Microbiology (2009), **155**, 733–740 **DOI** 10.1099/mic.0.023960-0

Short motif sequences determine the targets of the prokaryotic CRISPR defence system

F. J. M. Mojica, C. Díez-Villaseñor, J. García-Martínez and C. Almendros

Departamento de Fisiología, Genética y Microbiología, Facultad de Ciencias, Universidad de Alicante, E-03080 Alicante, Spain

Correspondence
F. J. M. Mojica
fmojica@ua.es

Received 22 September 2008
Revised 18 November 2008
Accepted 3 December 2008

Clustered regularly interspaced short palindromic repeats (CRISPR) and their associated CRISPR-associated sequence (CAS) proteins constitute a novel antiviral defence system that is widespread in prokaryotes. Repeats are separated by spacers, some of them homologous to sequences in mobile genetic elements. Although the whole process involved remains uncharacterized, it is known that new spacers are incorporated into CRISPR loci of the host during a phage challenge, conferring specific resistance against the virus. Moreover, it has been demonstrated that such interference is based on small RNAs carrying a spacer. These RNAs would guide the defence apparatus to foreign molecules carrying sequences that match the spacers. Despite this essential role, the spacer uptake mechanism has not been addressed. A first step forward came from the detection of motifs associated with spacer precursors (proto-spacers) of *Streptococcus thermophilus*, revealing a specific recognition of donor sequences in this species. Here we show that the conservation of proto-spacer adjacent motifs (PAMs) is a common theme for the most diverse CRISPR systems. The PAM sequence depends on the CRISPR-CAS variant, implying that there is a CRISPR-type-specific (motif-directed) choice of the spacers, which subsequently determines the interference target. PAMs also direct the orientation of spacers in the repeat arrays. Remarkably, observations based on such polarity argue against a recognition of the spacer precursors on transcript RNA molecules as a general rule.

INTRODUCTION

Prokaryotes belonging to the most varied groups contain a peculiar type of repetitive DNA, recognized in 2000 as a family (Mojica *et al.*, 2000), distinguished by the regular spacing of the recurrent motif, and consequently defined as short regularly spaced repeats (SRSR). A majority are partially palindromic, a feature that was later incorporated in the present denomination of CRISPR (clustered regularly interspaced short palindromic repeats), proposed by Jansen and co-workers in agreement with our group (Jansen *et al.*, 2002). Repeat units are 24–47 bp long, and alternate with unique sequences (spacers) of similar size (27–72 bp). Despite considerable divergence, CRISPR can be classified based on sequence similarity (Kunin *et al.*, 2007), defining 12 major groups (henceforth referred in the text as CRISPR-*n*, where *n* is the group identification number). Arrays of the same CRISPR are sometimes immediately followed by a conserved AT-rich sequence (Mojica *et al.*, 2000) known as the leader (Jansen *et al.*, 2002). In arrays with a degenerated terminal repeat, the leader is typically located on the opposite edge. Although their role remains undetermined, various reports suggest that leaders promote transcription towards the repeats (Brouns *et al.*, 2008; Lillestøl *et al.*, 2006; Mandin *et al.*, 2007; Tang *et al.*, 2002, 2005; Willkomm *et al.*, 2005). Also, the preferential incorporation of new spacers at the leader-proximal side of the array is consistent with a participation in the recognition of the incoming spacers (Barrangou *et al.*, 2007). A typical CRISPR system within a genome is made up of one or several arrays of the same repeat (with up to 250 units), the adjacent leader sequences, and 6–20 CAS (CRISPR-associated sequence) genes usually in close proximity to one of the arrays. For detailed descriptions of the CRISPR systems see Lillestøl *et al.* (2006) and Sorek *et al.* (2008).

While a large guild (25–45 families) of CAS genes has been found (Haft *et al.*, 2005; Jansen *et al.*, 2002; Makarova *et al.*, 2006), only motifs of *cas1* and *cas2* (core genes) are present in the nine different types of CAS operons. The activities of most CAS proteins are just predicted on the basis of sequence homology (Makarova *et al.*, 2006), structural similarity (Ebihara *et al.*, 2006), or the effects of their

Abbreviations: CAS, CRISPR-associated sequence; CRISPR, clustered regularly interspaced short palindromic repeats; PAM, proto-spacer adjacent motif; PAME, proto-spacer adjacent motif end; SRSR, short regularly spaced repeats.

The GenBank/EMBL/DDBJ accession numbers for the original sequences reported in this paper are FJ232365–FJ232375.

Four supplementary tables and three supplementary figures are available with the online version of this paper.

inactivation (Barrangou *et al.*, 2007; Brouns *et al.*, 2008). Exceptionally, it has been demonstrated that Cas2 family members act as sequence-specific endoRNases (Beloglazova *et al.*, 2008).

