the V583 genome. For all tested sites, a PCR product was obtained. The size of the PCR products allowed to easily classify regions analyzed as either 'filled' or 'empty' (i.e. containing or lacking EFAR sequences) sites. PCR results are summarized in Fig. 3a. Data revealed a poor conservation of EFAR$^+$ regions on the whole. Only 2/12 elements were retained in all the isolates: EFAR 16, located between the genes encoding the phenylalanyl-tRNA synthetase (ORF 1116) and an ABC transporter permease (ORF 1117), and EFAR 31, located between ftsK (ORF 2052) and the gene encoding a pyridine nucleotide-dissulphide oxidoreductase (ORF 2055).

Conservation at other loci varied from 80% to 70% (see loci defined by EFARs 43, 47 and 22) down to 15–10% (see loci defined by EFAR 10 and 29, respectively). In 11/16 filled regions containing EFAR 43, amplification yielded a PCR product larger than expected. As revealed by sequence analysis of the amplimer derived from the isolate 617, size increase is due to a type 1 insertion. The nature of the sequences inserted into EFAR 52 in the 412 and 595 isolates was not investigated.

In view of the results emerging from PCR surveys, the amount of EFAR DNA in each isolate was determined by slot-blot hybridization (Fig. 3b), and shown to vary from a minimum of *c.* 10 copies, as in isolates 183 and 1070, to a maximum of *c.* 30–35 copies, as in isolates 67 and 413.

No correlation could be drawn between the relative abundance and genomic conservation of EFAR repeats. Thus, for example, isolates 921 and 1185, which resulted to be EFAR$^+$ at all of the loci tested, had less EFAR DNA than

isolates 67, 413 and 1226, which in contrast seemed lacking EFAR sequences at several of the loci analyzed.

Analysis of EFAR empty sites

On the basis of the sizes of the amplification products, it was postulated that EFAR sequences are missing at several expected chromosomal regions. The analysis of the sequence content of 10 different empty sites (marked in Fig. 3a by asterisks) confirmed the lack of EFAR sequences (Fig. 4a). Sequence data unexpectedly revealed the presence, at the site of EFAR insertions, of short SLSs ranging in size from 24 to 37 bp. These alternative sequence elements did not belong to specific DNA families, as they exhibited poor homology to each other. Furthermore, no related sequences were identified in the V583 genome by BLAST searches. The alternative presence of EFAR and shorter SLSs at the same genomic sites may be interpreted in two different ways. According to one view, EFAR sequences may have been excised from the genome, and replaced at each site by a small SLS (Fig. 4b). A different view suggests, in contrast, that each of the small SLSs functioned as entry sites for the genomic integration of an EFAR repeat. In support of the latter hypothesis, the sequence analysis of the same empty sites from different isolates showed the same alternative SLS in all the specific EFAR regions analyzed.

It is worth noting that sequences at the edges of the alternative SLSs coincided with base-paired regions found at filled sites at the termini of EFAR repeats (highlighted residues in Fig. 4a). This finding can be rationalized by hypothesizing that the enzyme(s) mediating the insertion of EFAR might leave behind part of the AT-rich stems of targeted SLSs upon cleavage.

Discussion

EFAR are relatively large palindromic repeats exhibiting a highly modular structure. Unit-length sequences measure 170 bp and can fold into characteristic L-shaped structures where two distinct hairpins, SLS1 and SLS2, are connected by a 20 nt-long single-stranded region. The identification of elements carrying just sequences spanning SLS1 supports the hypothesis that EFAR may result from the combination of independent sequence types having the ability to fold into SLSs.

None of the mobile DNA sequences found in the *E. faecalis* V583 strain (Paulsen *et al.*, 2003) is related, as shown by BLAST homology searches, to EFAR. All the intergenic regions of V583 were compared by the SEQMATCHALL program of the EMBOSS package. Surprisingly, EFAR make up the only family of small (< 300 bp) DNA sequences spread in the genome of enterococci.

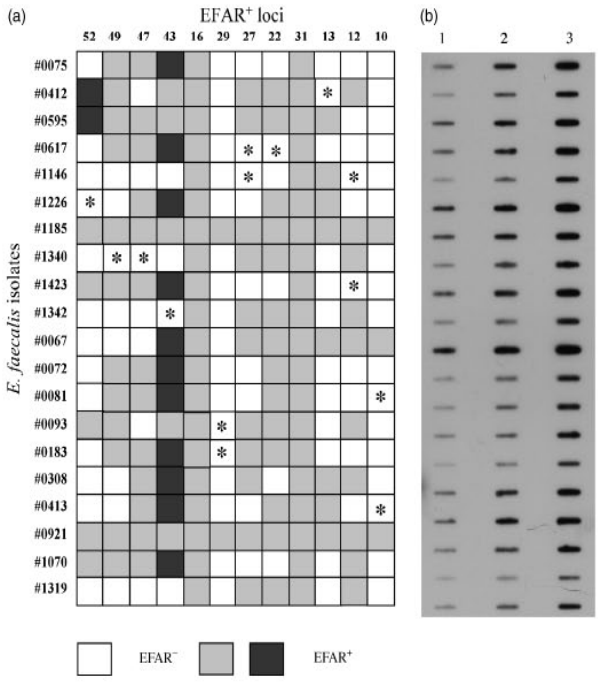


Fig. 3. (a) Analyses of EFAR+ loci. The conservation of EFAR sequences among clinical isolates as assessed by PCR is shown. Numbers to the top refer to chromosomal loci defined by unit-length EFARs identified in the V583 strain (see Table 2), numbers to the left denote *Enterococcus faecalis* isolates. White and grey boxes denote empty and filled sites, respectively. Dark grey boxes mark filled sites with a size larger than expected. Asterisks mark amplimers selected for sequence analysis. (b) The genomic abundance of EFAR sequences in the 20 *E. faecalis* isolates analyzed in (a) was evaluated by slot-blot hybridization. 0.25, 0.5 and 1 μg of DNA from each isolate (lanes 1 to 3, respectively) were loaded onto a Hybond filter, and hybridized to a ^{32}P-radiolabeled EFAR DNA probe (see 'Materials and methods').

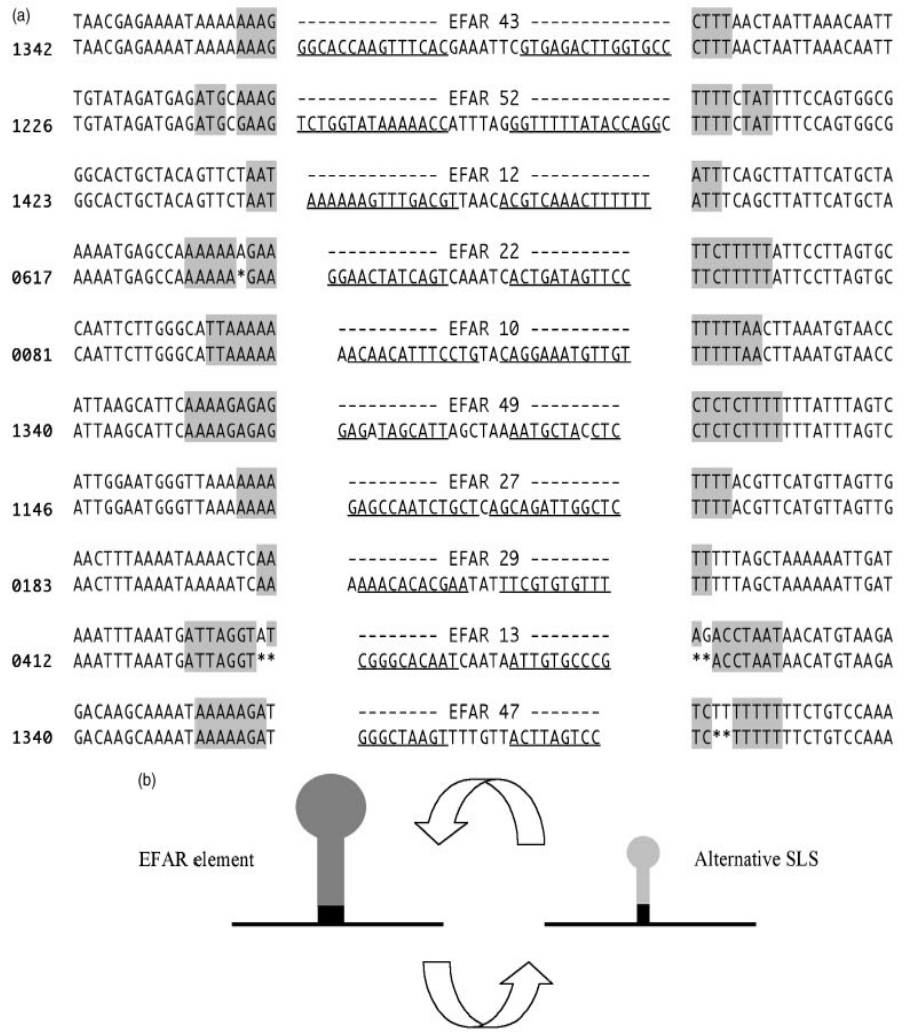


Fig. 4. Analyses of empty EFAR sites. (a) Alignments of homologous DNA regions containing an EFAR repeat in the V583 strain (top sequences) and lacking it in the indicated *Enterococcus faecalis* isolates (bottom sequences). Regions of dyad symmetry are underlined. Asterisks have been introduced to maximize homologies. (b) EFAR and alternative SLSs are sketched not in scale. Black boxes mark regions of base pairing found at the termini of both types of elements, and correspond to the bases highlighted in (a).

