

Short repeats and IS elements in the extremely radiation-resistant bacterium *Deinococcus radiodurans* and comparison to other bacterial species

Kira S. Makarova^{a, b, d}, Yuri I. Wolf^{a, b}, Owen White^c, Ken Minton^d, Michael J. Daly^{d*}

^aNational Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

^bPermanent address: Institute of Cytology and Genetics, Russian Academy of Sciences, Novosibirsk 630090, Russia

^cThe Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA

^dDepartment of Pathology, Rm. B3153, Uniformed Services University of the Health Sciences, 4301 Jones Bridge Road, Bethesda, MD 20814, USA

Abstract— Computer analysis of the complete genome of *Deinococcus radiodurans* R1 has shown that the number of insertion sequences (ISs) and small noncoding repeats (SNRs) it contains is very high, and comparable with those of *Escherichia coli*. IS elements and several families of SNRs are described, together with their possible function in the *D. radiodurans* genome. © 1999 Éditions scientifiques et médicales Elsevier SAS

Deinococcus radiodurans / repeated sequences / genome analysis

1. Introduction

Rapid sequencing of complete genomes of living organisms from a wide variety of taxonomic groups lead to the origin of comparative genomics. Recent improvements in database management and sequence comparison methods have facilitated fast and efficient genome annotation [17]. Until recently, gene sequence analysis has been the major focus, primarily because of the broad scientific interest in characterizing the relationship between amino acid sequence, protein structure, function, and evolution. Much less is understood about the function of intergenic regions, where the sequence complexity and potential function is only now being revealed and analyzed for the first time, at the level of whole genomes.

Escherichia coli has been the subject of whole genome sequencing recently and is now the most functionally characterized bacterium, for

both intra- and intergenic sequences [6]. Analyses of its intergenic regions have been reported extensively [3], and show that they contain promoters, regulatory sites, and numerous repeats. While the functions of these promoters and regulatory sites have been characterized in great detail in *E. coli*, there is comparatively little understanding of the structure and function of intergenic regions, and especially repeats, in other bacteria. *E. coli* is currently the only organism for which abundant intergenic repeats have been described; however, their function(s) remain largely undefined [3]. Within the sequenced bacterial genomes, generally, intergenic regions have been shown to be highly variable in promoter organization and repeat composition [6, 8, 15, 19, 21, 22, 30]. This intergenic heterogeneity, seen even between closely related species, suggests to us that the organization of such regions is highly species-specific. [9, 24, 31].

Deinococcus radiodurans is well known for its extreme radiation resistance phenotype; cells can survive acute doses of γ -radiation up to 1 700 000 rad without any lethality or increasing

* Correspondence and reprints
Tel.: +1 301 295 3750; fax: +1 301 295 1640;
mdaly@usuhs.mil

mutation frequency [10]. By comparison, *E. coli* is 100–500 times more sensitive [29]. *D. radiodurans*' supreme resistance has been the subject of numerous studies examining its DNA repair mechanisms [11, 12], and currently, it is being engineered for bioremediation of radioactive waste sites [23]. Both of these areas of research have attracted much interest and were the impetus for sequencing the *D. radiodurans* strain R1 genome, that consists of a chromosome (DR_MAIN:2.65 Mbp), two megaplasmids (DR412: 412 kbp; DR177: 177 kbp), and a plasmid (46 kbp) (O. White et al. Science 286 (1999) 1571–1577). Among this genome's interesting features are 52 insertion sequences (ISs) and 247 small noncoding repeats (SNRs), found by computer analysis. The number of repeats and their distribution in *D. radiodurans* is comparable to those found in the evolutionarily distant *E. coli*; repeats appear to be a complex genetic trait, and their evolutionary significance and role in genome function remain unclear.

2. Results and discussion

2.1. ISs in the *D. radiodurans* R1 genome and comparison with other bacteria

During the annotation of the *D. radiodurans* genome, 12 distinct open reading frames (ORFs) were found that have sequence similarity to transposases of several different IS families [7] (table 1), and several of these ORFs exist in multiple copies. This is the first report and analysis of IS elements in *D. radiodurans*, with the exception of IS2621, that was identified previously [27]. For most of these elements (IS4_DR, TCL9, TCL121, TCL23, IS3_DR, and AXL_DR) we were able to identify the precise length. All of these elements have the usual features compared to those ISs identified in other organisms [16]. For example, they contain one or two ORFs that encode a transcriptional regulator and a transposase, as well as having inverted terminal repeats and/or internal repeats (data not shown). Three elements (TCL9,

Table 1. Distribution of ISs.

Name	Family	Length (bp)	Copy number				Total length (bp)
			Plasmid	DR177	DR412	DR_MAIN	
IS2621	IS4	1322	0	6	1	6	17186
IS2621 5' fragment	-	25	0	1	2	4	N/A
IS4_DR	IS4	1207	4	6	0	3	15942
IS605_DR	IS605	~ 1060	0	0	0	8	8480
TCL9	Tc1/mariner	1048	0	1	0	4	5250
TCL121	Tc1/mariner	1073	0	2	0	1	3210
TCL23	Tc1/mariner	1069	1	1	0	1	3207
AXL_DR	Tc1/mariner	912	1	0	0	1	1824
IS3_DR	IS3	1304	0	1	0	0	1300
SCL_DR	Related to mini-circle element of <i>Streptomyces coelicolor</i>	~ 600	0	0	0	1	600
VCL_DR	IS15	~ 500	1	0	0	0	1500
DNIIV_DR	DNA invertase	~ 600	1	0	0	0	600
TNPA_DR	TNPA	~ 3000	1	0	0	0	3000
Total 52 copies			9	17	1	25	62099
Number of copies per 10000 nucleotides			1.97	0.96	0.02	0.09	

TCL121, TCL23) of the Tc1-mariner family are closely related and likely to be a product of a recent duplication, probably specific to the *Deinococcus* lineage (data not shown). For elements that are heterogeneous, it was not possible to determine the specific length for these (e.g., IS605) [20] or for those that are present in the *D. radiodurans* genome in only a single copy and that do not contain typical structural features (VCL_DR, SCL_DR, and DNIV_DR).

Overall, we detected 52 copies of IS elements in the *D. radiodurans* genome (table 1). The three most abundant ISs are IS4_DR (13 copies), IS2621 (11 copies), and IS200_DR (eight copies). The distribution of IS elements within the chromosome and plasmids is variable; the number of copies per 10 000 nucleotides in the plasmid and megaplasmid DR177 is more than ten times higher than the number found in the megaplasmid DR412 or the chromosome. Only one IS element is present in DR412, whereas nine IS elements are present in the 46-kbp plasmid. There are five elements in the *D. radiodurans* genome that exist as single copies and these may be transpositionally inactive or, perhaps, were only recently acquired by strain R1. It is noteworthy that three of these single copy elements are located in the 46-kbp plasmid. Together, these observations indicate that the IS distribution in *D. radiodurans* is nonuniform.

