J Mol Evol (2005) 60:174–182
DOI: 10.1007/s00239-004-0046-3

JOURNAL OF MOLECULAR EVOLUTION

Intervening Sequences of Regularly Spaced Prokaryotic Repeats Derive from Foreign Genetic Elements

Francisco J.M. Mojica, César Díez-Villaseñor, Jesús García-Martínez, Elena Soria

División de Microbiología, Departamento de Fisiología, Genética y Microbiología, Universidad de Alicante, Campus de San Vicente, E-03080, Spain

Received: 6 February 2004 / Accepted: 1 October 2004 [*Reviewing Editor:* Dr. John Huelsenbeck]

Abstract. Prokaryotes contain short DNA repeats known as CRISPR, recognizable by the regular spacing existing between the recurring units. They represent the most widely distributed family of repeats among prokaryotic genomes, suggesting a biological function. The origin of the intervening sequences, at present unknown, could provide clues about their biological activities. Here we show that CRISPR spacers derive from preexisting sequences, either chromosomal or within transmissible genetic elements such as bacteriophages and conjugative plasmids. Remarkably, these extrachromosomal elements fail to infect the specific spacer-carrier strain, implying a relationship between CRISPR and immunity against targeted DNA. Bacteriophages and conjugative plasmids are involved in prokaryotic population control, evolution, and pathogenicity. All these biological traits could be influenced by the presence of specific spacers. CRISPR loci can be visualized as mosaics of a repeated unit, separated by sequences at some time present elsewhere in the cell.

Key words: CRISPR — DNA repeats — Conjugative plasmids — Bacteriophages — Lateral gene transfer — Prokaryotic evolution — Pathogenicity

Correspondence to: Francisco J.M. Mojica; *email:* fmojica@ua.es

Introduction

Prokaryotic genomes contain a peculiar family of repeated DNA sequences. They consist of 24- to 40-nucleotide (nt) recurrent motifs regularly spaced by intervening sequences of sizes similar to that of the repeated unit. These repetitive elements were defined as short regularly spaced repeats (SRSR) (Mojica et al. 2000) and more recently named CRISPR (clustered regularly interspaced short palindromic repeats) (Jansen et al. 2002). CRISPR are widespread among the various physiological and phylogenetic groups of prokaryotes including *Archaea* (both *Crenarchaeota* and *Euryarchaeota*) and lineages of Gram-negative and Gram-positive bacteria (Jansen et al. 2002; Mojica et al. 2000). Thus they represent the most widely distributed family of repeats among prokaryotic genomes. A biological function is predicted by the broad distribution and the remarkable structural conservation of CRISPR, but this has not been firmly established. The only experimental insight reported in this sense relates CRISPR to partitioning on the basis of incompatibility between replicons containing loci of these repeats (Mojica et al. 1995). The origin of the intervening sequences, at present also unknown, could provide the clue to determine the CRISPR function. With this aim, we have carried out a systematic search for spacer identities, finding significant similarities to a variety of DNA molecules. The highest identities are with genetic elements, including chromosomes, bacteriophages, and conjugative plasmids, of strains closely related to the one containing the spacer. Interestingly, these targeted

viruses are unable to infect the spacer-carrier cell but succeed with closely related strains lacking the specific CRISPR spacer. Likewise, plasmids efficiently transferred among various species in the same phylogenetic group cannot be stably maintained in members with a CRISPR spacer matching a sequence in the replicon. The relationship between CRISPR and immunity against targeted foreign DNA is discussed in relation to its functional and evolutionary significance.

Materials and Methods

PCR and Nucleotide Sequencing

PCR reactions for *E. coli* CRISPR loci amplification were performed under standard conditions with recombinant *Taq* polymerase from Invitrogen (Carlsbad, CA), using the oligonucleotide primers 5′TGGTGAAGGAGTTGGCGAAGG3′ and 5′AAAAT GTCCCTCCGCGCTTACG3′ or 5′CGATCCAGAGCTGGTCG AATG3′ and 5′CGCTGACCGATGATAAAC3′. PCR products were purified with the QIAquick PCR purification kit (QiaGen, Valencia, CA) and sequenced with the Big Dye Terminator Cycle sequencing kit in an ABI PRISM 310 DNA Sequencer following the manufacturer instructions (Applied Biosystems, Foster City, CA). The sequence data for CRISPR regions of the *E. coli* strains ECOR42, ECOR44, ECOR47, and ECOR49 have been submitted to the GeneBank database under accession numbers AY490777, AY490778, AY490779, and AY490780, respectively.

Sequence Analyses

CRISPR loci of available prokaryotic genomes were detected with a specifically designed computer program (Mojica et al. 2000). Similarities to the spacers were searched for in the GeneBank nucleotide sequence database using the BLASTn program (Altschul et al. 1997) at the NCBI Web site (www.ncbi.nlm.nih.gov/BLAST/). The default parameters were used, but for a word size of 7. Given the short length of the spacers (<50 nt) and the considerably large database (over 10^{10} nt), the significance of the aligments was empirically determined by an iterative process. First, only identical matches were considered, which gave Expect values (E-values) from 0.02 (20-nt sequence) to 10^{-18} (for the largest spacers). According to this and to the abrupt discontinuity usually found between sequences related and unrelated to the identical one, the cutoff for the E-value in subsequent searches was rigorously set at 0.02. Although searches were done against all organisms, alignments with values below the threshold corresponded to sequences related to the organism hosting the query (with identities over 80%), and the first match after the discontinuity mainly to eukaryotic sequences, validating this cutoff value. Thus, eventually, sequence identity, E-value, and relatedness were used together as criteria in each search to confirm positives and recognize false negatives. Sequences fitting these established criteria are designated CRISPR-spacer homologs in the text.