EFAR may be interrupted by different types of insertions. All these repeats are inserted in a sequence-specific manner, and most are homologous to each other (Fig. 1c). EFAR insertions seem to have strictly coevolved with EFAR elements, as no homologous sequences were identified outside the mapped EFAR$^+$ loci in the V583 genome. However, it cannot be formally ruled out that insertions rather represent remnants of larger repeats measuring 560–600 bp (see EFAR subfamilies I–II in Fig. 2). Size heterogeneity of EFAR repeats is correlated to the presence/absence of both insertions and modules. Interestingly, most of the clinical isolates analyzed in Fig. 3 exhibited quite distinct different EFAR-PCR patterns. Changes in the distribution of repetitive sequences among bacterial strains are monitored using

PCR primers, which hybridize to the conserved ERIC repeats (Versalovic *et al.*, 1991). A major bias in this type of analyses is data reproducibility, the degeneracy of the primers allowing the detection of amplification patterns also in species lacking ERIC DNA (see Gillings & Holley, 1997). EFAR spread in the genomes of enterococci likely by transposition, and empty and filled homologous chromosomal regions can be distinguished among clinical isolates (Fig. 3). The genomic integration of mobile elements is frequently associated to the generation of target site duplications (TSDs) ranging in size from 2 to 13 bp at the point of insertion. TSDs are not found at the termini of EFAR repeats, and the mechanism of integration of EFAR seems to be indeed rather unusual. The analysis of a representative set of empty sites (Fig. 4a) unequivocally showed that EFAR target sites coincided with 25–40 bp-long DNA regions able to fold into SLSs, which featured AT-rich complementary tracts at their ends. This type of SLS is overrepresented in the genomes of low-GC firmicutes, and may serve multiple functions (Petrillo *et al.*, 2006). Several ISs tend to insert into regions of dyad symmetry (Odaert *et al.*, 1998; Calcutt *et al.*, 1999; Hu *et al.*, 2001; Choi *et al.*, 2003), and rho-independent transcriptional terminator-like sequences are privileged sites of integration for the small (130–170 bp) YPAL repeats in *Yersiniae* (De Gregorio *et al.*, 2006). Interestingly, YPAL induce the duplication of 8–25 bp of target sequences, and this results in the formation of long complementary terminal regions (De Gregorio *et al.*, 2006). In contrast, the integration of EFAR was accompanied by the deletion of most of the target. Yet, EFAR were similarly flanked by base-paired residues provided by the targeted SLS (Fig. 4). Complementary termini are crucial for the recognition of SLSs formed by YPAL RNAs by the RNAseIII (De Gregorio *et al.*, 2006), and may plausibly be important for the mobilization of EFAR.

The isolates of *E. faecalis* analyzed in this work feature distinct PFGE patterns (Table 1) and exhibit differences in the organization, or the interspersion of EFAR sequences (Figs 2 and 3). PCR assays similar to those reported in Fig. 2 revealed that isolates with identical PFGE types showed close or identical EFAR profiles (R. Zarrilli, E. De Gregorio and PP. Di Nocera, in preparation). Thus, changes in the structural organization of the EFAR family may be an additional tool to investigate the epidemiology and population structure of *E. faecalis*. The standardization of PCR and electrophoresis conditions should enable different labs to easily obtain validated EFAR-PCR profiles for genotype analysis.

Acknowledgements

M.S. Carlomagno is thanked for critical reading of the manuscript.

References

Angeletti S, Lorino G, Gherardi G, Battistoni F, De Cesaris M & Dicuonzo G (2001) Routine molecular identification of enterococci by gene-specific PCR and 16S ribosomal DNA sequencing. *J Clin Microbiol* **39**: 794–797.

Brugger K, Redder P, She Q, Confalonieri F, Zivanovic Y & Garrett RA (2002) Mobile elements in archaeal genomes. *FEMS Microbiol Lett* **206**: 131–141.

Calcutt MJ, Lavrrar JL & Wise KS (1999) IS1630 of *Mycoplasma fermentans*, a novel IS30-tYPAL insertion element that targets and duplicates inverted repeats of variable length and sequence during insertion. *J Bacteriol* **181**: 7597–7607.

Carlomagno MS, Chiariotti L, Alifano P, Nappo AG & Bruni CB (1988) Structure and function of the *Salmonella typhimurium* and *Escherichia coli* K-12 histidine operons. *J Mol Biol* **203**: 585–606.

Choi S, Ohta S & Ohtsubo E (2003) A novel IS element, IS621, of the IS110/IS492 family transposes to a specific site in repetitive extragenic palindromic sequences in *Escherichia coli*. *J Bacteriol* **185**: 4891–4900.

De Gregorio E, Abrescia C, Carlomagno MS & Di Nocera PP (2003a) Ribonuclease III-mediated processing of specific *Neisseria meningitidis* mRNAs. *Biochem J* **374**: 799–805.

De Gregorio E, Abrescia C, Carlomagno MS & Di Nocera PP (2003b) Asymmetrical distribution of *Neisseria* miniature insertion sequence DNA repeats among pathogenic and nonpathogenic *Neisseria* strains. *Infect Immun* **71**: 4217–4221.

De Gregorio E, Silvestro G, Petrillo M, Carlomagno MS & Di Nocera PP (2005) Enterobacterial repetitive intergenic consensus sequence repeats in *Yersiniae*: genomic organization and functional properties. *J Bacteriol* **187**: 7945–7954.

De Gregorio E, Silvestro G, Venditti R, Carlomagno MS & Di Nocera PP (2006) Structural organization and functional properties of miniature DNA insertion sequences in *Yersiniae*. *J Bacteriol* **188**: 7876–7884.

Feil EJ & Spratt BG (2001) Recombination and the population structures of bacterial pathogens. *Annu Rev Microbiol* **55**: 561–590.

Gillings M & Holley M (1997) Repetitive element PCR fingerprinting (rep-PCR) using enterobacterial repetitive intergenic consensus (ERIC) primers is not necessarily directed at ERIC elements. *Lett Appl Microbiol* **25**: 17–21.

Hu WY, Thompson W, Lawrence CE & Derbyshire KM (2001) Anatomy of a preferred target site for the bacterial insertion sequence IS903. *J Mol Biol* **306**: 403–416.

Malani PN, Kauffman CA & Zervos MJ (2002) Enterococcal disease, epidemiology, and treatment. *The Enterococci: Pathogenesis, Molecular Biology, and Antibiotic Resistance* (Gilmore MS, Clewell DB, Courvalin PM, Dunny GM, Murray BM & Rice LB, eds), pp. 385–408. ASM Press, Washington, DC.

Murray BE (2000) Vancomycin-resistant enterococcal infections. *N Engl J Med* **342**: 710–721.

Nallapareddy SR, Wenxiang H, Weinstock GM & Murray BE (2005) Molecular characterization of a widespread,

pathogenic, and antibiotic resistance-receptive *Enterococcus faecalis* lineage and dissemination of its putative pathogenicity island. *J Bacteriol* **187**: 5709–5718.

Odaert M, Devalckenaere A, Trieu-Cuot P & Simonet M (1998) Molecular characterization of IS1541 insertions in the genome of *Yersinia pestis*. *J Bacteriol* **180**: 178–181.

Oggioni MR & Claverys JP (1999) Repeated extragenic sequences in prokaryotic genomes, a proposal for the origin and dynamics of the RUP element in *Streptococcus pneumoniae*. *Microbiology* **145**: 2647–2653.

Paulsen IT, Banerjei L, Myers GS *et al.* (2003) Role of mobile DNA in the evolution of vancomycin-resistant *Enterococcus faecalis*. *Science* **299**: 2071–2074.

Petrillo M, Silvestro G, Di Nocera PP, Boccia A & Paolella G (2006) Stem-loop structures in prokaryotic genomes. *BMC Genomics 2006* **7**: 170.

Ruiz-Garbajosa P, Bonten MJ, Robinson DA *et al.* (2006) Multilocus sequence typing scheme for *Enterococcus faecalis* reveals hospital-adapted genetic complexes in a background of high rates of recombination. *J Clin Microbiol* **44**: 2220–2228.

Sambrook J, Fritsch EF & Maniaits T (1989) *Molecular Cloning: A Laboratory Manual*, 2nd edn. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

Versalovic J, Koeuth T & Lupski JR (1991) Distribution of repetitive DNA sequences in eubacteria and application to fingerprinting of bacterial genomes. *Nucleic Acids Res* **19**: 6823–6831.

Zarrilli R, Tripodi MF, Di Popolo A, Fortunato R, Bagattini M, Crispino M, Florio A, Triassi M & Utili R (2005) Molecular epidemiology of high-level aminoglycoside-resistant enterococci isolated from patients in a university hospital in southern Italy. *J Antimicrob Chemother* **56**: 827–835.

Systematic identification of stem-loop containing sequence families in bacterial genomes

Luca Cozzuto[1,2+], Mauro Petrillo[1,3+], Giustina Silvestro[4], Pier Paolo Di Nocera[4], Giovanni Paolella[1,3*]

[1] CEINGE Biotecnologie Avanzate scarl, Via Comunale Margherita 482, 80145 Napoli Italy.

[2] S.E.M.M. - European School of Molecular Medicine - Naples site, Italy

[3] DBBM Dipartimento di Biochimica e Biotecnologie Mediche, Universita' di Napoli FEDERICO II, Via S. Pansini 5, 80131 Napoli, Italy.

[4] DBPCM Dipartimento di Biologia e Patologia Cellulare e Molecolare, Universita' di Napoli FEDERICO II. Via S. Pansini 5, 80131 Napoli, Italy.

* *corresponding author*

+ These authors equally contribute to the work and should be regarded as joint First Authors.

email:
Luca Cozzuto, LC: cozzuto@ceinge.unina.it
Mauro Petrillo, MP: petrillo@ceinge.unina.it
Giustina Silvestro, GS: gsilvest@unina.it
Pier Paolo Di Nocera, PPDN: dinocera@unina.it
Giovanni Paolella, GP: paolella@dbbm.unina.it

Abstract

Background

Analysis of non-coding sequences in several bacterial genomes brought to the identification of families of repeated sequences, able to fold as secondary structures. These sequences have often been claimed to be transcribed and fulfill a functional role. A previous systematic analysis of a representative set of 40 bacterial genomes produced a large collection of sequences, potentially able to fold as stem-loop structures (SLS). Computational analysis of these sequences was carried out by searching for families of repetitive nucleic acid elements sharing a common secondary structure.

Results

The initial clustering procedure identified clusters of similar sequences in 29 genomes, corresponding to about 1% of the whole population. SLSs selected in this way have a substantially higher aptitude to fold into a stable secondary structure than the initial SLS set. Regrouping of the selected sequences by sequence similarity, strand reciprocity and genomic location allowed to remove redundancies. HMM analysis was used to define a final set of 92 families. 25 of them include all well-known SLS containing repeats and some families reported in literature, but not analyzed in detail. The remaining 67 families have not been previously described. Two thirds of the families share a common predicted secondary structure and are located within intergenic regions.

Conclusions

Systematic analysis of 40 bacterial genomes revealed a large number of repeated sequence families, including known and novel ones. Their predicted structure and genomic location suggest that even in compact bacterial genomes, a relatively large fraction of the genome consists of non-protein-coding sequences, possibly functioning at RNA level.