We also have compared the number of IS copies in *D. radiodurans* to the numbers found in other sequenced bacterial genomes [6, 8, 21, 22]. *D. radiodurans* contains the largest number of these elements (52; figure 1A) and *E. coli*, with its 36 IS elements, ranks in second place. Considering their respective genome sizes, *D. radiodurans* has 16.3 ISs per 1 000 genes, whereas *E. coli* has only 8.4. Since *Bacillus subtilis* does not contain any IS elements, it is apparent that these sequences are not an indispensable part of a bacterial genome. If the number of elements is a reflection of transposition activity, then this would be expected to influence genome instability, and it would seem that *D. radiodurans* is able to survive high levels of genomic rearrangement. Yet there is little experimental evidence for transposition in *D. radiodurans*; in the

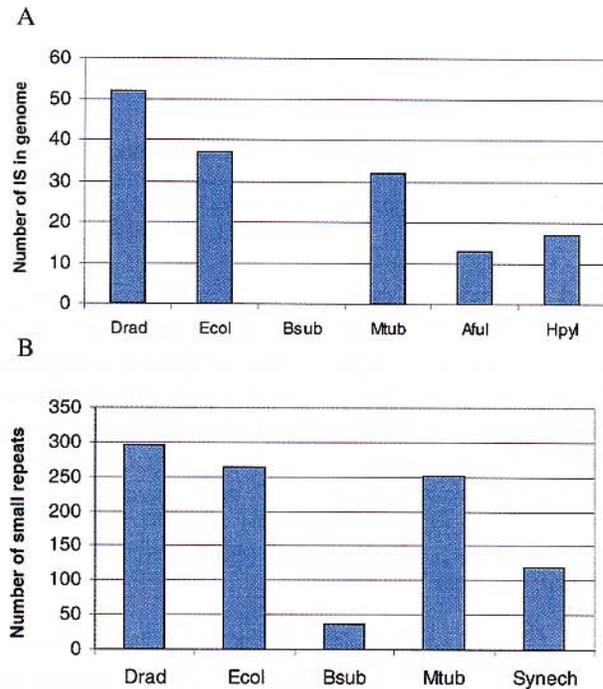


Figure 1. A. Abundance of IS elements in complete genomes. Drad - *Deinococcus radiodurans*; Ecol - *Escherichia coli*; Bsub - *Bacillus subtilis*; Mtub - *Mycobacterium tuberculosis*; Aful - *Archaeoglobus fulgidus*; Hpyl - *Helicobacter pylori*. **B.** Number of small noncoding extragenic repeats. Abbreviations are as in A; Synech - *Synechocystis* sp.

whole *D. radiodurans* R1 genome there is only one example of a gene disruption by an IS element – IS2621 is inserted into the gene for alkaline serine exoprotease A (aqualysin I). Similarly, there has been only one experimentally detected IS-induced mutation in *D. radiodurans* (*uvrA* [27]). Whether the *D. radiodurans* insertional elements are transpositionally functional or involved or not in genome instability, is the subject of ongoing investigations.

In the analysis of IS2621, it was determined that it is the only element in *D. radiodurans* that has several smaller fragments of similarity dispersed throughout the genome (table 1). All of these fragments have the same length and are identical in sequence. They consist of the first 25 nucleotides of IS2621 with a single A to T substitution at position 21, near the end of the 5' terminal repeat (figure 2). This substitution may

Table II. Distribution of small noncoding repeats.

Name	Length (bp)	Copy number			
		Plasmid	DR177	DR412	DR_MAIN
SRE	160	0	3	4	32
SNR1	139	0	0	1	39
SNR2	114	0	0	8	76
SNR4	147	0	1	2	4
SNR5	215	0	0	1	27
SNR7	140	0	2	0	14
SNR8	131	0	0	1	19
SNR9	105	0	0	1	6
SNR10	60	0	0	0	6
Total sum		0	6	18	223
Number of copies per 10000 nucleotides		0	0.3	0.4	0.8

overall trend in their abundance within the different genomes evaluated (*figure 1B*). For *D. radiodurans*, we confirmed all SNR hits reported by BLASTN individually, and had to reduce the total number of SNRs from 295 to 247. A similar correction should be performed on the other genomes.

2.2.2. The mosaic organization of SNRs in *D. radiodurans*

D. radiodurans SNRs have complex configurations. One way to evaluate this complexity is to consider SNRs as being composed of several relatively conserved modules, as is shown in *figure 3*. In these examples, BLASTN reported distinct modules that are 18–20 nucleotides long, with hits that span the length of the SNRs. Surprisingly, it appears that these modules are poorly associated with certain distinct internal features of some SNR elements, such as inverted and/or reverse repeats (*figures 4* and *5*).

A more detailed analysis of the modular architecture of SNRs was done for the most prevalent SNR family (SNR2; *figure 5*). From the analysis of BLASTN results, we identified five modules in the longest representative of this family (*figure 5A*). Modules I (also part of the SRE family; *figure 4*) and V contain two parts of the inverted repeat present in SNR2. The different configurations of the SNR2 family are shown in *figure 5B*. This data suggests that in the evolution of SNRs, deletions and insertions are likely to have played an important role. For

example, module III is prone to be missing when both modules II and IV are present.

2.2.3. Distribution of SNRs along the chromosome of *D. radiodurans*

If the placement of SNRs is determined by some functional demand or their propagation history, then one would expect to find clear trends in their distribution; for example, a tendency to form clusters or, conversely, to be evenly spaced. Alternatively, one can postulate a purely random placement of SNRs within intergenic regions, i.e., that the probability of SNR occurrence after any nucleotide within a given intergenic region is uniform. If this 'random' hypothesis is true, then the distance between adjacent SNRs (minus the length of genes that intervene between them) should follow a geometric distribution. To test this hypothesis, we mapped the positions of all SNRs in the chromosome (*figure 6*) and calculated the distances between adjacent repeats. This test for position randomness was performed for all repeats considered together, as well as individually for the four largest families (*table III*, *figure 7*). Both analyses showed that, with one exception, there is no significant deviation from the random placement model. The exception is the SNR5 family that shows a tendency ($P < 0.05$) to occur closer to each other than predicted by the random model. Further, as shown in *table IV*, there is no significant corre-