Results

We have searched for identities to about 4500 CRISPR spacers from 67 strains representing 36 genera of prokaryotes (Table 1). Significant similarities to known sequences (Fig. 1; see Materials and Methods for criteria) were found for 88 spacers from 4 strains of *Archaea*, 12 strains of Gram-negative bacteria, and 9 strains of Gram-positive bacteria (Table 2). Interestingly, 47 spacers of the 88 matched sequences within genes corresponding to bacteriophages, 10 within plasmidic DNA, and 31 within chromosomal DNA not directly related to foreign genetic elements.

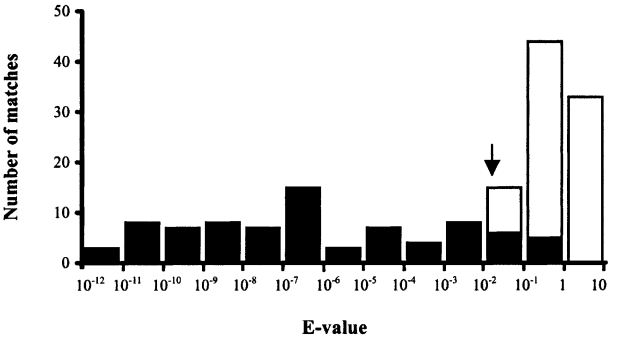

Fig. 1. Distribution of E-values corresponding to validated similarities (solid bars) and best-score discarded alignments for the positive searches (open bars). Cutoff significance value is indicated by an arrow. See Materials and Methods for details.

The spacers from the crenarchaea *Sulfolobus*, the euryarchaea *Methanothermobacter*, the Gram-positive bacteria *Streptococcus pyogenes*, and the Gram-negative bacteria *Escherichia coli* were selected as representatives of the four prokaryotic groups for further analysis. Given the substantial differences and evolutionary distance between these microorganisms, the conclusions drawn from this study span the prokaryotes.

Analysis of Sulfolobus *CRISPR Spacers*

The members of the genus *Sulfolobus* are sulfur metabolizing aerobic crenarchaea, growing optimally at 80°C and pH 2–4.

Sulfolobus solfataricus P2 has about 400 CRISPR spacers (She et al. 2001), 9 of which showed similarity to known sequences (Table 3), consistently within extrachromosomal genetic elements of *Sulfolobus*, either SIRV viruses (Prangishvili et al. 1999) or the conjugative plasmid pNOB8 (Schleper et al. 1995). It is noteworthy that, among the viral genes, only ORF121 (a putative resolvase [Birkenbihl et al. 2001]) has a homolog (*hje*) in the *S. solfataricus* P2 chromosome, although with a much lower similarity to the spacer sequence (13 instead of 32 identities in 37 nt). The plasmid-borne genes containing the most similar sequences to spacers are involved in transposition (ORF406; homologous to transposases [Zillig et al. 1998]), replicon partitioning (ORF315; homologous to *parA* [Easter et al. 1997]), or plasmid