Background

The availability of a massive amount of sequence data stimulated in-depth analyses on the organization of bacterial genomes [1-6]. Although less prominent than in eukaryotic genomes, sequence repeats are found in most bacterial species. According to their sizes, sequence repeats may be roughly classified into two main classes. Large repeats (0.8-2kb) are mostly insertion sequences (IS), and encode proteins mediating their genomic mobility. The terminal inverted repeats (TIRs) and the nature of their gene products allow sorting ISs into specific classes [7,8]. Smaller repeats (50-300 bp) make up a much less defined and more variegate set of genomic sequences. Some of them contain palindromic sequences, demonstrated or proposed to be structured as stem-loops able to function as regulatory elements at DNA or RNA level. For example, *E. coli* PU-BIME elements have been shown to interact with the DNA gyrase [9] and the integration host factor protein [10], but also to function as mRNA stabilizers [11] and transcriptional attenuators [12]. Similarly, palindromic sequence repeats have been shown to function as mRNA stability determinants in *Neisseriae* [13-15] and *Yersiniae* [16,17].

Following these observations, and given the current availability of a large number of sequenced bacterial genomes, a systematic analysis of stem-loop containing repeated sequences appeared of interest. In a previous article [18], high stability stem-loop structures (SLS) were studied within a representative set of bacterial genomes and some of them were shown to have strong similarity with each other. Here we extend this study to detect all families of SLS-containing sequences in the same bacterial set. To this aim, a pipeline, combining sequence clustering and Hidden Markov Model (HMM) based searches, was developed. This strategy led to the definition of a large number of sequence families, sharing sequence similarity and, in most cases, a common predicted secondary structure.

Results

Identification of initial SLS clusters

In a previous work a large number of SLSs were identified within 40 bacterial genomes [18]. For each bacterial species, SLSs obtained from this study and predicted to fold with a free energy lower than –5 Kcal/mol were selected. In order to avoid obvious structured repeated sequences, SLSs were filtered to eliminate those falling within either mature RNA species (tRNAs, rRNAs) or known ISs. An all-against-all BLAST comparison was performed on the selected SLSs for the creation of a distance matrix, where distance is reported as the E-value of the found matches. Since SLSs are strand-specific, BLAST was run without searching for the complementary strand. Links between overlapping SLSs were cut, by eliminating the corresponding matches from the matrix. The resulting matrix was then fed to a Markov Clustering algorithm (MCL) based tool [19] to produce a set of clusters. This clustering step was performed by using stringent parameters (see Materials and Methods) in order to favour the selection of more homogeneous clusters.

To avoid repeated analyses on the same genomic sequence, overlapping clustered SLSs were subsequently joined into larger SLS-containing regions (SCRs). Clusters composed of at least 7 SCRs were selected and are reported in Table 1. Of the 40 analyzed genomes, 29 contain at least one cluster. The procedure led to the identification of 523 clusters, which together contain 28,904 SLSs, corresponding to 12,254 non-overlapping SCRs. No clusters were identified for the remaining 11 genomes: *L. innocua, L. monocytogenes, S. pyogenes, C. pneumoniae, C. trachomatis, U. urealyticum, R. prowazekii, T. pallidum, Buchnera, C. jejuni* and *H. pylori*.

The clusters identified in each positive genome range between 1 and 75, for a total of 8 to over 4,000 clustered SCRs per genome. All together, clustered SLSs correspond to about 1,3% of the originally selected population of over 2 million SLSs. In order to evaluate the quality of the described clustering procedure, grouped SCRs were aligned by using the PCMA multiple alignment tool [20], and the resulting alignments were evaluated by ALISTAT [21]. Over 80% of the clusters showed an average identity better than 60%. The established consensus was larger than 90 bp for the about half of them, while the others produced consensus sequences between 27 and 90 bp (not shown).

Clustered SLSs show high ability to form a stable secondary structure

SLS ability to fold into a reliable secondary structure was analyzed by using RANDFOLD [22], which compares the predicted minimum folding energy (MFE) of a sequence with those of a large number of random shuffles of the same sequence. Results are expressed as a p-value, representing the probability

of the predicted MFE being truly different from the others. In this test, sequences were shuffled by preserving dinucleotide frequencies, as proposed by Workman and Krogh [23].

For each genome, the test was performed on clustered SLSs, as well as on SLSs randomly picked from the initial population and random sequences of equal size extracted from the same genome. The results are reported in Figure 1, where sequences are assigned to specific "folding aptitude" classes, according to the p-value calculated by RANDFOLD. Most clustered sequences (panel A) show a non-random probability of folding below 0.01 (dark grey bars), and, very often, also below 0.001 (black bars), whereas only about 20% of the original SLS population reach these p-values (Figure 1, panel B). Only for *M. leprae*, *L. johnsonii*, *M. genitalium* and *M. pneumoniae*, the two SLS populations do not show statistically different folding aptitudes. A very small fraction (less than 5%) of control sequences showed a non-random folding probability higher than 0.1% (light grey bars in Figure 1, panel C).

Evaluation and refinement of the initial clustering

Various grouping procedures were used to combine the initial 523 clusters, according to sequence similarity, strand reciprocity and position on the genome. The results are reported in Table 2.

In order to group clusters sharing sequence similarity, the clustered SCRs were re-clustered by using the above described BLAST-MCL based procedure, under less stringent conditions. The initial 523 clusters could be associated into 301 new clusters, most of them characterized by a larger number of elements (see column 'sequence' in Table 2). Within each new cluster, overlapping SCRs were fused as described above.

The ability to form SLSs is generally shared by the two complementary strands of a given DNA sequence, the only exception being sequences where GU pairing is essential to form a stem-loop satisfying the minimum requirements. A number of clusters should therefore be composed of elements from the opposite strands of the same genomic region. Such clusters were identified, again by using the BLAST-MCL procedure, this time allowing BLAST searches also on the complementary strand. About two thirds of the clusters could be paired in this way, thus reducing the total number to 205 'unrelated' clusters (see column 'strand' in Table 2).

A third refinement was directed to connect clusters, which might represent different parts of a larger DNA repeat. For this reason, paired clusters, whose elements resulted to be overlapping or located at short distance (< 150 bp), were identified and joined within one group. This led to a further reduction to 137 cluster groups (see column 'location' in Table 2).

The resulting set was pruned by comparing SCRs from each cluster against the IS sequences collected in the ISfinder database [8] by using BLAST, in order to remove insertion sequences, possibly missed in the initial filtering. Similarly rRNA- and tRNA-related clusters were removed by evaluating the

genomic localization of their elements, relative to genes encoding stable RNAs. These tests revealed that 28 cluster groups were related to insertion sequences, mainly not known at the time of the initial filtering, and 11 cluster groups were composed of sequence elements contained within rRNA precursors. These 39 cluster groups, reported in the columns 'IS' and 'rRNA' of Table 2, were flagged and not used for further analysis.

The whole procedure above described led to the identification of 98 candidate SLS-containing repeated DNA families.

Characterization of families expanded by Hidden Markov Model searches

The candidate families were identified starting from small SLS sequences, which may be contained within regions of sequence similarity larger than the originally detected SLSs. In addition, genomic sequences may exist which, although similar, do not contain a SLS able to match the threshold used in the original search. For these reasons, a combined iterative procedure, based on HMM genome searches, was developed and applied to each family. In the procedure, a HMM built on the alignment of all family members is used to scan the parental genome and the detected sequences are aligned to the model. Alignments are extended by attaching neighboring sequences in order to define larger models, when possible. Multiple cycles of alignment, elongation, model building and genome search were performed until the borders of the repeated sequence were reached (see Material and Methods).

A final, manual refinement was performed to combine essentially identical models. At the end of this procedure 92 models were obtained, which define the families reported in Table 3, where the length of the model and the number of detected sequences, both covering the entire model or part of it, are indicated. 67 models range in size between 31 and 200 bp, while the rest are larger, but only two extend over 1 Kb.

BLAST comparison of all family elements, against the consensus sequences for DNA repeats described in literature, revealed that 25 families are already known, corresponding to essentially all previously identified SLS-containing families. For each of them, size and copy number are reported in Table 3, along with the corresponding values derived from literature data.

The remaining 67 families are not described as such in literature. Their sizes range from 31 to over 2,000 bp for a number of elements varying between 9 and 164. Nine of these families (Bhal-2, Clot-2, Clot-3, Myt-5 Sal-2, Myt-11, Nem-4, Pam-1, Hin-1) contain known DNA sequence motifs, such as CRISPR [34], MIRU [35] and DUS [36]: the combination of two or more specific elements, matching these motifs, generates larger, SLS-containing repeated sequences, not previously described. Sixteen families are made up of sequences contained within larger sequence blocks, either coding for abundant protein motifs or located with larger, ill-defined redundant intergenic sequences. Forty-two families

appear to be unrelated to previously described sequence elements.

Secondary structure analyses

Three different approaches were used to evaluate the aptitude of sequences from the detected families to fold into a common secondary structure (results are reported in Table 4):

1) ability to form conserved secondary structures, evaluated, for each family, by RNAz [37] analysis of the alignment of six representative sequences to the family HMM (column "conserved structure");

2) presence of aligned SLSs and agreement with the structure predicted by RNAz (column "conserved SLS position");

3) probability of non-random folding for SLSs contained within each family, calculated by using RANDFOLD [22] (column "SLS folding aptitude").

Only families with either a predicted conserved secondary structure or aligned SLSs are reported. Of the 92 described families, 57 generate a common secondary structure, when analyzed by RNAz. For most (47) of them, marked as "s", the predicted structure contains a stem-loop compatible with the original search. In all but Cod-2, the position of the originally found SLSs is in agreement with the structure predicted by RNAz. These SLSs tend to be positive also to the RANDFOLD test: in 36 of the 47 families, most members contain SLSs, able to fold into a non-random secondary structure ($P <= 0.005$). For ten of the 57 families, indicated by "c", a more complex common structure is predicted by RNAz, not including a stem-loop compatible with the original search. Most of them, accordingly, do not feature aligned SLSs. Yet, the presence of aligned SLSs in three families (Lac-1, REPLEP, BoxC) may be seen as an indication for SLS-containing alternative folding.

RNAz failed to predict a common structure for 35 of the 92 families: for most of these families (29 out of 35) no aligned SLSs are available, indicating the absence of common secondary structures. Aligned SLSs are present in 6 families (Myg-1, Myp-1, Myp-4, Eco-1, Pae-3, RPE-6), which score negative at the RNAz test: for all but RPE-6, aligned SLSs show a low folding aptitude (see Table 4).

Genomic localization

Genomic localization of the families is reported in Table 3 where, in column "type", families are classified, according to the position of the vast majority of their members, relative to annotated coding sequences. 41 families are intergenic (I), 30 genic (G) and 7 tend to span the borders between coding and non coding sequences, and are therefore indicated as border spanning (S).