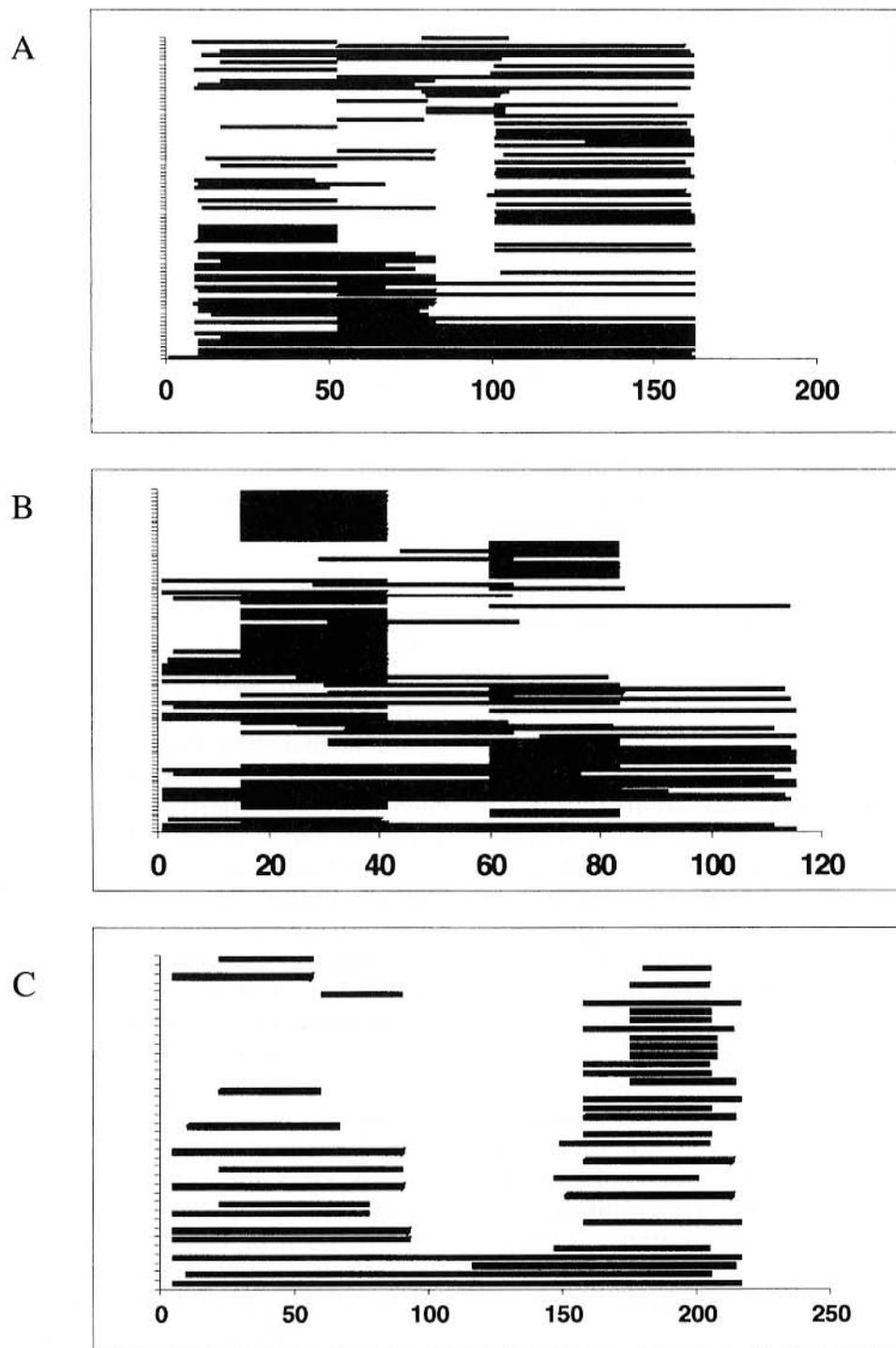


Figure 3. Mosaic organization of SNRs - Map of BLASTN hits to a full-length query sequence. **A.** SRE; **B.** SNR2; **C.** SNR5.

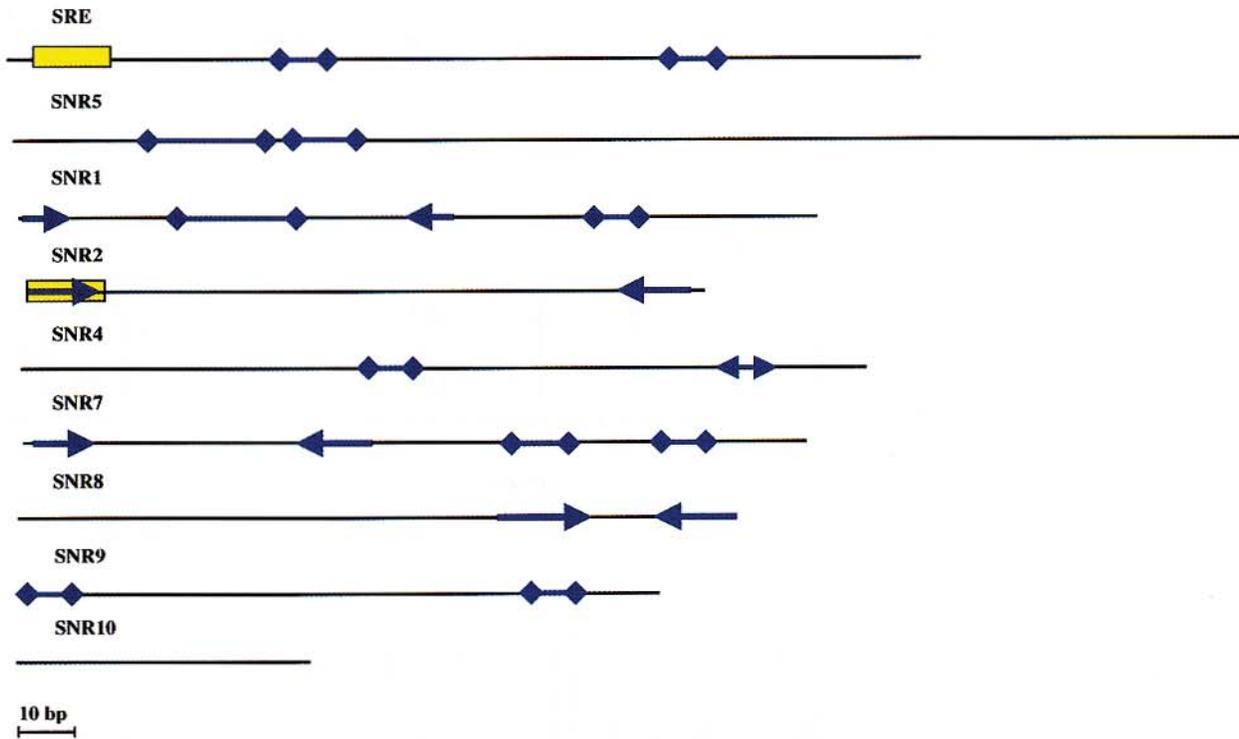


Figure 4. Architecture of SNRs. Inverted repeats are marked by arrows. Reverse repeats are defined by regions bound by diamonds. SRE and SNR2 share the module represented as a rectangle.

lation between the direction of a repeat and the direction of an adjacent gene.