Table 1. List of strains analyzed

Strain	Phylogenetic group
Aeropyrum pernix K1	Crenarchaeota
Aquifex aeolicus VF5	Aquificales
Archaeoglobus fulgidus DSM-4304	Euryarchaeota
Bacillus cereus ATCC-14579	Gram-positive bacteria
Bacillus halodurans C-125	Gram-positive bacteria
Calothrix sp. D253	Cyanobacteria
Campylobacter jejuni subsp. *jejuni* NCTC11168	Epsilon-proteobacteria
Chlorobium tepidum TLS	CFB/green-sulfur bacteria
Clostridium difficile 630	Gram-positive bacteria
Clostridium tetani Massachusetts E88	Gram-positive bacteria
Corynebacterium diphtheriae gravis NCTC13129	Gram-positive bacteria
Corynebacterium efficiens YS-314T	Gram-positive bacteria
Escherichia coli UPEC-CFT073	Gamma-proteobacteria
Escherichia coli O157:H7 Sakai	Gamma-proteobacteria
Escherichia coli O157:H7 EDL933	Gamma-proteobacteria
Escherichia coli K12-MG1655	Gamma-proteobacteria
Escherichia coli ECOR42	Gamma-proteobacteria
Escherichia coli ECOR44	Gamma-proteobacteria
Escherichia coli ECOR47	Gamma-proteobacteria
Escherichia coli ECOR49	Gamma-proteobacteria
Geobacillus stearothermophilus 10	Gram-positive bacteria
Geobacter sulfurreducens PCA	Delta-proteobacteria
Haloferax mediterranei ATCC-33500	Euryarchaeota
Haloferax volcanii DS2	Euryarchaeota
Listeria monocytogenes EGD-e	Gram-positive bacteria
Listeria innocua Clip 11262	Gram-positive bacteria
Methanothermobacter thermoautotrophicum ΔH	Euryarchaeota
Methanococcus jannaschii DSM-2661	Euryarchaeota
Methanopyrus kandleri AV19	Euryarchaeota
Mycoplasma gallisepticum R	Gram-positive bacteria
Nanoarchaeum equitans Kin4-M	Nanoarchaeota
Neisseria meningitidis Z2491 (serogroup A)	Beta-proteobacteria
Nitrosomonas europaea ATCC-19718	Beta-proteobacteria
Nostoc sp. PCC 7120	Cyanobacteria
Pasteurella multocida Pm70	Gamma-proteobacteria
Photorhabdus luminescens laumondii TT01	Gamma-proteobacteria
Porphyromonas gingivalis W83	CFB/green-sulfur bacteria
Pyrobaculum aerophilum IM2	Crenarchaeota
Pyrococcus furiosus DSM3638	Euryarchaeota
Pyrococcus abyssi GE5	Euryarchaeota
Pyrococcus horikoshii (shinkaj) OT3	Euryarchaeota
Salmonella bongori 12149	Gamma-proteobacteria
Salmonella enterica Typhi Ty2	Gamma-proteobacteria
Salmonella enterica Typhi CT18	Gamma-proteobacteria
Salmonella enterica subsp. *enterica serovar* Dublin	Gamma-proteobacteria
Salmonella enteritidis PT4	Gamma-proteobacteria
Salmonella paratyphi A	Gamma-proteobacteria
Salmonella typhimurium LT2 SGSC1412	Gamma-proteobacteria
Salmonella typhimurium DT104	Gamma-proteobacteria
Salmonella typhimurium SL1344	Gamma-proteobacteria
Shigella flexneri 2a301	Gamma-proteobacteria
Shigella flexneri 2a2457T	Gamma-proteobacteria
Shigella sonnei 53G	Gamma-proteobacteria
Streptococcus agalactiae NEM316	Gram-positive bacteria
Streptococcus agalactiae 2603V/R	Gram-positive bacteria
Streptococcus mutans UA159	Gram-positive bacteria
Streptococcus pyogenes M1 GAS SF370	Gram-positive bacteria
Sulfolobus solfataricus P2	Crenarchaeota
Sulfolobus tokodaii 7	Crenarchaeota
Thermoanaerobacter tengcongensis MB4T	Gram-positive bacteria
Thermoplasma acidophilum DSM1728	Euryarchaeota
Thermoplasma volcanium GSS1	Euryarchaeota
Thermotoga maritima MSB8	Thermotogales
Thermus thermophilus HB8	Thermus/Deinococcus
Vibrio vulnificus YJ016	Gamma-proteobacteria

Table 2. Distribution of CRISPR-spacer homologs

Strain	No. of spacers analyzed	No. of spacers with homologs in		
		Phages[a]	Plasmids	NF[b]
Chlorobium tepidum TLS	62		1	
Clostridium tetani Massachusetts E88	62	1		6
Corynebacterium efficiens YS-314T	22		1	2
Escherichia coli ECOR42	14		1	
Escherichia coli ECOR44	10	1		
Escherichia coli ECOR47	17	1		
Escherichia coli ECOR49	11		1	
Listeria innocua Clip11262	9	3		
Listeria monocytogenes EGD-e	4	1		
Methanothermobacter thermoautotrophicum ΔH	169	9		
Mycoplasma gallisepticum R	71			1
Neisseria meningitidis Z2491 (serogroup A)	16			4
Photorhabdus luminescens laumondii TT01	65	7		3
Porphyromonas gingivalis W83	44			4
Pyrobaculum aerophilum IM2	129			1
Salmonella typhimurium LT2 SGSC1412	57	1		
Shigella sonnei 53G	3			1
Streptococcus agalactiae NEM316	13	1		1
Streptococcus agalactiae 2603V/R	25	1	1	3
Streptococcus pyogenes MI GAS SF370	9	8		
Sulfolobus solfataricus P2	424	6	3	
Sulfolobus tokodaii 7	471	2	2	
Thermoanaerobacter tengcongensis MB4T	306			5
Yersinia pestis CO-92 (Biovar Orientalis)	16	4		
Yersinia pestis KIM5P12 (Biovar Mediaevalis)	10	1		

[a]Prophages are included.
[b]Number of spacers with homology to chromosomal sequences not directly related to foreign DNA (prophages are excluded).

transfer (ORF1025; homologous to *traG* [Cabezón et al. 1997; Firth and Skurray 1992]). Only ORF406 of pNOB8 has homologous genes within the *S. solfataricus* chromosome, and like *hje* (see above), the similarity to the spacer diminishes from 30 to fewer than 21 identities in 37 nt.

Sulfolobus tokodaii strain7 has about 450 CRISPR spacers (Kawarabayasi et al. 2001). Four of them showed significant similarity to known sequences, invariably within uncharacterized ORFs located in genetic elements of *Sulfolobus*, either viruses (SIRV1 and SSV1 [Prangishvili et al. 1999; Stedman et al. 2003]) or conjugative plasmids (pNOB8 and pING1 [She et al. 1998; Stedman et al. 2000; Zillig et al. 1998]) (see Table 3). No homolog to these genes was detected in the *S. tokodaii* genome.