For 14 families no clear predominance of genic or intergenic sequences was observed, and therefore the family was not assigned to a class.

Genomic localization of the families predicted to fold in a secondary structure is reported in Table 4; for

all families, genomic localization, correlated with the predicted ability of the family members to fold into a common, stable secondary structure, is summarized in Table 5. For most intergenic families a secondary structure is predicted (31 out of 41). Genic families, in contrast, are predominantly not structured: only about one third (9 out of 30) have a structure predicted by RNAz and only for 5 of them aligned SLSs support its existence. Border spanning and unclassified sequence families feature a predicted secondary structure with frequencies similar to intergenic ones.

Characterization of specific families

The described procedure led to the identification of a large number of families of repeated bacterial sequences, some already known, other previously undescribed. For many of them, a number of tests showed the potential folding of the majority of their members into a shared secondary structure. Four such families are reported in Figure 2, where the predicted secondary structure is shown along with the aligned, originally found, SLSs. One of them, the ERIC family from *E. coli*, had previously been described, while the other three are new. ERIC, as anticipated from literature reports [31,32], is predicted to fold into a single, long stem-loop structure. Sta-1 folds into a simple, shorter SLS. Pae-1 and Efa-1 families feature more complex structures, composed of a pair of adjacent SLSs. The structures predicted for these four families may be predicted on both strands, with complementary sequences generally, but not necessarily, folding into corresponding stems. For Pae-1, the prediction of different structures on the two strands indicates the likely presence of multiple foldings of comparable stability, which, on each strand, are alternatively selected as the best one, because of minor base pair differences.

For some of the identified families, secondary structure predictions, although supported by high RNAz scores, are not consistent with the originally found SLSs. Generally this stems from the prediction, by RNAz, of structures not including SLSs fitting with the original SLS definition. PU-BIME and dRS3, shown in Figure 3, are examples of such families: in PU-BIME the stem includes a five base internal loop, while in dRS3 the 8 bp stem is too short. Both cases are not compatible with the original search (see Materials and Methods).

Finally, for about one third of the 92 identified families, it is unlikely that a RNA secondary structure may play a relevant role, as shown by the absence of either a common predicted structure or alignment of originally found SLSs. An example of such families is Myt-10, reported in Figure 3.

Discussion

In a previous study, a systematic analysis of putative SLSs found in bacterial genomes showed that they tend to be more abundant and stable than SLSs randomly formed in shuffled sequences of comparable size and base composition [18]. This observation led to the hypothesis that, along with SLSs stochastically formed because of sequence composition, a sizeable quota is possibly the result of selective pressure, due to the need to preserve a biological function. SLS-containing secondary structures are known to play a relevant role in several aspects of gene expression and its regulation. Structured RNAs are a functional component of enzymes like RNAse P [38], or contribute to the formation of regulatory cis-acting regions such as riboswitches [39], thermosensors [40], transcriptional attenuators and terminators [41,42]. Palindromic RNA sequence repeats may also influence mRNA stability [11].

In this work, we describe a systematic procedure to identify and classify families of repeated sequences, characterized by a shared secondary structure, in the genomes of a representative set of bacteria, most of which of medical interest. To this aim, SLSs were first clustered by sequence similarity and subsequently evaluated for their potential to form secondary structures. In most analyzed genomes, a fraction of SLSs could be grouped into clusters, containing at least 7 non-overlapping SLSs. No clusters were found in 11 of the 40 analyzed genomes.

Clustering by sequence similarity selected 523 clusters corresponding to just above 28,000 SLSs, about 1% of the whole SLS population: this figure may vary quite a lot in specific species, being sensibly larger, up to 6%, in *N. meningitidis*, and substantially lower in *B. subtilis* and *P. multocida*, where less than 0.1% of the SLSs fall within clusters. Clustering ended up by selecting a subset of SLSs different from the original population and characterized by a much higher probability of non-random folding (see Figure 1), indicating that selection based on sequence similarity was very effective in enriching for structured regions.

Various refinement steps produced the final set of 137 clusters, reported in Table 2. Although mature rRNA and tRNA genes were initially masked within the searched genomic sequences, some clusters were identified, which correspond to unmasked parts of ribosomal RNA precursor genes (Table 2). Similarly, some clusters correspond to SLSs contained within ISs, which escaped the initial filtering for various reasons. Removal of these two subsets and other redundancies reduced the number of identified families to 92.

Notwithstanding the starting population of SLS-containing sequences, within these families regions sharing primary structure similarity, but not a common SLS, might, in principle, still be found, and 35

families with no recognizable shared secondary structure, were indeed identified. Most of these sequences are, not surprisingly, found within coding regions, where the formation of secondary structures is expected to be limited by the translation machinery. However, some of these families coincide with intergenic sequence repeats, such as the *S. pneumoniae* BOX and *P. putida* REP sequences unable to form structures compatible with the originally searched ones.

Families sharing common secondary structures

Most identified families, 57 out of 92, are predicted by RNAz to share a common secondary structure. This group includes well-known intergenic families, such as the *E. coli* PU-BIME and ERIC repeats, and their homologues in other species, as well as a number of less known families, most of which described in isolated reports, but not characterized in detail (see Table 3). Practically all intergenic repeats, previously shown or predicted to fold into a RNA secondary structure, have been found. The only exceptions are the *S. pneumoniae* RUP and the *R. conorii* RPE-6 repeats, which, although identified by the pipeline, do not fall into this group, because RNAz could not predict a shared secondary structure better than the defined threshold.

For known families, the sequence boundaries, as predicted by the pipeline, are essentially coincident with those previously reported in literature. Specific discrepancies were found only in two families. In the *N. meningitidis* NEMIS elements, the present search identified the central 46 bp core, but failed to extend the similarity to either the partial 108 or the complete 158 bp repeats described by Mazzone et al. [13]. Similarly, for the *S. pneumoniae* RUP family, only 63 bases were detected out of the complete 108 bp elements [26].

Known and novel families

In well characterized genomes, such as those of enterobacteria, practically all known families have been detected, along with a few new ones. In *E. coli,* the known PU-BIME, ERIC and BoxC families were recognized and feature shared secondary structures, while the only new one identified, the Eco-1 family, is predicted as unable to fold. PU-BIME repeats were also detected in *S. typhi* as two related variants (a full size and a shorter one, only the former predicted to fold) and in *S. typhimurium*, along with two novel families, Sal-1 and Sal-2 (Table 3). For both of them RNAz could predict a shared secondary structure of the complex type.

As expected, ERIC sequences were detected not only in *E. coli,* but also in *Y. pestis* and *V. cholerae* [16,31]: *Y. pestis* repeats are predicted to fold with a structure closely similar to the *E. coli* elements. In contrast, ERIC sequences detected in *V. cholerae* are not predicted to fold, being 20 bp shorter than both *E. coli* and *Y. pestis* homologues, because of selective erosion of their TIRs. *Yersiniae* ERIC sequences

have been shown to regulate the level of expression of neighboring genes by folding into RNA harpins [16]. *V. cholerae* ERIC, being unable to fold, may thus not function as RNA stability determinants.

Most potentially structured new families have been found in species less analyzed experimentally or whose genome was more recently sequenced, such as pseudomonaceae, bordetellae, mycobacteria.

For both novel and known families, the predicted common secondary structure is often a stem-loop (see Sta-1 and ERIC in Figure 2). In a fraction of cases, however, RNAz analysis proposes different structures. Some families feature a double hairpin (see EFA-1 and Pae-1 in Figure 2) and others feature a complex structure containing a SLS (not shown).

Genomic localization

Genomic localization highlights the preferential tendency of repeated sequences with a predicted common secondary structure to lye within intergenic regions; this is true for both known and novel ones. In contrast, families found within coding sequences (CDSs) of genomes are often not structured. This is in agreement with the results of RANDFOLD analysis: most (19 out of 27) intergenic families with aligned SLSs (Table 4) are enriched in highly structured SLSs, while this is true for only one genic family, Myp-2. These observations support the overall hypothesis that many of these sequence families fold in a secondary structure at the RNA level, particularly those located in intergenic regions, where the translation machinery is not expected to interfere with secondary structure formation.

Three novel intergenic structured families, Hin-1 in *H. influenzae*, Nem-4 in *N. meningitidis* and Pam-1 in *P. multocida* are composed of similar sequences, characterized by the repetition of short, abundant oligonucleotides, known as DUS [36]. The recurrence, at specific short distances, of this basic oligonucleotide module, shorter than the searched pattern, produces a conserved SLS larger than the required threshold. It is possible that these sequences function as transcriptional terminators, and it has been recently reported that terminator hairpins are indeed frequently formed by closely spaced, complementary instances of exogenous DNA uptake signal sequences [43].

Some novel structured families are located within CDSs. They often contain repetitive motifs of one or a few coding regions, such as Lac-1 in *L.johnsonii*, Pae-3 in *P. aeruginosa* and Efa-2 in *E. faecalis*. Interestingly, the Cod-2 family defines a very small repeat, found within various CDSs, encoding different peptides in different frames. Cod-2 repeats resemble repetitive sequence elements found by Claverie and coworkers in protein coding genes of *R. conorii* [44]. Five genic families found in *M. pneumoniae* are part of large (1.5-5.4 kb), possibly mobile repeated DNA sequences having coding capacity [45].

About one third of the identified families are found to be "unstructured". These sequences were not the object of the original search; a possible explanation of their detection is the incidental presence of SLSs

within large repeated sequences. Most such families fall within CDSs (see Table 4, and Myt-10 in Figure 3 as an example). Ten of them are contributed by only two genomes: *M. tuberculosis* and *M. pneumoniae*. Other unstructured families are clustered within the same CDS (Bor-3 and Bor-6 in *B. bronchiseptica*) or are dispersed within multiple CDSs, sharing a common protein domain (Bor-4 and Bor-5 in *B. bronchiseptica,* Pae-2 and Ppu-3 in *P. aeruginosa* and *P. putida*, respectively).

Conclusions

A systematic analysis of 40 bacterial genomes is presented, aimed to identify repeated sequence families, sharing a common secondary structure. This procedure identified practically all already described families meeting these constraints, as well as a larger number of novel, undescribed nucleic acid repeats.

About two thirds of the families shared a conserved secondary structure, often a stem-loop based one. Interestingly, these families are mostly composed by elements located within intergenic regions. This localization reflects the hypothesis that RNA folding, within these regions, is more likely to occur, not being affected by the translation machinery.

The identification of repetitive sequence families, able to fold into secondary structures and preferentially located within intergenic regions, reinforces the notion that also in prokaryotic genomes, typically more compact than eukaryotic ones, a relatively large fraction, not coding for proteins, is likely to play a biological role, by encoding functional RNAs.