Thus, SNRs are not likely to play a direct role in the regulation of transcription or translation. This conclusion is based on the following observations: 1) SNRs appear to be randomly distrib-

uted in intergenic regions; 2) their direction cannot be correlated with the direction of flanking genes; and 3) there is no apparent relationship between a particular SNR family and the function of genes flanking these repeats. It should be noted in this context that, while some

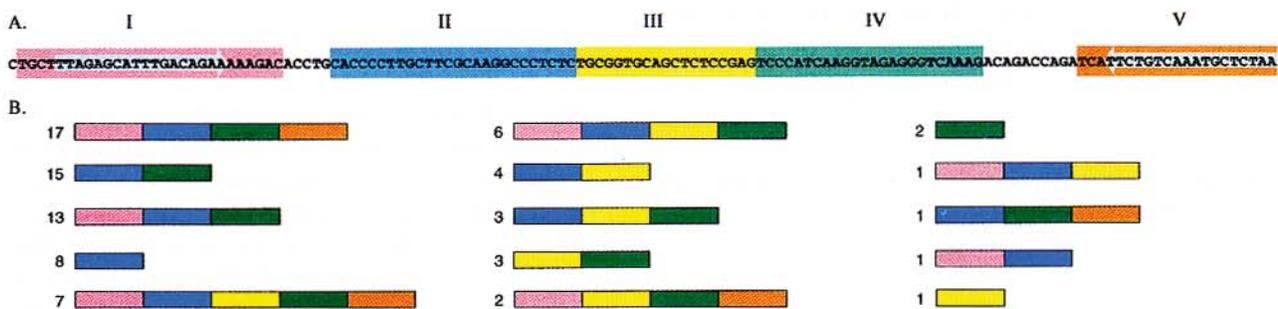


Figure 5. A. Structure of the full-length repeated member of the SNR2 family. Inverted repeats are marked by arrows. Roman numerals and different colors mark the five conserved modules. **B.** Number of SNR2 members with the indicated modular configuration. Each class of module is represented by a different color.

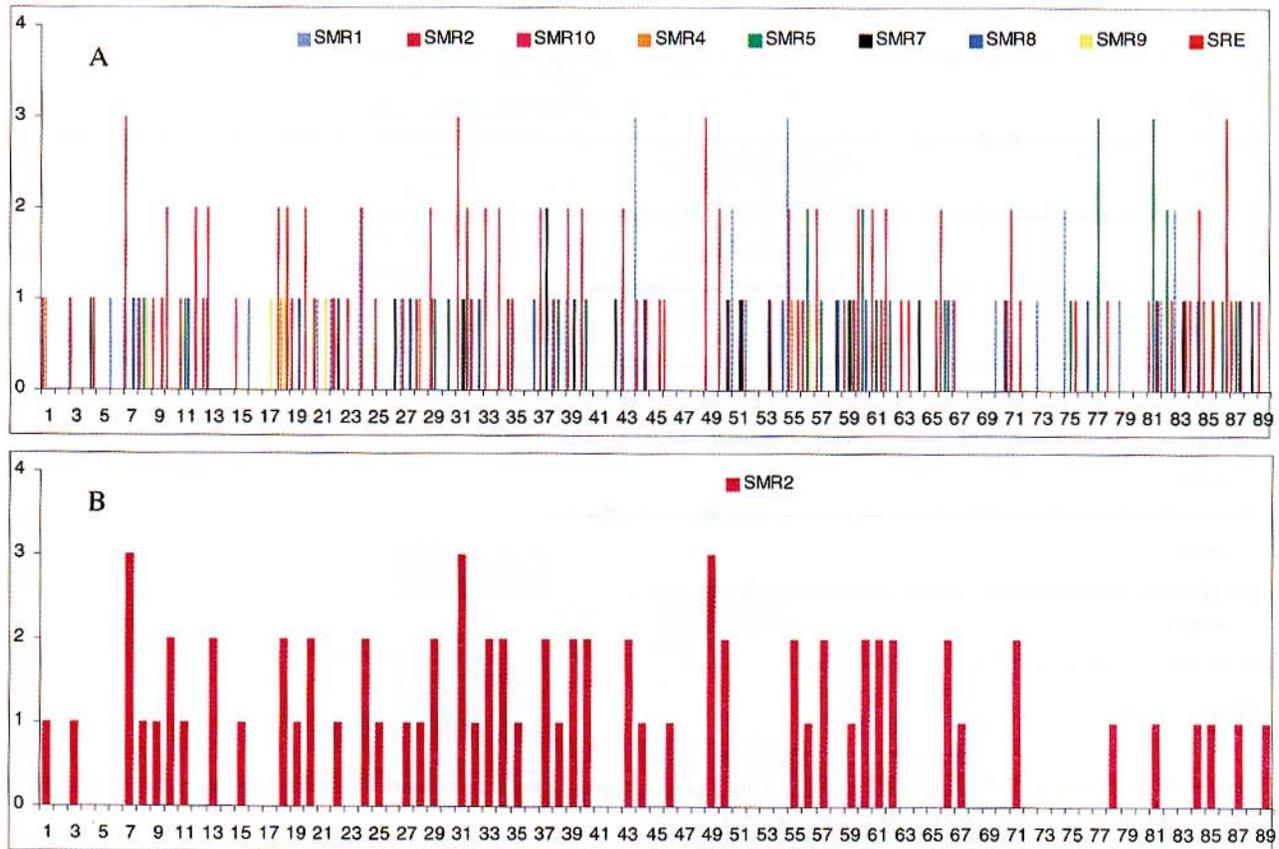


Figure 6. Map of SNRs in the chromosome. Horizontal axis, chromosome fragment number (each fragment equals 30,000 bp). Vertical axis, number of repeats in the corresponding chromosome fragment. A: All repeats; B: SNR2 family.

D. radiodurans SNRs have characteristics similar to the *E. coli* families of small repeats (BIMEs, or bacterial interspersed mosaic elements [18]), SNRs do not share features with *E. coli* translation *rho*-independent terminators (Ter repeats [5]). In addition, the energy of potential RNA secondary structures, predicted for

D. radiodurans SNRs, does not differ from the values obtained for coding regions or other sequence fragments unrelated to SNRs (data not shown).