Analysis of Methanothermobacter thermoautotrophicum *CRISPR Spacers*

Methanothermobacter (formerly *Methanobacterium*) *thermoautotrophicum* is a lithoautotrophic, thermophilic euryarchaeon that grows optimally at 65°C. *M. thermoautotrophicum* ΔH has 169 CRISPR spacers (Mojica et al. 2000), 9 of which showed similarity to sequences in the databases (Table 4), all of them within phages of *Methanothermobacter*, either the *M. wolfeii* prophage ΨM100 (four spacers [Luo et al. 2001]) or the *M. marburgensis* phage ΨM2 (six spacers including one duplicated [Pfister et al. 1998]).

Among the genes containing sequences similar to spacers, only *peiP* (a lytic enzyme [Luo et al. 2001]) has a homolog (MTH412) in the *M. thermoautotrophicum* ΔH chromosome. As in the *Sulfolobus* cases (see above), the spacer sequence is greatly degenerated (13 instead of 37 identities in 37 nt).

Analysis of Streptococcus pyogenes *CRISPR Spacers*

Streptococcus pyogenes is a strict human pathogen (Cunningham 2000) member of the low-G+C group of Gram-positive bacteria. Among the four at present fully sequenced *S. pyogenes* strains, we have only detected CRISPR loci in SF370 (Ferretti et al. 2001), which contains two clusters of repeats, one of them with four CRISPR units (Cluster4), and the other (Cluster7) with seven repeats (Jansen et al. 2002). The three spacers of Cluster4 (here named 4-1, 4-2, and 4-3) and five of six spacers of Cluster7 (here named 7-1, 7-2, 7-3, 7-4, and 7-5) showed significant similarity to known sequences (Table 5). The greatest similarities were found within pro-

Table 3. Features of the sequences most similar to CRISPR spacers from the genus *Sulfolobus*

Strain	ORF	Replicon	Activity	Alignment[a]
S. solfataricus P2	ORF406	pNOB8	Transposase	tgaatagcaacatcgtgtaacctcatcctcagccttc taaaaggcaacatcgtgcaacctcatcctcat-cttc
	ORF1025	pNOB8	NTPase	ttgtctgtcggtgaagcagtagtatttctaagaggccgtcc ctatctgtcggtgaagccgtagcatttctaagaggccgtcc
	ORF315	pNOB8	Resolvase	cctaatatcctcgggtacttatagaaccctccttctggtc cctaatattctaggatacttatagaaccctccttctggtc
	ORF121	SIRV1	Resolvase	aaagcggttgttttccagttccagaaactggaattcttat gaagtgcctgttttccagttccagaaactggaattcttaa
	ORF510	SIRV1	Unknown	atgttctttttccagaactgtaactataattttgttgatat atgttctttttccagatgtgtaactataattttgttgatat
	ORF134	SIRV1	Structural	tggtaaatagctctgttaggcccagttattccatattctg tgatatatagctctgttaggcccagttattccatattgtg
	ORF356	SIRV1	Glycosyl transferase	atcatttatgcacatttcaactccatttccaatatgaat ttttcctatacacatttgaactccatttccaatatgaat
	ORF98	SIRV1	Unknown	aagataccacaacacttggagttaatgcattattgaatatggatacat aagatactacagtaactggaataaatgcattattgaatatggatactt
	ORF268	SIRV1	Unknown	cgtaagaataatttttacaaaatcttagtaattatatagtttatatc agttagaataatttttgcaaaatttaagtaattatattgtatatatc
S. tokodaii strain7	ORF F-92	SSV1	Unknown	aaaatgagcttgaaaaaatcgaaaacgaactcaggttaatgag aaaaggagcttgaaaaaatcgaaaacgaactcaggttaatgag
	ORF436	SIRV1	Unknown	aacttatgcaaatctcatctatgacgcgagaaataatatgta aacttatgcaaatttgatctatgatgctagaaataatatgta
	ORF94	pING1	Unknown	ctacagaaagctaagatgataagggcctaccttattg ctgcagaaagctaaaatgataagggcctatcttattg
	ORF246	pNOB8	Unknown	atcatcctcctcattgtcgatactccctttgtgaattt atcattctcttcattgttgatactcccttagtgaattt

[a]CRISPR-spacer sequence (top line) and best-match homologous sequence (bottom line).