Materials and Methods

Selection of SLS clusters

SLSs previously identified in 40 bacterial genomes by Petrillo at al. [18] were taken as the starting population. Only SLSs predicted to fold with a free energy <=–5 Kcal/mol were used for the present study.

For each genome, selected SLSs were clustered according to a procedure, based on BLAST and MCL programs [46,19]. An all-against-all BLAST comparison was performed on the SLS population, to create an E-value based distance matrix. The BLAST result matrix was pruned by removing hits linking overlapping SLSs, and subsequently fed to MCL to produce a set of clusters. BLAST was performed with an E-value cut-off of 1E-4 and only on the sequence top strand. MCL was run by setting the inflation parameter (I) equal to 4. The alignments of clustered elements were produced by PCMA [20] used with default parameters.

Aptitude to form a stable secondary structure

The aptitude of SLSs and control sequences to form a stable secondary structure was tested by running RANDFOLD [22]. The '–d' option was used, in order to preserve dinucleotide frequencies. RANDFOLD was set to shuffle each sequence 1,000 times. In the tests reported in Figure 1, all clustered SLSs (panel A) were compared to a number of SLSs representing the 5% of initial SLS population (panel B) and to a number of genomic sequences having the same size of clustered SLSs, randomly extracted from the corresponding genomes (panel C). Control sequences analyzed in panels B and C, were selected three times, in order to evaluate average and standard deviations.

Cluster refinement

The regrouping procedures summarized in Table 2 were made as follows:

- **Regrouping by sequence** was made by using the BLAST-MCL procedure (see above) on all SCRs, but in a less stringent way, i.e. setting parameter 'I' to 1.4.
- **Regrouping by strand** was performed by using the BLAST-MCL procedure, but allowing searches on the complementary strand and setting parameter 'I' to 1.4.
- **Regrouping by location** was obtained by merging clusters in which SCRs were partially overlapping, or within a distance of 150 bp, according to their genomic coordinates.

For each regrouping procedure, groups of clusters, sharing at least 50% of the elements, were fused into a larger one.

Identification of families by cycles of HMM searches

In order to identify all family members of each cluster, a procedure was developed, based on cycles of alignment by PCMA and search on the genome by HMMER package tools [21]. First, SCRs of clusters regrouped by sequence (see Table 2) were aligned by PCMA with option 'ave_grp_id' set to 50. The procedure can be summarized as it follows:

1. The alignment is used to build a HMM by HMMBUILD and HMMCALIBRATE, with the default options.
2. The produced HMM is used to search new elements within the genome, by using HMMSEARCH. E-value cut-off was set to 1E-10. Independent searches are run on each genomic sequence strand.
3. Identified sequences are extracted and aligned to their parental HMM by HMMALIGN. Pairs of overlapping sequences on the opposite strands are avoided by discarding the one with the worse score and E-value.
4. The aligned sequences are extended by 10% of the length of the parental HMM. Only the extensions are aligned by PCMA.
5. The alignment of the extended sequences is then used for the construction of a new model, returning to step 1.

The loop ends when one of the following criteria is met:

- The detected sequences, which cover the entire model, are less than 7;
- The new model is shorter in terms of length than the previous one.
- The alignment does not extend the HMM any further (within a tolerance of 3 bp).
- The alignment contains a number of gaps higher than 30% of the aligned bases.
- The extreme value distribution, derived from the model calibration, is in the range Average_Score ± 3*Standard_Deviation, derived from HMMBUILD.

The HMM and the final alignment are used as definition of the family.

Secondary structure analyses

SLSs contained in sequences of each family were analyzed by RANDFOLD as described above and taken as positive if their p-value is < 0.005. Families were divided in four categories, according to the fraction of sequences containing at least one positive SLS ('+++' if 90% or above; '++' if 70-90%; '+' if 50-70%; '-' if less than 50%).

Representative sequences of the families shown in Figures 2 and 3 were chosen in the following way:

- All sequences able to cover the entire model are sorted by the E-value determined by HMMSEARCH.
- Six sequences are picked from this population by selecting the best model-fitting one and five (if available) more with progressively increasing of the E-value.

Sequences were aligned to corresponding HMM by using HMMALIGN and the resulting alignments were analyzed by RNAz (version 0.1.1) [37]. For RNAz analysis, alignments with length <=200 bp were used as a single block, while alignments with length >200 bp were screened in sliding windows (length 120 and slide 40), according to Washietl et al. [47].

RNAz was used with standard parameters. All alignments with RNAz classification score $P > 0.5$ were considered. Overlapping hits, i.e. resulting from hits in overlapping windows, were analyzed again by using larger sliding windows able to contain structures obtained with different hits.

List of abbreviations used

bp, base pair

CDS, coding sequence

CRISPR, clustered regularly interspaced short palindromic repeats

DUS, DNA uptake sequence

HMM, Hidden Markov Model

IS, insertion sequence

MCL, Markov Clustering algorithm

MFE, minimum folding energy

MIRU, mycobacterial interspersed repeated unit

nt, nucleotide

SCR, SLS-containing region

SLS, stem-loop-structure

TIR, terminal inverted repeat

Authors' contributions

LC and MP designed the procedure for clustering of SLSs into families. They also analyzed sequence family members, respect to folding aptitude and secondary structure prediction. GS retrieved the literature information and provided manual annotation and analysis. PPDN and GP conceived and coordinated the study. All authors read and approved the final manuscript.

Acknowledgments

We wish to thank Angelo Boccia for suggestions and useful discussions, Tommaso Russo for critically reading the manuscript. Informatic support by Gianluca Busiello is also acknowledged. This work has been supported by Ministero dell'Istruzione dell'Universita' e della Ricerca (MIUR) under the PON2004 (SCoPE), FIRB (LITBIO), PRIN 2005 and BioinfoGRID European Projects.

Conflicts of interest. The authors have declared that no conflicts of interest exist.

References

1. van Belkum A, van Leeuwen W, Scherer S, Verbrugh H: **Occurrence and structure-function relationship of pentameric short sequence repeats in microbial genomes.** *Res Microbiol* 1999, **150**:617-626.

2. Audit B, Ouzounis CA: **From genes to genomes, universal scale-invariant properties of microbial chromosome organization.** *J Mol Biol* 2003, **332**:617-633.

3. Rocha EP, Danchin A: **Gene essentiality determines chromosome organisation in bacteria.** *Nucleic Acids Res* 2003, **31**:6570-6577.

4. Ussery DW, Binnewies TT, Gouveia-Oliveira R, Jarmer H, Hallin PF: **Genome update: DNA repeats in bacterial genomes.** *Microbiology* 2004, **150**:3519-3521.

5. Rocha EP: **Inference and analysis of the relative stability of bacterial chromosomes**. *Mol Biol Evol* 2006, **23**:513-522.

6. Field D, Wilson G, van der Gast C: **How do we compare hundreds of bacterial genomes?** *Curr Opin Microbiol* 2006, **9**:499-504.

7. Leplae R, Hebrant A, Wodak SJ, Toussaint A: **ACLAME: a CLAssification of Mobile genetic Elements.** *Nucleic Acids Res* 2004, **32**:D45-49.

8. Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M: **ISfinder: the reference centre for bacterial insertion sequences.** *Nucleic Acids Res* 2006, **34**:D32-36.

9. Yang Y, Ames GF: **DNA gyrase binds to the family of prokaryotic repetitive extragenic palindromic sequences.** *Proc Natl Acad Sci U S A* 1988, **85**:8850-8854.

10. Boccard F, Prentki P: **Specific interaction of IHF with RIBs, a class of bacterial repetitive DNA elements located at the 3' end of transcription units.** *EMBO J* 1993, **12**:5019-5027.

11. Higgins CF, McLaren RS, Newbury SF: **Repetitive extragenic palindromic sequences, mRNA stability and gene expression: evolution by gene conversion? A review.** *Gene* 1988, **72**:3-14.

12. Espeli O, Moulin L, Boccard F: **Transcription attenuation associated with bacterial repetitive extragenic BIME elements.** *J Mol Biol* 2001, **314**:375-386.

13. Mazzone M, De Gregorio E, Lavitola A, Pagliarulo C, Alifano P, Di Nocera PP: **Whole-genome organization and functional properties of miniature DNA insertion sequences conserved in pathogenic Neisseriae.** *Gene* 2001, **278**:211-222.

14. De Gregorio E, Abrescia C, Carlomagno MS, Di Nocera PP: **Ribonuclease III-mediated processing of specific Neisseria meningitidis mRNAs.** *Biochem J* 2003, **374**:799-805.

15. Rouquette-Loughlin CE, Balthazar JT, Hill SA, Shafer WM: **Modulation of the mtrCDE-encoded efflux pump gene complex of Neisseria meningitidis due to a Correia element insertion sequence.** *Mol Microbiol* 2004, **54**:731-741.

16. De Gregorio E, Silvestro G, Petrillo M, Carlomagno MS, Di Nocera PP: **Enterobacterial repetitive intergenic consensus sequence repeats in Yersiniae: genomic organization and functional properties.** *J Bacteriol* 2005, **187**:7945-7954.

17. De Gregorio E, Silvestro G, Venditti R, Carlomagno MS, Di Nocera PP: **Structural organization and functional properties of miniature DNA insertion sequences in Yersiniae.** *J Bacteriol* 2006, **188**:7876-7884.

18. Petrillo M, Silvestro G, Di Nocera PP, Boccia A, Paolella G: **Stem-loop structures in prokaryotic genomes.** *BMC Genomics* 2006, **7**:170.

19. Enright AJ, Van Dongen S, Ouzounis CA: **An efficient algorithm for large-scale detection of protein families.** *Nucleic Acids Res* 2002, **30**:1575-1584.

20. Pei J, Sadreyev R, Grishin NV: **PCMA: fast and accurate multiple sequence alignment based on profile consistency.** *Bioinformatics* 2003, **19**:427-428.

21. Bateman A, Birney E, Durbin R, Eddy SR, Finn RD, Sonnhammer EL: **Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins.** *Nucleic Acids Res* 1999, **27**:260-262.

22. Bonnet E, Wuyts J, Rouze P, Van de Peer Y: **Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences.** *Bioinformatics* 2004, **20**:2911-2917.

23. Workman C, Krogh A: **No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution.** *Nucleic Acids Res* 1999, **27**: 4816–4822.

24. Okstad OA, Tourasse NJ, Stabell FB, Sundfaer CK, Egge-Jacobsen W, Risoen PA, Read TD, Kolsto AB: **The bcr1 DNA repeat element is specific to the Bacillus cereus group and exhibits mobile element characteristics.** *J Bacteriol* 2004, **186**:7714-7725.