2.2.4. Comparison of SREs from *D. radiodurans* strains R1 and SARK

The first report of a *D. radiodurans* sequence that contains a SNR was published long before the complete genome sequence for *D. radiodurans* R1 became available [25]. In this early work, the repeated sequence was named SRE (for *D. radiodurans* strain SARK repetitive element). We therefore used the name SRE for the corresponding SNR family in *D. radiodurans* strain R1. These two sources of SNR sequence data, from the closely related strains SARK and

Table III. Testing the hypothesis of random repeat distribution along the main chromosome.

Repeat	Number	$P(\chi^2)$
All	223	0.058
SRE	32	0.997
SNR1	39	0.607
SNR2	76	0.682
SNR5	27	0.028

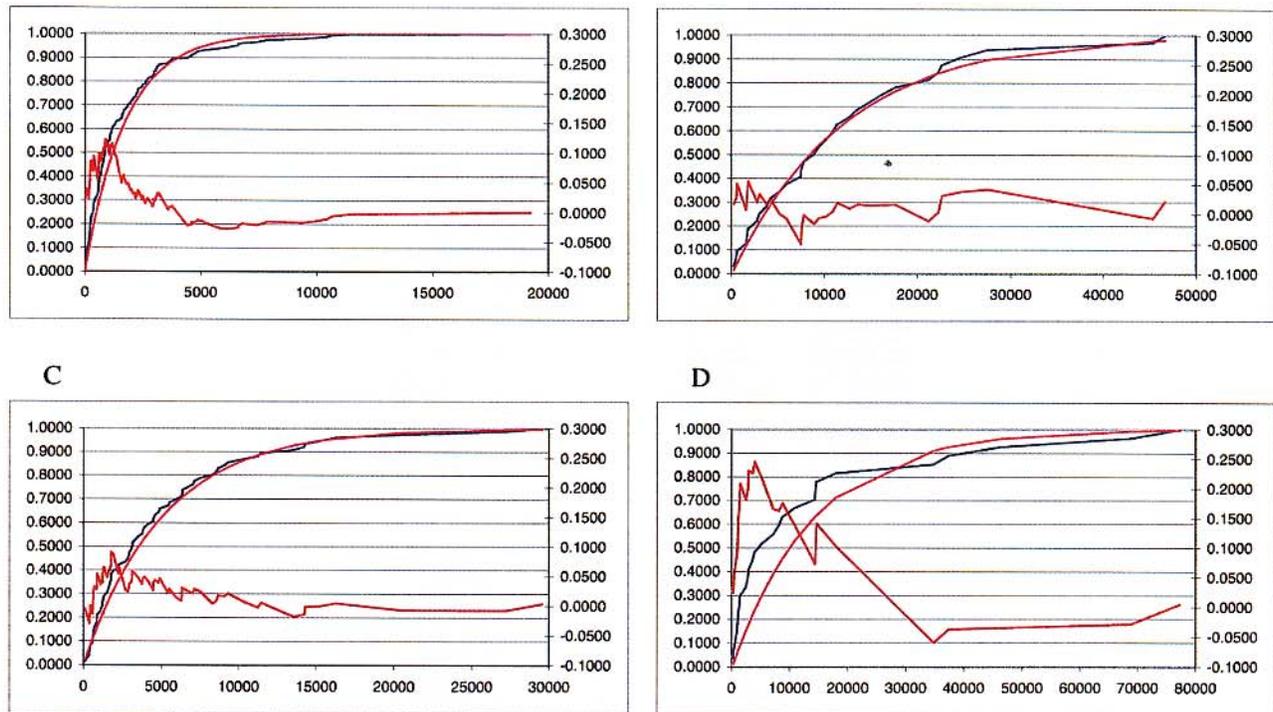


Figure 7. Distribution of distances between repeats in the chromosome (blue, experimental data; magenta, exponential approximation - left axis), and difference between experimental and approximating distribution functions (red - right axis). **A.** All repeats; **B.** SRE; **C.** SNR2; **D.** SNR5

R1, provide an opportunity to compare two evolutionary distinct, but closely related SNRs. *Figure 8* shows a multiple alignment of SRE sequences from both strains, and a maximum likelihood tree constructed from this data is shown in *figure 9*. Within this tree, there is a separate branch for SREs derived from SARK, even though the SARK SRE1 element contains an insertion relative to SRE4, SRE5, SRE11, and

SRE29. Notably, most of the strain-specific substitutions are located in the central regions of two pairs of inverted repeats; these repeats may form two hairpin-like structures [25]. However, the R1 substitutions may disrupt these hairpins. The predicted energy for the consensus hairpin I in SARK is -11.2 kcal/mole, but only -6.1 kcal/mole in R1, as estimated by the Mfold program [26]; for hairpin II it is -17.0 kcal/mole and -11.0 kcal/mole, respectively.

Table IV. Correlation between the repeat direction and the direction of adjacent genes.

Repeat	Number	$P(\chi^2)$ for the left gene	$P(\chi^2)$ for the right gene
SRE	32	0.50	0.70
SNR1	39	1.00	0.54
SNR2	76	0.88	0.70
SNR5	27	0.47	0.07
SNR7	14	0.08	0.53
SNR8	19	0.21	0.76

The nonrandom clustering of the strain-specific nucleotide substitutions in strain R1 is tantalizing and numerous explanations for this difference can be proposed. One can speculate that there is a strain-specific selection pressure for either strengthening (in SARK) or disrupting (in R1) these hairpins within this particular repeat family. If so, this would suggest that the SRE repeat affects genome function. The second possibility is that the multiple substitutions in