phages present in the CRISPR-negative *S. pyogenes* strains SSI-1 (Nakagawa et al. 2003), MGAS8232 (Smoot et al. 2002), and MGAS315 (Beres et al. 2002). Those genes containing sequences identical to spacers 7-1 (*spyM3_1215* of prophage 315.4, *spyM3_0930* of prophage 315.2, and *spyM3_119* of prophage 315.6, all in MGAS315, and homologous genes in the other strains) and 7-4 (*hylp* within prophage 315.3 in MGAS315) are related to cell lysis. Genes with homologs to spacers 4-3 and 7-3 are involved in DNA modification, probably protecting the phage against host restriction. Gene *spyM3_0941* of prophage 315.2 in MGAS315 and homologs in SSI-1 and MGAS8323, all of them containing a sequence similar to spacer 4-2 (28 identities in 31 nt), encode a capside structural protein (Beres et al. 2002). The gene with similarity to spacer 7-2 encodes the exotoxin SpeM, which acts as a superantigen (Profit et al. 2003; Smoot et al. 2002). It is noteworthy that the *Streptococcus* prophages containing sequences similar to CRISPR spacers are usually absent from the SF370 genome. Moreover, while the homologous sequence is highly conserved in CRISPR-negative *Streptococcus* strains, in the exceptional cases where a prophage homolog is inserted into the SF370 genome, either the spacer-homologous gene is absent (spacer 4-1) or the corresponding sequence similar to the spacer is degenerated (spacers 7-1 and 7-4).

Analysis of Escherichia coli *CRISPR Spacers*

E. coli is an extensively studied Gram-negative bacteria that includes both saprophytic and pathogenic strains. No homology to CRISPR spacers in the genomes of the *E. coli* strains UPEC-CFT073, O157:H7 Sakai, 0157:H7-EDL933, and K12-MG1655 was detected in this work. We have sequenced additional CRISPR loci from strains of the

Table 4. Features of the sequences most similar to CRISPR spacers from *Methanothermobacter thermoautotrophicum*

Location	Phage	Activity	Alignment[a]
40 bp 3′ from ORF31	ΨM100	Not applicable	cttctagcaagagacattgacgatatacacaagtac |||||||||||||||||||||||||||||||||| cttctagcaagagacattgacgatatacacaagtac
ORF31	ΨM100	Unknown	aagcgccgggcagacagcacacatacaagacttcacaa |||||||||||||||||||||||||||||||||| aagcgccgggcagacagcacacatacaagacttcacaa
ORF31	ΨM100	Unknown	tttcacgatgactctgttgagttcatcgtattctttcc |||||||||||||||||||||||||||||||||| tttcacgatgactctgttgagttcatcgtattcttccc
ORF21	ΨM100	Tail protein	tgatgttgggaaggtttggccatctgaatgatttga |||||||||||||||||||||||||||||||||| tgatgttgggaaggtttggccatctgaatgatttga
peiP	ΨM2	Pseudomurein endoisopeptidase	aatattgaaacgttcaaggacatgttgaagaggtatg |||||||||||||||||||||||||||||||||| aatattgaaacgttcaaggacatgttgaagaggtatg
ORF6	ΨM2	Unknown	agtatgtgcagtatcctctctatgtccccttcattc |||||||||||||||||||||||||||||||||| agtatgtgcagtatcctctctatgtccccttcattc
ORF6	ΨM2	Unknown	aacttcacagaaaagcctccatggagcaagtgctct |||||||||||||||||||||||||||||||||| aacttcacagaaaagcctccatggagcaggtgctcc
103 bp 3′ from ORF6	ΨM2	Not applicable	gattttgacggtgagtacaatctctgctgtcagaactg |||||||||||||||||||||||||||||||||| gattttgacgctgagtacaatctctgctgtcagaactg
ORF17	ΨM2	Unknown	ccggtccttgacgggaaaatctacagggccacaatag |||||||||||||||||||||||||||||||||| ccggtccttgatgggaaaatctacagggccacaatag

[a]CRISPR-spacer sequence (top line) and best-match homologous sequence (bottom line).

Table 5. Features of the sequences most similar to CRISPR spacers from *S. pyogenes*

Spacer	Gene	Prophage[a]	Activity	Alignment[b]
4-1	*spyM3_1239*	315.4	Unknown	gctgtgacattgcgggatgtaatcaaagtaaaaa ||||||||||||||||||||||||||||||||| gctgtgacattgcggaatgtaatcaaagcaaaaa
4-2	*spyM3_0941*	315.2	Capside protein	taaagcaaacctagcagaagcagaaaatgac |||||||||||||||||||||||||||||||||| taaagcgaacctagtagaagcagaaaacgac
4-3	*spyM18_0741*	Φ_{speC}	Methyltransferase	ctgatgtaattggtgattttcgtgatatgcttt |||||||||||||||||||||||||||||||||| ctgatgtaattggtgattttcgtgatatgcctt
7-1	*spyM3_1215*	315.4	Endopeptidase	gcgctggttgatttcttcttgcgcttttt |||||||||||||||||||||||||||||||||| gcgctggttgatttcttcttgcgcttttt
7-2	*speM*	Φ_{speLM}	Exotoxin	tatatgaacataactcaatttgtaaaaaa |||||||||||||||||||||||||||||||||| tatatgaacataactcaatttgtaaaaaa
7-3	*spyM18_0742*	Φ_{speC}	Methyltransferase	aggaatatccgcaataattaattgcgctct |||||||||||||||||||||||||||||||||| aggaatatccgcaataattaattgcgctct
7-4	*hylP*	315.3	Hyaluronidase	agtgccgaggaaaaattaggtgcgcttggc |||||||||||||||||||||||||||||||||| agtgccgaggaaaaattaggtgcgcttggc
7-5	*spyM3_1347*	315.5	Unknown	aaatttgtttagcaggtaaaccgtgcttt |||||||||||||||||||||||||||||||||| aaatttgtttagcaggtaaaccgtgcttt

[a]Prophages 315.2-5 are integrated into *S. pyogenes* MGAS315. Φ_{speC} and Φ_{speLM} are integrated into *S. pyogenes* MGAS8232.
[b]CRISPR-spacer sequence (top line) and best-match homologous sequence (bottom line).