25. Martin B, Humbert O, Camara M, Guenzi E, Walker J, Mitchell T, Andrew P, Prudhomme M, Alloing G, Hakenbeck R *et al*: **A highly conserved repeated DNA element located in the chromosome of Streptococcus pneumoniae.** *Nucleic Acids Res* 1992, **20**:3479-3483.

26. Oggioni MR, Claverys JP: **Repeated extragenic sequences in prokaryotic genomes: a proposal for the origin and dynamics of the RUP element in Streptococcus pneumoniae.** *Microbiology* 1999, **145**:2647-2653.

27. Halling SM, Bricker BJ: **Characterization and occurrence of two repeated palindromic DNA elements of Brucella spp.: Bru-RS1 and Bru-RS2.** *Mol Microbiol* 1994, **14**:681-689.

28. **RicBase Rickettsia genome database.** [http://igs-server.cnrs-mrs.fr/mgdb/Rickettsia]

29. Cole ST, Supply P, Honore N: **Repetitive sequences in Mycobacterium leprae and their impact on genome plasticity.** *Lepr Rev* 2001, **72**:449-461.

30. Parkhill J, Achtman M, James KD, Bentley SD, Churcher C, Klee SR, Morelli G, Basham D, Brown D, Chillingworth T *et al*: **Complete DNA sequence of a serogroup A strain of Neisseria meningitidis Z2491.** *Nature* 2000, **404**:502-506.

31. Bachellier S, Clement JM, Hofnung M: **Short palindromic repetitive DNA elements in enterobacteria: a survey.** *Res Microbiol* 1999, **150**:627-639.

32. Sharples GJ, Lloyd RG: **A novel repeated sequence located in the intergenic regions of bacterial chromosomes.** *Nucleic Acids Res* 1990, **18**:6503–6508.

33. Aranda-Olmedo I, Tobes R, Manzanera M, Ramos JL, Marques S: **Species-specific repetitive extragenic palindromic (REP) sequences in Pseudomonas putida.** *Nucleic Acids Res* 2002, **30**:1826-1833.

34. Godde JS, Bickerton A: **The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes.** *J Mol Evol* 2006, **62**:718-729.

35. Supply P, Mazars E, Lesjean S, Vincent V, Gicquel B, Locht C: **Variable human minisatellite-like regions in the Mycobacterium tuberculosis genome.** *Mol Microbiol* 2000, **36**:762-771.

36. Davidsen T, Rodland EA, Lagesen K, Seeberg E, Rognes T, Tonjum T: **Biased distribution of DNA uptake sequences towards genome maintenance genes.** *Nucleic Acids Res* 2004, **32**:1050-1058.

37. Washietl S, Hofacker IL, Stadler PF: **Fast and reliable prediction of noncoding RNAs.** *Proc Natl Acad Sci U. S. A.* 2005, **102**:2454-2459.

38. Kazantsev AV, Pace NR: **Bacterial RNase P: a new view of an ancient enzyme.** *Nat Rev Microbiol* 2006, **4**:729-740.

39. Nudler E, Mironov AS: **The riboswitch control of bacterial metabolism.** *Trends Biochem Sci* 2004, **29**:11-17.

40. Johansson J, Mandin P, Renzoni A, Chiaruttini C, Springer M, Cossart P: **An RNA thermosensor controls expression of virulence genes in Listeria monocytogenes.** *Cell* 2002, **110**:551-561.

41. Merino E, Yanofsky C: **Transcription attenuation: a highly conserved regulatory strategy used by bacteria.** *Trends Genet* 2005, **21**:260-264.

42. Ermolaeva MD, Khalak HG, White O, Smith HO, Salzberg SL:**Prediction of transcription terminators in bacterial genomes.** *J Mol Biol* 2000, **301**:27-33.

43. Kingsford CL, Ayanbule K, Salzberg SL: **Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake.** *Genome Biol* 2007, **8**:R22.

44. Claverie JM, Ogata H: **The insertion of palindromic repeats in the evolution of proteins.** *Trends Biochem Sci* 2003 **28**:75-80.

45. Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li BC, Herrmann R: **Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae.** *Nucleic Acids Res* 1996, **24**:4420-4449.

46. Altschul, SF, Gish, W, Miller, W, Myers, EW, Lipman, DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403-410.

47. Washietl S, Hofacker IL, Lukasser M, Huttenhofer A, Stadler PF: **Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome.** *Nat Biotechnol* 2005, **23**:1383-1390.

Figure Legends

Figure 1. Fraction of sequence elements positive to RANDFOLD test. RANDFOLD test was run onto groups of clustered SLSs (panel A), total SLSs (panel B) and random sequences (panel C) from the 29 genomes listed in Table 1. The fraction of elements scoring positive with the indicated probability is diagrammed. Standard deviation bars are shown in panels B and C.

Figure 2. Alignment of ERIC, Pae-1, Sta-1 and Efa-1 family members. (A) A representative set elements from each family was aligned by using the HMM model as a guide. In each panel, one row corresponds to one family member (indicated on the right with its genomic position). Within each row, sequence conservation is indicated by increasing gray levels and gaps by dotted spaces; overlapping SLSs are reported as red and blue lines, the red ones indicating SLSs used to define the original HMM model for the family, the blue all the others. Darker colors indicate the SLS folding aptitude, i.e. positivity to RANDFOLD for P<=0.005. Common secondary structures, predicted by RNAz, are reported at the bottom, just above the ruler in nucleotides: green triangles indicate stems produced by pairing complementary regions on the same strand as the identified SLSs, while brown triangles indicate the same from the opposite strand. The boxed regions highlight areas where aligned SLSs and predicted structures are in agreement. **(B)** Graphic representation of the RNAz predicted secondary structures.

Figure 3. Alignment of PU-BIME, dRS3 and Myt-10 family members. Panels **A** and **B** legends are as in Figure 2**.**

Table 1. Sequence-based clustering of SLSs. BLAST-MCL based clustering of SLSs from bacterial genomes described in Petrillo et al. [18]: only species featuring at least one cluster, with a minimum of 7 elements, are listed. For each species, the number of SLSs within the starting populations, the number of clusters and the number of clustered SLSs are reported. The number of SLS-containing regions (SCRs), obtained by fusing overlapping clustered SLSs, is also reported.

Table 2. Regrouping of SLS clusters. Clusters reported in Table 1 were tested for sequence similarity, strand reciprocity and relative genomic position of their elements, and grouped accordingly. The number of clustered groups is reported in columns marked "Grouped by". The number of groups, whose elements are part of ISs or rRNA genes, is shown in the last two columns.

Table 3. Families of SLS-containing repeated sequences. The final set of 92 families of repeated sequences is reported, grouped by species. For each family, the length of the model and the number of sequences fitting the model are given. The number of complete sequences, i.e. covering the model from end to end, is also reported in parenthesis. Previously described sequence families have been named in column "Family", according to the current literature; for each of them, the number and typical size of its members are also provided, together with references. For novel families, a systematic name was built by fusing a shortened species name to a progressive number. In the column "type", I, G and S indicate the prevalent genomic location of the members of each families within intergenic, genic or border-spanning sequences. For some families, small previously described sequence motifs contribute to the formation of a substantially larger model; for others, their members are frequently located within larger previously described sequences. In both these cases, the fact is reported in column "notes".

Table 4. Secondary structure prediction analysis of families. Prediction scores of consensus secondary structure, calculated by RNAz, is reported for each family in column "P"; the type of predicted structure is indicated in column "conserved structure", where "s" indicates a stem-loop based structure, while "c" indicates a more complex structure where a stem-loop compatible with the original search is not present. For each family, the aligned localization of the original SLSs is indicated by '+' in column "conserved SLS position"; when SLS alignment is not in agreement with the RNAz prediction, a '°' is added to the '+' symbol. The column marked "SLS folding aptitude" reports the behavior of family elements in the RANDFOLD test: the number of '+' symbols describes the percent of positive elements ('+++' if 90% or above; '++' if 70-90%; '+' if 50-70%; '-' if less than 50%). The localization of family members, as already described in Table 3, is also reported.

Table 5. Correlation of structural properties of the described SLS families to genomic location. "Sec. Struct. +/-" indicates the presence or absence of a conserved secondary structure, as predicted by RNAz; "SLS +/-" indicates the presence or absence of aligned SLSs; "Total" means the sum of rows or columns.

Tables

Division	Species	SLSs	Clusters	Clustered SLSs	Clustered SCRs
low-GC Firmicutes	*Bacillus anthracis*	65,220	4	105	38
	Bacillus halodurans	55,624	6	182	93
	Bacillus subtilis	56,622	2	32	16
	Clostridium perfringens	35,027	6	149	81
	Clostridium tetani	29,883	14	178	123
	Enterococcus faecalis	40,991	7	317	142
	Lactobacillus johnsonii	25,668	3	173	26
	Staphylococcus aureus	32,372	11	275	144
	Streptococcus pneumoniae	25,095	28	825	386
Mollicutes	*Mycoplasma genitalium*	8,953	1	21	8
	Mycoplasma pneumoniae	13,926	20	372	165
high-GC Firmicutes	*Corynebacterium diphtheriae*	54,254	9	282	120
	Mycobacterium leprae	83,094	29	1,721	537
	Mycobacterium tuberculosis	170,502	59	2,182	636
α-Proteobacteria	*Brucella melitensis*	69,899	11	399	219
	Rickettsia conorii	14,933	19	797	383
β-Proteobacteria	*Bordetella bronchiseptica*	214,459	26	2,009	470
	Bordetella parapertussis	188,237	30	1,513	518
	Bordetella pertussis	158,592	52	7,212	4,602
	Neisseria meningitidis	56,605	44	3,595	991
γ-Proteobacteria	*Escherichia coli*	86,339	12	1,152	431
	Haemophilus influenzae	25,055	3	39	25
	Pasteurella multocida	31,209	1	24	8
	Pseudomonas aeruginosa	206,492	9	526	129
	Pseudomonas putida	175,088	75	3,640	1,352
	Salmonella typhi	90,027	8	177	116
	Salmonella typhimurium	91,844	7	157	94
	Vibrio cholerae	45,824	7	250	122
	Yersinia pestis	78,372	20	600	279
TOTAL		**2,230,206**	**523**	**28,904**	**12,254**