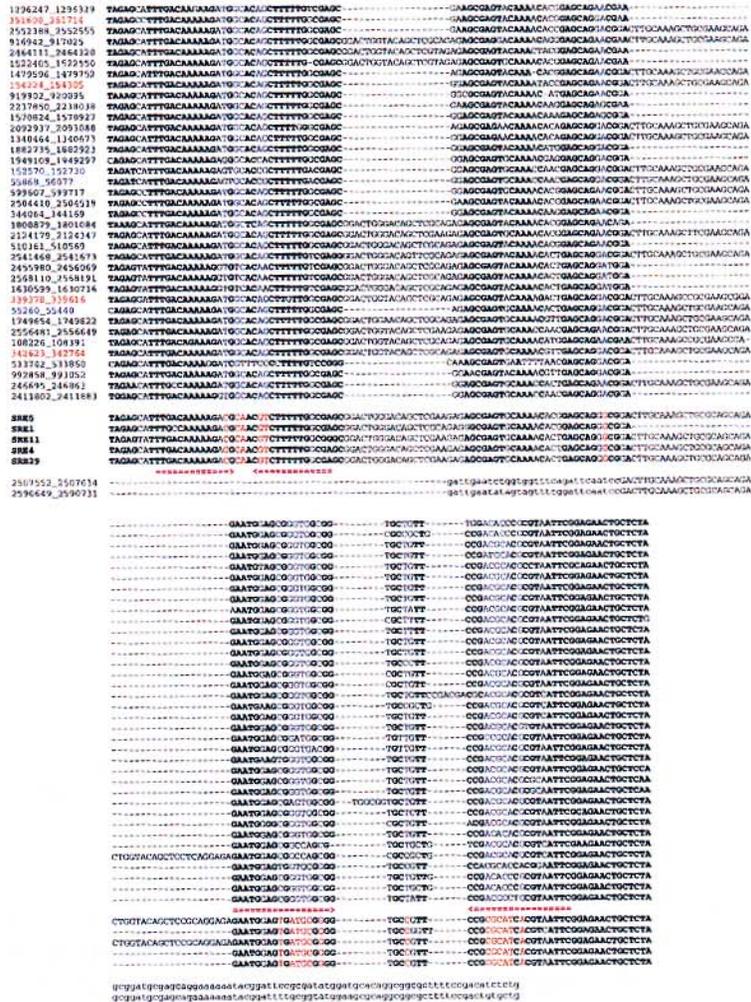


Figure 8. Multiple alignment of the SRE sequences from *D. radiodurans* strains SARK and R1. SREs are denoted by their start and end positions, within their genome partition. SREs found in the chromosome are marked in black; for DR412 in red, and for DR177 in blue; SREs from SARK are in bold. SARK-specific nucleotide alignment positions are shown in red; R1-specific positions in blue; and conserved positions (less than 4 substitutions) in bold. The positions of potential hairpins are shown by magenta arrows. The bottom two sequences are aligned only in the area of the conserved insertion.

the hairpin represent regions of the SRE that are hot-spots for spontaneous mutagenesis. Alternatively, these substitution clusters may have resulted from a single mutation event that occurred shortly after the divergence of R1 and SARK. If this is the case, then their occurrence in the hairpin region may be attributed to chance.

The familial tree of SREs shows that these elements do not cluster in the plasmid, megaplasmids, or the chromosome, in a particular way (figure 8). For example, all four SREs from

megaplasmid DR412 are located in different branches (figure 9), whereas three other SREs, found in DR177, occupy two distinct positions in the tree. However, two elements (start positions 152 570 and 558 68) in DR412 are identical, and there is one other pair of identical repeats in the chromosome (start positions 2 568 110 and 1 630 599). The absence of any clear correlation between the location of an element and its position in the tree, suggests that both *cis*- and *trans*-SRE propagation events may occur.

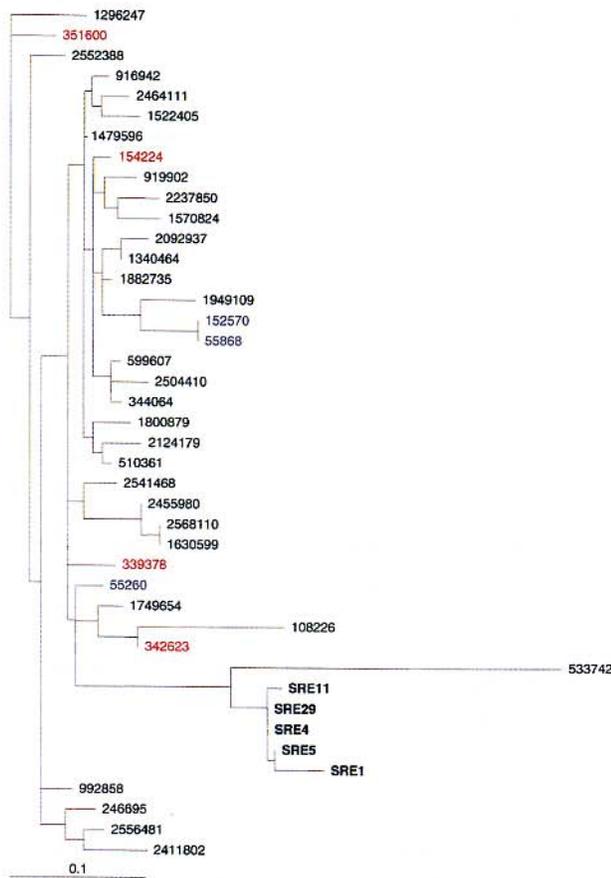


Figure 9. The maximum likelihood tree for the SRE repeat family, built using the PHYLIP program [14]. For R1 SREs, only the start positions are shown. Other designations are as in figure 8.

A comparison of SREs also reveals a modular organization similar to that shown for the SNR2 family (figure 5). In addition to module (I), shared by SRE and SNR2, there is another module present in a repeat family, that is not included in the main list. It was excluded because there are only two copies of it in the *D. radiodurans* genome (sequences 2507552 and 2590649; figure 8, bottom). As described earlier, there is no apparent correlation between insertions and deletions of modules and the indicated single-nucleotide sequence polymorphisms.

The first SRE in *D. radiodurans* was found within a cloned mitomycin C-inducible gene of strain SARK [25]. The sequenced part of the

clone contains several Shine-Dalgarno sequences and a putative start codon, located within the SRE. We compared this fragment of the SARK genome with the homologous fragment of R1 (figure 10), and found that the 5' region is considerably different. The two *D. radiodurans* ORFs being considered here are related to a gene encoding enolase in *Schizosaccharomyces pombe*. The 3' fragment of the available SARK sequence [25] is almost identical to the corresponding fragment of the R1 sequence (figure 10). A region of high similarity also exists at the far 5' terminus of the gene. The SARK fragment that intervenes in these regions of similarity cannot be aligned, ending almost exactly at the 3' terminus of the SRE (red sequence, figure 10). Since the 5' fragment of the available SARK sequence is located in the predicted coding sequence in R1, this strain is likely to contain an ancestral and intact copy of this gene. This supports the notion that a SRE was inserted into this SARK ORF and, also, that SREs are mobile. This insertion event in SARK also seems to have introduced an additional ≈ 100 bp of unknown DNA that could have been coinserted with the SRE. The apparent nonlethality of this SRE insertion can, perhaps, be explained by several Shine-Dalgarno-like sequences upstream of a potential initiation codon within the SRE; there are no frameshifts, nor in-frame stop codons introduced by this insertion.