E. coli reference (ECOR) collection (Ochman and Selander 1984) and found similarities ranging from 28 identities in 32 nt to 100% identity for four spacers (Table 6). Two of them have the highest similarity within genes of enterobacteria conjugative plasmids, and the other two within enterobacteria phages. The plasmid-borne genes involved are *resD* (ECOR49 spacer), related to plasmid replication and

Table 6. Features of the sequences most similar to CRISPR spacers from *E. coli*

Strain	Gene	Element	Activity	Alignment[a]
ECOR42	*traI*	Plasmid F	Helicase	gtttcccgtgcgtcgtatgaggcagaaaagag |||||||||||||||||||||||||||| gtttcccgtgcgtcgtatgaggcagaaaagag
ECOR44	Unannotated	Phage P1	Unknown	ctgttggcaagccaggatctgaacaataccgt ||||||||| ||||||||||||||||| ctgttggcaggccaggatctgaacaataccgt
ECOR47	*darB*	Phage P1	Methylase	gctggtggcgcgggcaaacggaacaatcccgc |||||||||||||| |||||||||||||| gctggtggcgcgggcgaacggaacaatcccgc
ECOR49	*resD*	Plasmid F	Resolvase	atcgacttatgccccatcaggctctgcaatac || |||||||||||||||||| |||| || atggacttatgtcccatcaggctttgcagaac

[a]CRISPR-spacer sequence (top line) and best-match homologous sequence (bottom line).

resolution (Lane et al. 1986), and a *traI* homolog (ECOR42 spacer), which participates in plasmid transfer (Traxler and Minkley 1988). One of the viral genes is located close to the replication origin of phage P1 (ECOR44 spacer) (Martin et al. 1991), and another is the *darB* gene of the same phage (ECOR47 spacer), which encodes for a DNA methylase involved in P1 protection against host restriction (Iida et al. 1987).

Discussion

In this paper we report the origin of the CRISPR intervening sequences, demonstrating that such spacers are not unique, as previously considered (Jansen et al. 2002; Mojica et al. 2000), but derive from preexisting sequences. The highest similarities to a given spacer are found within genetic elements of strains closely related to that carrying the spacer, reinforcing the conclusion that the CRISPR intervening sequences originate from such elements. About 65% of the spacer homologs encountered correspond to bacteriophages or conjugative plasmids, and the remaining 35% to chromosomal sequences not directly related to foreign DNA. This proportion of chromosomal DNA is most likely overestimated because of the limited accessory element sequences in the databases (Canchaya et al. 2003). The fact that only 88 spacers of about 4500 probed show significant identities could be interpreted as indicating a vast amount of uncharacterized genetic elements in nature. The case of *S. pyogenes* SF370 is exceptional. Eight of nine spacers showed over 90% identity to sequences within prophages integrated in other strains, suggesting a high sequence conservation of the corresponding genes among *S. pyogenes* phages and, at the same time, evoking a functional relevance for the genes targeted by spacers.

The incorporation of a given spacer in a CRISPR locus must not be devoid of functional significance. Although our current knowledge remains limited, multiple observations suggest that CRISPR could be involved in conferring specific immunity against foreign DNA: (i) SIRV viruses are unable to infect *S. solfataricus* (with CRISPR spacers similar to this phage sequences), although they can penetrate the cell (She et al. 2001), and while a SSV-type element is integrated into the *S. solfataricus* genome (which has no CRISPR spacer similar to SSV sequences), SSV-related prophages are absent from *S. tokodaii*; (ii) *M. thermoautotrophicum* ΔH, with CRISPR spacers similar to ΨM100, has two putative attachment sites of this phage, but there is no recognizable prophage in its genome (Smith et al. 1997); (iii) in the exceptional cases where a prophage homologous to those that contain sequences matching a given spacer is present in the carrier strain, the particular sequence used to be either absent or degenerated; (iv) pNOB8 can be transferred into *S. solfataricus*, but this plasmid is not integrated, even though there is a perfectly conserved integration site (She et al. 2002), neither stably maintained as a replicon (Schleper et al. 1995); and (v) although there is evidence that a pNOB8-like plasmid was once integrated into the *S. tokodaii* strain7 genome, only a few sequences homologous to pNOB8 genes remain scattered in the chromosome (Kawarabayasi et al. 2001). Indeed, the preferential occurrence of CRISPR spacers derived from genetic elements that fail to infect the corresponding spacer-carrier strain, but not from those successfully propagated in the population, strongly suggests a relationship between CRISPR and such immunity. Most targeted genes detected in this study are directly involved in plasmid transference, DNA replication, virion assembly, DNA protection against restriction, replicon partitioning, pili synthesis, replicons resolution, transposition, or phage integration and excision. Inhibition of any of these genes would be enough to hinder infectivity. The incompatibility with foreign DNA could be explained by the inhibition of any of these gene functions, necessary for the efficient transfer or infection. The transcription of the

CRISPR loci (Tang et al. 2002) suggests that such activity could be executed by CRISPR-RNA molecules, acting as regulatory RNA that specifically recognizes the target through the homologous RNA-spacer sequence, similarly to the eukaryotic interference RNA.