Table 1

Species	Clusters	Grouped by			Located within	
		sequence	strand	location	IS	rRNA
B. anthracis	4	3	2	2		
B. halodurans	6	6	4	3		1
B. subtilis	2	2	1	1		1
C. perfringens	6	2	1	1		
C. tetani	14	13	10	6	3	
E. faecalis	7	5	3	3	1	
L. johnsonii	3	3	2	2	1	
S. aureus	11	7	5	4		
S. pneumoniae	28	22	13	9	6	
M. genitalium	1	1	1	1		
M. pneumoniae	20	20	18	12		
C. diphtheriae	9	7	5	4	1	
M. leprae	29	18	11	5		
M. tuberculosis	59	36	21	15	3	
B. melitensis	11	7	5	4		
R. conorii	19	6	4	4		
B. bronchiseptica	26	8	5	4		
B. parapertussis	30	16	10	5	4	
B. pertussis	52	28	16	4	3	
N. meningitidis	44	9	7	6		
E. coli	12	8	6	6		2
H. influenzae	3	1	1	1		
P. multocida	1	1	1	1		
P. aeruginosa	9	5	4	4		
P. putida	75	35	26	14	4	2
S. typhi	8	4	3	3		2
S. typhimurium	7	6	4	4		1
V. cholerae	7	7	5	4		2
Y. pestis	20	15	11	5	2	
Total	523	301	205	137	28	11

Table 2

Species	Family	This work		Literature			Type	Notes
		size	copies	size	copies	ref.		
B. anthracis	Bant-1	72	104 (29)				I	
	Bcr1	167	31 (21)	147	12	[24]	I	
B. halodurans	Bhal-1	74	36 (32)				I	
	Bhal-2	76	50 (41)				I	contains CRISPR repeats
C. perfringens	Clop-1	93	44 (28)				I	
C. tetani	Clot-1	74	19 (16)				I	
	Clot-2	31	34 (32)					contains CRISPR repeats
	Clot-3	90	24 (17)				I	contains CRISPR repeats
E. faecalis	Efa-1	163	65 (18)				I	
	Efa-2	292	11 (9)				G	
L. johnsonii	Lac-1	231	34 (6)				G	
S. aureus	Sta-1	105	25 (25)				I	
	Sta-2	460	9 (8)				S	
	Sta-3	136	24 (15)				I	
	Sta-4	99	46 (27)				I	
S. pneumoniae	BOX	84	205(105)	100-200	127	[25]	I	
	RUP	63	110 (99)	108	54	[26]	I	
	Stre-1	45	241(225)				G	
B. melithensis	Bru-RS	118	222 (69)	103-105	35-40	[27]	I	
R. conorii	Rpe-4	100	97 (74)	95	94	[28]	I	
	Rpe-5	115	45 (35)	115	55	[28]	I	
	Rpe-6	108	123 (74)	136	168	[28]		
	Rpe-7	123	186 144)	99	223	[28]		
M. genitalium	Myg-1	259	10 (7)				I	
M. pneumoniae	Myp-1	143	25 (18)				G	part of REPMP1 repeat
	Myp-2	158	42 (16)				G	part of REPMP4 repeat
	Myp-3	558	11 (8)				G	part of REPMP5 repeat
	Myp-4	364	8 (7)				G	part of REPMP5 repeat
	Myp-5	426	8 (8)				G	part of REPMP5 repeat
	Myp-6	468	11 (11)				G	part of REPMP2/3 repeat
	Myp-8	674	9 (9)				G	part of REPMP2/3 repeat
	Myp-9	226	9 (9)				G	part of REPMP2/3 repeat
	Myp-10	330	12 (12)				G	part of REPMP2/3 repeat
	Myp-7	131	42 (22)				G	
C. diphtheriae	Cod-1	140	17 (16)				I	
	Cod-2	32	43 (39)				G	
	Cod-3	170	23 (20)					
	Cod-5	74	35 (29)				I	
M. tuberculosis	Myt-1	72	75 (70)					
	Myt-2	115	769(223)				G	located within PE genes
	Myt-3	81	81 (77)				G	located within PE genes
	Myt-4	83	196 (68)				G	located within PE genes
	Myt-5	71	41 (2)				G	contains CRISPR repeats
	Myt-7	136	278 (68)				G	located within PE genes
	Myt-8	92	33 (25)					
	Myt-9	67	53 (15)					
	Myt-10	154	62 (59)				G	located within PE genes
	Myt-11	65	56 (21)					contains MIRU repeats
M. leprae	REPLEP	740	29 (9)	400-880	15	[29]	I	
	RLEP	641	38 (30)	601-1075	37	[29]	S	
	Myl-1	371	7 (4)				S	part of LEPREP repeat
	Myl-2	1979	9 (7)				S	part of LEPREP repeat

Table 3 (part 1)

Species	Family	This work		Literature			Type	Notes
		size	copies	size	copies	ref.		
B. bronchiseptica	Bor-1	117	196 (92)				I	
	Bor-2	167	17 (6)				I	
	Bor-3	134	34 (32)				G	
	Bor-4	81	164(114)				G	
	Bor-5	112	135(101)				G	
	Bor-6	147	37 (31)				G	
B. pertussis	Bor-1	93	128 (78)				I	
N. meningitidis	ATR	206	14 (9)	183	13	[30]	I	
	Nem-2	341	11 (7)					
	Nem-3	127	10 (9)				G	
	Nem-4	36	412(362)				I	contains DUS repeats
	dRS3	33	755(708)	20	770	[30]	I	
	NEMIS	46	262 (81)	106-158	250	[13]	I	
	Rep2	65	22 (18)	59-154	26	[30]	I	
P. multocida	Pam-1	155	12 (12)				S	contains DUS repeats
E. coli	BoxC	50	22 (20)	56	32	[31]		
	Eco-1	734	9 (7)				G	
	ERIC	140	19 (19)	127	21	[32]	S	
	PU-BIME	108	301(199)	40	485	[31]		
H. influenzae	Hin-1	31	53 (51)				I	contains DUS repeats
P. aeruginosa	Pae-1	84	133 (61)				I	
	Pae-2	287	65 (24)				G	
	Pae-3	220	16 (13)				G	
	Pae-4	52	41 (35)					
P. putida	Ppu-1	617	39 (28)				I	
	Ppu-2	2056	10 (8)				S	
	Ppu-3	251	27 (23)				G	
	Ppu-4	81	41 (24)				I	
	Ppu-9	124	57 (31)				I	
	REP	39	588(496)	30	804	[33]	I	
S. typhi	PU-BIME	43	146(126)	40	100	[31]	I	
	PU-BIME*	80	59 (37)	40	>100	[31]		
S. typhimurium	PU-BIME	78	142 (94)	40	82	[31]		
	Sal-1	115	27 (17)				I	
	Sal-2	120	33 (3)				G	contains CRISPR repeats
V. cholerae	ERIC	103	97 (66)	127	80	[31]	I	
	Vic-1	184	14 (1)				I	
Y. pestis	ERIC	115	241(128)	69-127	167	[16]	I	
	YPAL	168	101 (68)	169	30	[17]	I	
	YPAL*	136	26 (13)	130	10	[17]	I	

Table 3 (part 2)

Species	Family	P	Conserved structure	Conserved SLS position	SLS folding aptitude	Type
B. anthracis	Bcr1	0.99	s	+	+	I
B. halodurans	Bha1-1	0.98	s	+	++	I
	Bha1-2	0.99	c		-	I
C. perfringens	Clop-1	0.96	s	+	+	I
C. tetani	Clot-1	0.95	s	+	++	I
E. faecalis	Efa-1	0.85	s	+	+++	I
	Efa-2	1.00	s	+	-	G
L. johnsonii	Lac-1	0.97	c	+°	-	G
S. aureus	Sta-1	0.84	s	+	+++	I
	Sta-2	1.00	s	+	++	S
	Sta-3	0.97	s	+	+	I
B. melithensis	Bru-RS	0.98	s	+	+	I
R. conorii	Rpe-4	0.73	s	+	-	I
	Rpe-5	1.00	s	+	+	I
	Rpe-6	0.45	-	+°	+	
	Rpe-7	0.99	s	+	++	
M. genitalium	Myg-1	0.06	-	+°	-	I
M. pneumoniae	Myp-1	0.00	-	+°	-	G
	Myp-2	0.95	s	+	++	G
	Myp-3	0.89	s	+	-	G
	Myp-4	0.09	-	+°	-	G
	Myp-5	0.74	s	+	-	G
	Myp-6	0.55	c		-	G
	Myp-7	0.67	s	+	-	G
C. diphtheriae	Cod-1	0.97	s	+	+++	I
	Cod-2	0.98	s		-	G
	Cod-3	0.99	s	+	+++	
M. tuberculosis	Myt-1	0.74	s	+	+++	
	Myt-8	0.90	s	+	++	
M. leprae	REPLEP	1.00	c	+°	-	I
	RLEP	1.00	s	+	++	S
	Myl-1	0.61	s	+	++	S
	Myl-2	0.97	s	+	+	S
B. bronchiseptica	Bor-1	0.86	s	+	++	I
	Bor-2	1.00	s	+	-	I
B. pertussis	Bor-1	0.93	s	+	++	I
N. meningitidis	ATR	1.00	s	+	-	I
	Nem-2	0.93	s	+	+	
	Nem-4	0.93	s	+	+++	I
	dRS3	0.98	c		-	I
	NEMIS	1.00	s	+	+	I
	Rep2	0.98	s	+	+	I
P. multocida	Pam-1	0.96	s	+	+++	S
E. coli	BoxC	0.99	c	+°	-	
	Eco-1	0.18	-	+°	-	G
	ERIC	0.94	s	+	++	S
	PU-BIME	0.94	s	+	+	
H. influenzae	Hin-1	0.96	s	+	+	I
P. aeruginosa	Pae-1	0.97	s	+	++	I
	Pae-3	0.26	-	+°	-	G
	Pae-4	0.93	s	+	++	
P. putida	Ppu-1	0.97	s	+	+	I
	Ppu-2	1.00	s	+	+++	S
	Ppu-4	0.95	s	+	-	I
	Ppu-9	0.54	s	+	-	I
S. typhi	PU-BIME	0.97	c		-	I
	PU*-BIME	0.98	s	+	-	
S. typhimurium	PU-BIME	0.98	s	+	-	
	Sal-1	0.94	c		-	I
	Sal-2	1.00	c		-	G
Y. pestis	ERIC	0.90	s	+	-	I
	YPAL	1.00	s	+	+++	I
	YPAL*	0.96	c		-	I

Table 4

Structural data / Genomic location	Sec. Struct. +		Sec. Struct. -		Total
	SLS +	SLS -	SLS +	SLS -	
Genic	5	4	4	17	30
Border spanning	7	0	0	0	7
Intergenic	25	6	1	9	41
Others	9	1	1	3	14
Total	46	11	6	29	92