The distribution of BIME family repeats in *E. coli* strains has been described as being consistent with their ability to transpose [3]. Although the mechanism(s) of BIME mobility is not characterized, it has been reported that BIME dispersal can be correlated with transposition of IS elements [3]. Our example of the SREs (figure 10) appears to be analogous, where a transposition event could explain the insertion. If true, this could contribute to genome instability in *D. radiodurans*, also incurred by IS transposition. Also, it suggests that the SNR elements are not likely to be evolutionarily conserved parts of a widely distributed form of gene transcription or translation regulation. The overlapping of SNRs with genes within *D. radiodurans* is

```

GATCTTGGTGGCTCGGTCAAAGAGTTTCATGAGGTCTCCAGTCGAAGGGGGGGACGGCACAGGGGATGG
::: ::::::::::: ::::::::::: ::::::::::: ::::::::::: ::::::::::: ::::::::::: ::::::::::: :::::::::::
ATGCGCCCGAGAACTTCGGTCTGCGGCTGGCGGGGCCACAGTCGCCACGATTTGGTGGCACGGTCAAAGTGTTCATGAGGTCTCCAGAACAGCTTTGGAAGCACTTCAGCATCT
m r p r t s v l r l a g p t v a t i l v a r s k c f i g s p e q a l e a l p a s

m a h g p e p l e h l p k r r n v f l a
AAGCAGTTCAGGGAGCCATGAGCTGTGGCCATGAGCGATGCGGAAACCCCTACGGCTCATGGCCACGGCCCGGAGCCTTAGAGCATTGGCAAAGACGCAACGCTTTTGGC
TTAAAGAGGCAGCCATCCGACGGCGGTGAGCAAAGGCGGACGGCGGACGGCTGCCCGGAAGGCGGG-----
l k r q p s a g g q q r r t a d g c p e g g

s g l g q l a e g e c k t r s r a e n g v m r g a g s a h h v i r r t A L A Q R
GAGCGGACTGGGACAGCTCGCAGAGGGCGAGTGCAAAACACGGAGCAGGGCGGAGATGGAGTGAATGGCGGGGTCCGGTCCCGCATCACCTCATTCGGAGACTGCTCTAGCGCAGCG
::: :::::::::::
-----GCTTTAGCGCAGCG
A L A Q R

L A S R E H G A V T E L V L D T Q Q L V V L G D A V R A A G R A G L D L
CCTTGCCTCCGGGAACACGGCGCGGTACCGAGTTGGTCCCTCGATACGCAGCAGCTGGTGTACTTGGCGATGCGGTCCGAGCGGTGGCCGAGCCGCTTGATCTG
::: :::::::::::
CCTTGCCTCCGGGAACACGGCTCGGTCCGCGAGCTGGTCTCGATGCGCAGCAGCTGGTGTACTTGGCGATACGGTCCGAGCGGTGGCCGAGCCGCTTGATCTG
L A A R E H G S V A E L V F D A Q Q L V V L G D T V R A A G R A G L D L

```

Figure 10. Mobility of SRE elements. Alignment of the 5' part of the *D. radiodurans* (SARK and R1) gene homologous to enolase from *S. pombe*. The upper nucleotide sequence is derived from strain SARK, and the lower one from strain R1. The sequence of the SRE repeat is in red. Potential Shine-Dalgarno sequences are underlined once; a putative start codon is underlined twice [25].

quite prevalent; at least 46 out of 247 SNRs overlap ORFs. In all cases, the overlapping region is in either an extreme 5'- or 3'-terminus of the affected ORFs, and it appears not to influence protein function.

2.2.5. Repeated sequences and their significance to *D. radiodurans*

Since the isolation of *D. radiodurans* in 1956 [2], there have been only about 300 publications describing this bacterium, and this limitation has restricted us to considering the repeated genomic sequences in a completely theoretical manner. It is clear, however, that this situation will likely change in the years ahead. There is increasing interest in this bacterium in many research areas including: its supremely efficient DNA repair capabilities [11], bioremediation of radioactive waste sites [23], and even astrobiology. What has promoted this obscure organism to a level where it can be studied by a broad range of biologists, are recent developments in manipulating this bacterium by ge-

netic engineering. The preeminent characteristics of *D. radiodurans* that have made this progress possible are: 1) *D. radiodurans* is naturally transformable using DNA with homology to its genome; 2) it is extremely recombination proficient; and 3) it is prolific in its ability to amplify DNA sequences that are flanked by direct repeats [28].

Upon transformation, if a gene is integrated between direct repeats of *D. radiodurans* genomic DNA, selection pressure can yield recombinant strains with about 20 duplicated copies per chromosome [28], 8–10 identical chromosomes per cell. We have used this strategy to amplify vectors as large as 20 kb, resulting in a genome that contains about 4 Mbp more DNA than wild-type [28]; these expansions are stable and readily maintained for many generations, even without subsequent selection [12]. We have exploited this organism's ability to amplify genes in both our DNA repair studies [10, 12] and in engineering *D. radiodurans* for bioremediation [23].

D. radiodurans' propensity to amplify DNA sequences is extremely pertinent to our consideration of the *D. radiodurans* genome and its myriad repeated sequences. Among the relevant features for such consideration are 52 insertion sequences (12 families) and 247 small noncoding repeats (nine families), reported here. This exceptional degree of redundancy provides the potential for genome plasticity, particularly in the context of amplifying genomic regions located between direct repeats, of which there are many examples. In the context of this organism's recombination and DNA amplification capabilities, therefore, this genomic organization could endow cells with the ability to rapidly respond and adapt to changing environmental conditions, using amplification as a form as gene regulation. While there is a large body of experimental evidence showing that genes artificially introduced into the R1 genome are readily amplified if they become flanked by direct repeats [11, 29], there is no evidence currently showing that this occurs under natural conditions.

3. Conclusions

We have described 52 IS elements (12 families) and 247 SNRs (at least 9 families) in the *D. radiodurans* genome. Our analysis shows that the IS and SNR elements are nonuniformly distributed between the four genome partitions (the chromosome, two megaplasmids and the plasmid). We have identified a single mutation in the 5'-part of the inverted repeat of IS2621 [27], that may result in error-prone excision of this element. SNRs display similar modular configurations to the BIMEs and other small repeats in *E. coli*, despite the absence of sequence similarity. Furthermore, we also have presented circumstantial evidence that *D. radiodurans* SNRs, like BIMEs, are mobile and that there may be significant differences between allelic SNRs even in closely related strains. Since there is some evidence that certain SNRs of *E. coli* are recognized by DNA binding proteins [13], it is possible that the SNRs of *D. radiodurans* have similar protein binding capa-

bilities. We believe that our findings will help guide future *D. radiodurans* experimental studies on the potential transposition mechanisms of both IS and SNR elements, and their contribution to genome functioning.

Acknowledgments

This work was funded by grants DE-FG02-98ER62583, DE-FG02-97ER62492, and DE-FG07-97ER20293 from the U.S. Department of Energy; and grant 5R01-GM39933-09 from the National Institutes of Health. We would like to thank E. Koonin, N. Grishin, M. Galperin, and M. Gelfand for their critical review of this manuscript.