The susceptibility to foreign genetic elements has multiple consequences related to the pathogenic potential and evolution of the prokaryotes: (i) bacteriophages are involved in prokaryotic population control by inducing cell lysis (Young et al. 1992), (ii) antibiotic resistance is frequently transmitted by conjugative plasmids (Martínez and Baquero 2002), (iii) both bacteriophages and conjugative plasmids are vectors for lateral gene transfer (Boucher et al. 2003), and (iv) many virulence factors from bacteria are bacteriophage encoded (Boyd and Brussow 2002). Indeed, prophages account for most of the genetic variation among closely related strains and have greatly contributed to their evolution (Banks et al. 2002; Glaser et al. 2001).

Control of both mobility and maintenance of extrachromosomal genetic elements could be a relevant task of CRISPR loci, but the homologies found with chromosomal sequences unrelated to foreign DNA indicate that CRISPR may be constituents of a more versatile regulatory system. Experimental approaches have to be carried out to corroborate this functionality.

Acknowledgments. This work was financed by research grants from the Conselleria de Cultura, Educació i Ciència, Generalitat Valenciana (CTIDIB/2002/155 and gv04B-457). C.D. is supported by a graduate fellowship from the Ministerio de Educación, Cultura y Deporte. We are indebted to Kathy Hernández for assistance and to J. Antón for critical reading of the manuscript.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Banks DJ, Beres SB, Musser JM (2002) The fundamental contribution of phages to GAS evolution, genome diversification and strain emergence. Trends Microbiol 10:515–521

Beres SB, Sylva GL, Barbian KD, Lei BF, Hoff JS, Mammarella ND, Liu MY, Smoot JC, Porcella SF, Parkins LD, Campbell DS, Smith TM, McCormick JK, Leung DYM, Schlievert PM, Musser JM (2002) Genome sequence of a serotype M3 strain of group A *Streptococcus*: Phage-encoded toxins, the high-virulence phenotype, and clone emergence. Proc Natl Acad Sci USA 99:10078–10083

Birkenbihl RP, Neef K, Prangishvili D, Kemper B (2001) Holliday junction resolving enzymes of archaeal viruses SIRV1 and SIRV2. J Mol Biol 309:1067–1076

Boucher Y, Douady CJ, Papke RT, Walsh DA, Boudreau ME, Nesbo CL, Case RJ, Doolittle WF (2003) Lateral gene transfer and the origins of prokaryotic groups. Annu Rev Genet 37:283–328

Boyd EF, Brussow H (2002) Common themes among bacteriophage-encoded virulence factors and diversity among the bacteriophages involved. Trends Microbiol 10:521–529

Cabezón E, Sastre JI, de la Cruz F (1997) Genetic evidence of a coupling role for the TraG protein family in bacterial conjugation. Mol Gen Genet 254:400–406

Canchaya C, Proux C, Fournous G, Bruttin A, Brüssow H (2003) Prophage genomics. Microbiol Mol Biol Rev 67::38–276

Cunningham MW (2000) Pathogenesis of group A streptococcal infections. Clin Microbiol Rev 13:470–511

Easter CL, Sobecky PA, Helinski DR (1997) Contribution of different segments of the par region to stable maintenance of the broad-host-range plasmid RK2. J Bacteriol 179:6472–6479

Ferretti JJ, McShan WM, Ajdic D, et al. (2001) Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. Proc Natl Acad Sci USA 98:4658–4663

Firth N, Skurray R (1992) Characterization of the F plasmid bifunctional conjugation gene *traG*. Mol Gen Genet 232:145–153

Glaser P, Frangeul L, Buchrieser C, et al. (2001) Comparative genomics of *Listeria* species. Science 294:849–852

Iida S, Streiff MB, Bickle TA, Arber W (1987) Two DNA antirestriction systems of bacteriophage P1, darA, and darB: Characterization of darA− phages. Virology 157:156–166

Jansen R, Embden JD, Gaastra W, Schouls LM (2002) Identification of genes that are associated with DNA repeats in prokaryotes. Mol Microbiol 43:1565–1575

Kawarabayasi Y, Hino Y, Horikawa H, et al. (2001) Complete genome sequence of an aerobic thermoacidophilic crenarchaeon, *Sulfolobus tokodaii* strain7. DNA Res 8:123–140

Lane D, de Feyter R, Kennedy M, Phua SH, Semon D (1986) D protein of miniF plasmid acts as a repressor of transcription and as a site-specific resolvase. Nucleic Acids Res 14:9713–9728

Luo YN, Pfister P, Leisinger T, Wasserfallen A (2001) The genome of archaeal prophage Ψ M100 encodes the lytic enzyme responsible for autolysis of *Methanothermobacter wolfeii*. J Bacteriol 183:5788–5792