Table 5

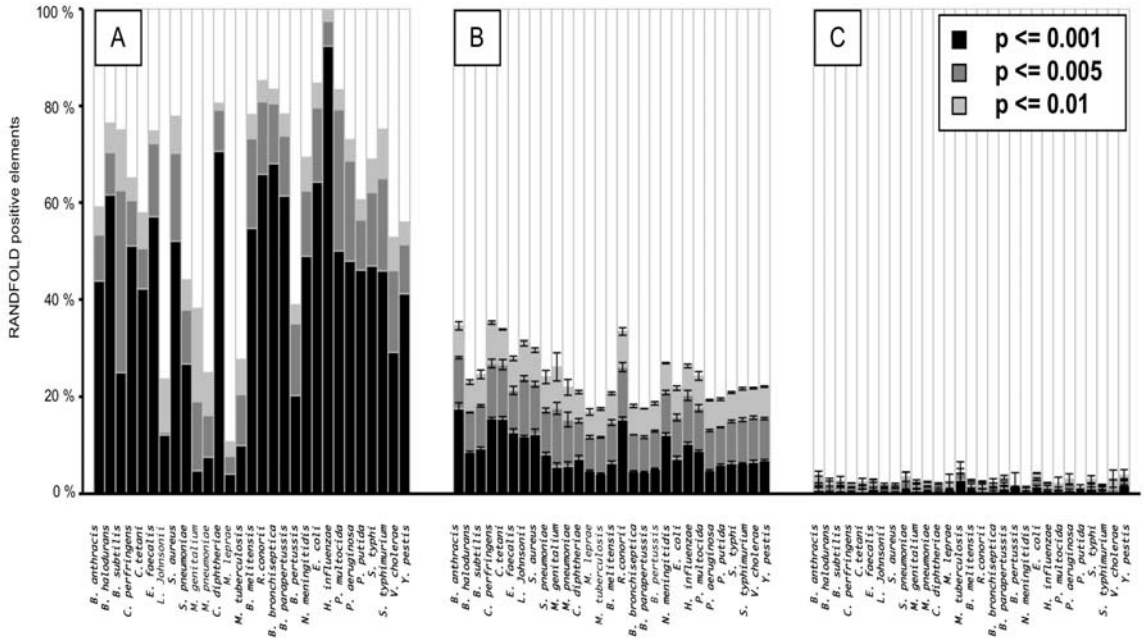

A
B
C
p <= 0.001
p <= 0.005
p <= 0.01
RANDFOLD positive elements
100 %
80 %
60 %
40 %
20 %
0 %
B. anthracis
B. halodurans
B. subtilis
C. perfringens
C.tetani
E. faecalis
L. Johnsonii
S. aureus
S. pneumoniae
M. genitalium
M. pneumoniae
C. diphtheriae
M. leprae
M. tuberculosis
B. melitensis
R.conorii
B. bronchiseptica
B. parapertussis
B. pertussis
N. meningitidis
E. coli
H. influenzae
P. multocida
P. aeruginosa
P. putida
S. typhi
S. typhimurium
V. cholerae
Y. pestis

Figure 1

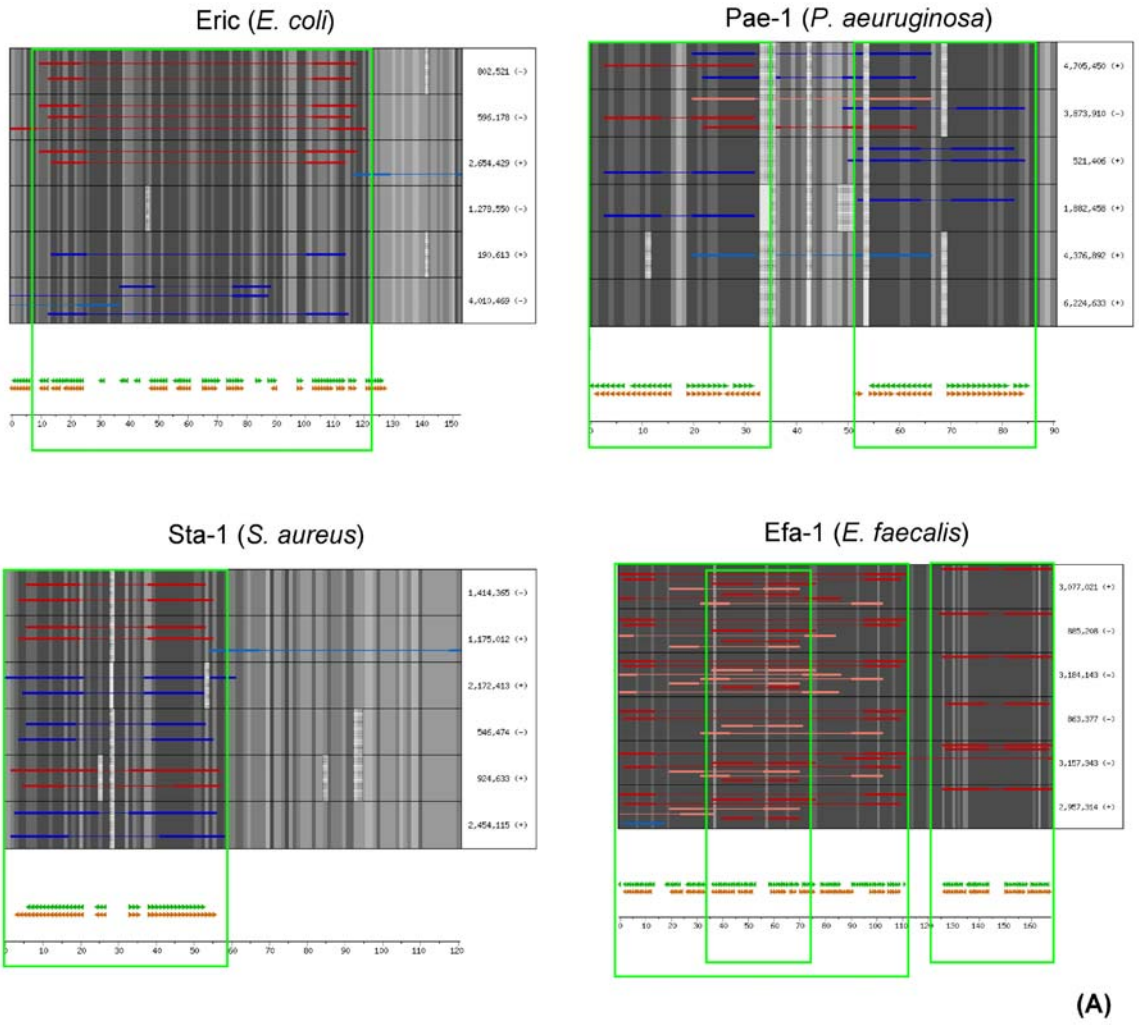

Eric (*E. coli*)
Pae-1 (*P. aeuruginosa*)
Sta-1 (*S. aureus*)
Efa-1 (*E. faecalis*)
(A)
Eric
Efa-1
Pae-1
Sta-1
(B)

Figure 2

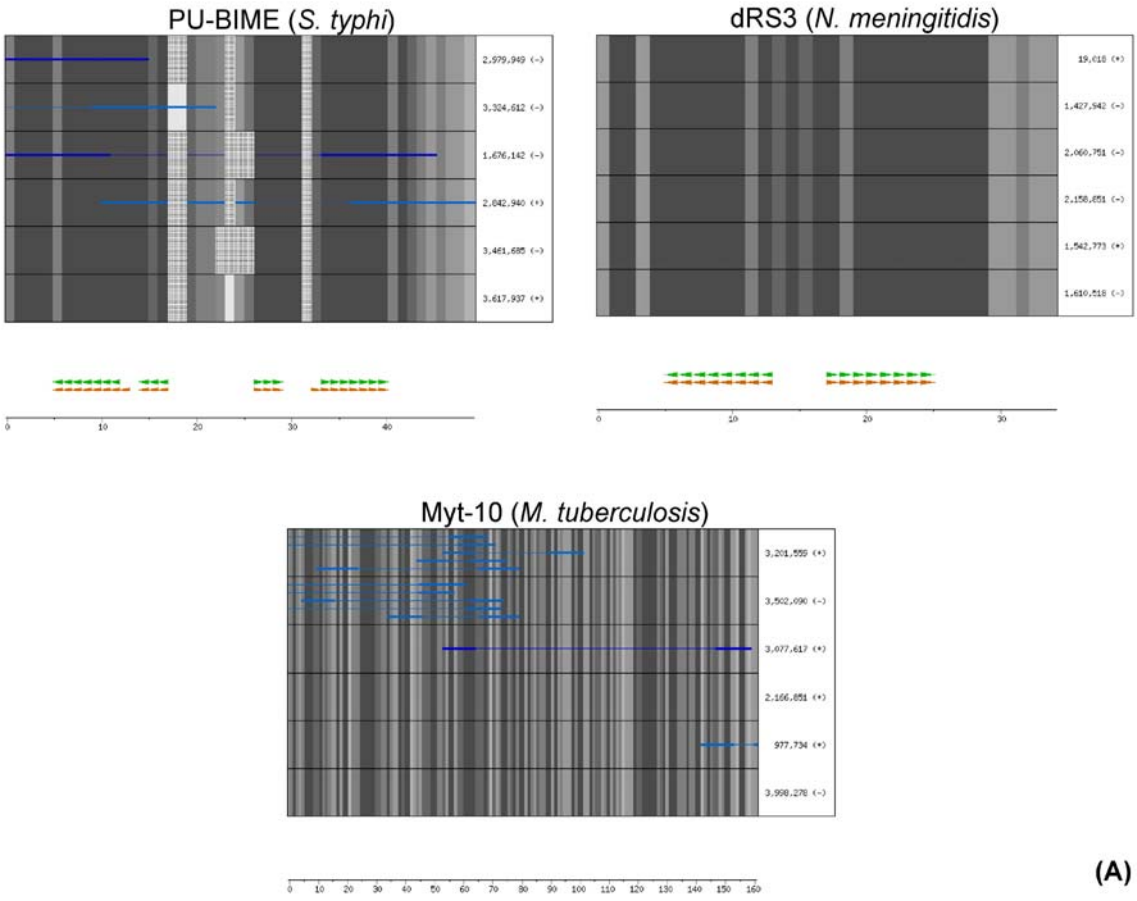

PU-BIME (S. typhi)
dRS3 (N. meningitidis)
Myt-10 (M. tuberculosis)
(A)
PU-BIME
dRS3
(B)

Figure 3