References

- [1] Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.* 25 (1997) 3389–3402.
- [2] Anderson A., Nordan H., Cain R., Parrish G., Duggan D., Studies on a radioresistant micrococcus. I. Isolation, morphology, cultural characteristics, resistance to gamma radiation, *Food Technology* 10 (1956) 575–578.
- [3] Bachellier S., Clement J.M., Hofnung M., Gilson E., Bacterial interspersed mosaic elements (BIMEs) are a major source of sequence polymorphism in *Escherichia coli* intergenic regions including specific associations with a new insertion sequence, *Genetics* 145 (1997) 551–562.
- [4] Bachellier S., Gilson E., Hofnung M., Hill *Escherichia coli* and *Salmonella*: Cellular and Molecular Biology, in: Neidhardt et al., (Eds.), ASM Press, Washington DC, 1996, pp. 2013–2040.
- [5] Blaisdell B.E., Rudd K.E., Matin A., Karlin S., Significant dispersed recurrent DNA sequences in the *Escherichia coli* genome. Several new groups, *J. Mol. Biol.* 229 (1993) 833–848.
- [6] Blattner F.R., Plunkett G., Bloch C.A., Perna N.T., Burland V., Riley M., Collado-Vides J., Glasner J.D., Rode C.K., Mayhew G.F., Gregor J., Davis N.W., Kirkpatrick H.A., Goeden M.A., Rose D.J., Mau B., Shao Y., The complete genome sequence of *Escherichia coli* K-12, *Science* 277 (1997) 1453–1474.
- [7] Capy P., Langin T., Higuete D., Maurer P., Bazin C., Do the integrases of LTR-retrotransposons and class II element transposases have a common ancestor?, *Genetica* 100 (1997) 63–72.
- [8] Cole S.T., Brosch R., Parkhill J. et al., Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence, *Nature* 393 (1998) 537–544.
- [9] Dahl M.K., Francoz E., Saurin W., Boos W., Manson M.D., Hofnung M., Comparison of sequences from the *malB* regions of *Salmonella typhimurium* and *Enterobacter aerogenes* with *Escherichia coli* K12: a potential new regulatory site in the interoperonic region, *Mol. Gen. Evol.* 218 (1989) 199–207.
- [10] Daly M.J., Ouyang L., Minton K.W., In vivo damage and recA-dependent repair of plasmid and chromosomal DNA in the radioresistant bacterium *Deinococcus radiodurans*, *J. Bacteriol.* 176 (1994) 3508–3517.
- [11] Daly M.J., Minton K.W., Interchromosomal recombination in the extremely radioresistant bacterium *Deinococcus radiodurans*, *J. Bacteriol.* 176 (1995) 7506–7515.

- [12] Daly M.J., Minton K.W., An alternative pathway for recombination of chromosomal fragments precedes recA-dependent recombination in the radioresistant bacterium *Deinococcus radiodurans*, *J. Bacteriol.* 178 (1996) 4461–4471.
- [13] 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.* 26 (1997) 767–777.
- [14] Felsenstein J., Inferring phylogenies from protein sequences by parsimony, distance, likelihood methods, *Methods Enzymol.* 266 (1996) 418–427.
- [15] Frothingham R., Meeker-O'Connell W.A., Genetic diversity in the *Mycobacterium tuberculosis* complex based on variable numbers of tandem DNA repeats, *Microbiology* 144 (1998) 1189–1196.
- [16] Galas D.J., Chandler M., in: Berg D.E., Howe M.M. (Eds), *Mobile DNA*, American society of Microbiology, Washington DC, 1989, 102–162.
- [17] Galperin M.Y., Frishman D., Towards automated prediction of protein function from microbial genomic sequences, *Meth. Microbiol.* 28 (1999) 245–263.
- [18] Gilson E., Saurin W., Perrin D., Bachellier S., Hofnung M., The BIME family of bacterial highly repetitive sequences, *Res. Microbiol.* 142 (1991) 217–222.
- [19] Karlin S., Mrázek J., Campbell A.M., Frequent oligonucleotides and peptides of the *Haemophilus influenzae* genome, *Nucleic Acids Res.* 24 (1996) 4263–4272.
- [20] Kersulyte D., Akopyants N.S., Clifton S.W., Roe B.A., Berg D.E., Novel sequence organization and insertion specificity of IS605 and IS606: chimaeric transposable elements of *Helicobacter pylori*, *Gene* 223 (1998) 175–186.
- [21] Klenk H.P., Clayton R.A., Tomb J.F. et al., The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*, *Nature* 390 (1997) 364–370.
- [22] Kunst F., Ogasawara N., Moszer I. et al., The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*, *Nature* 390 (1997) 249–256.
- [23] Lange C., Wackett L.P., Minton K.W., Daly M.J., Engineering a recombinant *Deinococcus radiodurans* for organopollutant degradation in radioactive mixed waste environments, *Nature Biotech.* 16 (1998) 929–933.
- [24] Lawson F.S., Billowes F.M., Dillon J.A., Organization of carbamoyl-phosphate synthase genes in *Neisseria gonorrhoeae* includes a large, variable intergenic sequence which is also present in other *Neisseria* species, *Microbiology* 141 (1995) 1183–1191.
- [25] Lennon E., Gutman P.D., Yao H.L., Minton K.W., A highly conserved repeated chromosomal sequence in the radioresistant bacterium *Deinococcus radiodurans* SARK, *J. Bacteriol.* 173 (1991) 2137–2140.
- [26] Mathews D.H., Andre T.C., Kim J., Turner D.H., Zuker M., An updated recursive algorithm for RNA secondary structure prediction with improved free energy parameters, *Am. Chem. Soc. Symp. Ser.* 682 (1998) 246–257.
- [27] Narumi I., Cherdchu K., Kitayama S., Watanabe H., The *Deinococcus radiodurans* *uvrA* gene: identification of mutation sites in two mitomycin-sensitive strains and the first discovery of insertion sequence element from deinobacteria, *Gene* 198 (1997) 115–126.
- [28] Smith M.D., Lennon E., McNeil L.B., Minton K.W., Duplication insertion of drug resistance determinants in the radioresistant bacterium *Deinococcus radiodurans*, *J. Bacteriol.* 170 (1998) 2126–2135.
- [29] Thornley M.J., Radiation resistance among bacteria, *J. Appl. Bacteriol.* 26 (1963) 334–345.
- [30] Tomb J.-F., White O., Kerlavage A.R. et al., The complete genome sequence of the gastric pathogen *Helicobacter pylori*, *Nature* 388 (1997) 539–547.
- [31] Velasco A., Acebo P., Flores N., Perera J., The *mer* operon of the acidophilic bacterium *Thiobacillus* T3.2 diverges from its *Thiobacillus ferrooxidans* counterpart, *Extremophiles* 3 (1999) 35–43.