Martin KA, Davis MA, Austin S (1991) Fine-structure analysis of the P1 plasmid partition site. J Bacteriol 173:3630–3634

Martínez JL, Baquero F (2002) Interactions among strategies associated with bacterial infection: Pathogenicity, epidemicity, and antibiotic resistance. Clin Microbiol Rev 15:647–679

Mojica FJM, Ferrer C, Juez G, Rodríguez-Valera F (1995) Long stretches of short tandem repeats are present in the largest replicons of the Archaea *Haloferax mediterranei* and *Haloferax volcanii* and could be involved in replicon partitioning. Mol Microbiol 17:85–93

Mojica FJM, Díez-Villaseñor C, Soria E, Juez G (2000) Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. Mol Microbiol 36:244–246

Nakagawa I, Kurokawa K, Yamashita A, Nakata M, Tomiyasu Y, Okahashi N, Kawabata S, Yamazaki K, Shiba T, Yasunaga T, Hayashi H, Hattori M, Hamada S (2003) Genome sequence of an M3 strain of *Streptococcus pyogenes* reveals a large-scale genomic rearrangement in invasive strains and new insights into phage evolution. Genome Res 13:1042–1055

Ochman H, Selander RK (1984) Standard reference strains of *Escherichia coli* from natural populations. J Bacteriol 157:690–693

Pfister P, Wesserfallen A, Stettler R, Leisinger T (1998) Molecular analysis of *Methanobacterium* phage ΨM2. Mol Microbiol 30:233–244

Prangishvili D, Arnold HP, Gotz D, Ziese U, Holz I, Kristjansson JK, Zillig W (1999) A novel virus family, the Rudiviridae: Structure, virus-host interactions and genome variability of

the *Sulfolobus* viruses SIRV1 and SIRV2. Genetics 152:1387–1396

Profit T, Sriskandan S, Yang L, Fraser JD (2003) Superantigens and streptococcal toxic shock syndrome. Emerg Infect Dis 9:1211–1218

Schleper C, Holz I, Janekovic D, Murphy J, Zillig W (1995) A multicopy plasmid of the extremely thermophilic archaeon *Sulfolobus* effects its transfer to recipients by mating. J Bacteriol 177:4417–4426

She Q, Phan H, Garrett RA, Albers SV, Stedman KM, Zillig W (1998) Genetic profile of pNOB8 from *Sulfolobus*: The first conjugative plasmid from an archaeon. Extremophiles 2:417–425

She Q, Singh RK, Confalonieri F, et al. (2001) The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2. Proc Natl Acad Sci USA 98:7835–7840

She Q, Brugger K, Chen L (2002) Archaeal integrative genetic elements and their impact on genome evolution. Res Microbiol 153:325–332

Smith DR, DoucetteStamm LA, Deloughery C, et al. (1997) Complete genome sequence of *Methanobacterium thermoautotrophicum* ΔH: Functional analysis and comparative genomics. J Bacteriol 179:7135–7155

Smoot JC, Barbian KD, Van Compel JJ, et al. (2002) Genome sequence and comparative microarray analysis of serotype M18 group A *Streptococcus* strains associated with acute rheumatic fever outbreaks. Proc Natl Acad Sci USA 99:4668–4673

Smoot LM, McCormick JK, Smoot JC, Hoe NP, Strickland I, Cole RL, Barbian KD, Earhart CA, Ohlendorf DH, Veasy LG, Hill HR, Leung DYM, Schlievert PM, Musser JM (2002) Characterization of two novel pyrogenic toxin superantigens made by an acute rheumatic fever clone of *Streptococcus pyogenes* associated with multiple disease outbreaks. Infect Immun 70:7095–7104

Stedman KM, She Q, Phan H, Holz I, Singh H, Prangishvili D, Garrett R, Zillig W (2000) pING family of conjugative plasmids from the extremely thermophilic archaeon *Sulfolobus islandicus*: Insights into recombination and conjugation in Crenarchaeota. J Bacteriol 182:7014–7020

Stedman KM, She Q, Phan H, Arnold HP, Holz I, Garrett RA, Zillig W (2003) Relationships between fuselloviruses infecting the extremely thermophilic archaeon *Sulfolobus*: SSV1 and SSV2. Res Microbiol 154:295–302

Tang TH, Bachellerie JP, Rozhdestvensky T, Bortolin ML, Huber H, Drungowski M, Elge T, Brosius J, Huttenhofer A (2002) Identification of 86 candidates for small non-messenger RNAs from the archaeon *Archaeoglobus fulgidus*. Proc Natl Acad Sci USA 99:7536–7541

Traxler BA, Minkley EG Jr (1988) Evidence that DNA helicase I and oriT site-specific nicking are both functions of the F TraI protein. J Mol Biol 5:205–209

Young R (1992) Bacteriophage lysis: Mechanism and regulation. Microbiol Rev 56:430–481

Zillig W, Arnold HP, Holz I, Prangishvili D, Schweier A, Stedman K, She Q, Phan H, Garrett R, Kristjansson JK (1998) Genetic elements in the extremely thermophilic archaeon *Sulfolobus*. Extremophiles 2:131–140