Since its first discovery in *Escherichia coli* (Ishino *et al.*, 1987), the involvement of the CRISPR system in several processes has been proposed, including replicon partitioning (Mojica *et al.*, 1995), thermal adaptation (Riehle *et al.*, 2001), DNA repair (Makarova *et al.*, 2002) and chromosome rearrangements (DeBoy *et al.*, 2006). These studies evidenced a certain activity of the CRISPR loci, further supported by the identification of the CAS genes (Jansen *et al.*, 2002), and the detection of transcription of the CRISPR loci (Tang *et al.*, 2002). Afterwards, the finding that spacers derive from pre-existing sequences (proto-spacers) in foreign genetic elements of the spacer holder (Bolotin *et al.*, 2005; Mojica *et al.*, 2005; Pourcel *et al.*, 2005) provided a clue for unravelling the function of CRISPR. Signs of incompatibility between CRISPR spacers and transmissible molecules carrying homologous sequences (Mojica *et al.*, 2005), and an overall correlation between susceptibility to phages and the number of spacers per genome (Bolotin *et al.*, 2005), gave cause to predict the involvement of CRISPR in an immunity-like system. That role was first proved in 2007 for *Streptococcus thermophilus*: CRISPR-harbouring strains became resistant to infection by phages after the acquisition of new spacers derived from the virus (Barrangou *et al.*, 2007). More recently (Brouns *et al.*, 2008), a decreased sensitivity to λ phage has been reported for *E. coli* strains carrying artificial CRISPR systems with spacers targeting essential genes of the virus. At present, the mechanisms underlying the acquisition of spacers and the subsequent resistance are not characterized. Models have been postulated in reference to the prokaryotic regulatory RNAs and the eukaryotic interference RNA systems (Makarova *et al.*, 2006; Mojica *et al.*, 2005). As a common theme of the proposals, RNA molecules carrying at least one spacer would guide the silencing CAS proteins against genetic elements with sequences matching the spacer. In good agreement with this, processing of the CRISPR transcripts has been shown to be essential for the antiviral response (Brouns *et al.*, 2008). With reference to the acquisition of spacers, the first insight into the process came from the finding of motifs associated with proto-spacers of *S. thermophilus* (Bolotin *et al.*, 2005; Deveau *et al.*, 2008; Horvath *et al.*, 2008), implying a specificity in the recognition of the spacer precursors. Here we have explored the presence of proto-spacer adjacent motifs (hereafter referred to as PAMs), corresponding to arrays of every CRISPR type defined by Kunin *et al.* (2007). For the six main groups, we have found the conservation of di- or trinucleotides, starting immediately or one position after the proto-spacer. PAMs are revealed as CRISPR-type-specific motifs, irrespective of the spacer carrier or the proto-spacer holder. The existence of a recognizing motif, together with the polarized arrangement of the spacers in the CRISPR arrays, and the conservation of their relative orientation with respect to the PAM sequences, have fundamental implications for the spacer uptake process, providing at the same time a foundation for testing the elements involved and, in general, for the characterization of the CRISPR-CAS system.

METHODS

PCR, nucleotide sequencing and sequence analysis. CRISPR arrays were detected in publicly available prokaryotic genomes (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_taxtree.html) with a specifically designed computer program (Mojica *et al.*, 2000). In the case of *E. coli*, spacer sequences additional to those from complete genomes were obtained from *E. coli* Reference (ECOR) collection strains (Ochman & Selander, 1984) after PCR amplification and sequencing of CRISPR loci as previously described (Mojica *et al.*, 2005). Putative spacer precursors (proto-spacers) were identified as sequences with at least 90 % identity to spacers, located outside CRISPR loci. Searches were performed with the BLASTN program (Altschul *et al.*, 1997) run against the nr database at the NCBI Website (http://blast.ncbi.nlm.nih.gov/Blast.cgi), with the parameters the application automatically sets for short queries.

Generation of sequence logos. Regions containing sequences with 100 % identity to spacers of related CRISPRs were aligned using either end of the spacer as reference. Gaps were not added in any case. Alignments were visualized with WebLogo (http://weblogo.berkeley.edu/logo.cgi), a Web-based application that generates graphical representations (logos) of the patterns within a multiple sequence alignment (Crooks *et al.*, 2004). Each logo consists of stacks of letters, one stack for each position in the sequence. The height of each stack is measured in bits, with a maximum of 2, and indicates the sequence conservation at that position. The height of letters within the stack reflects the relative frequency of the corresponding nucleotide at that position. The sequence conservation is defined as the difference between the maximum possible entropy and the entropy of the observed symbol distribution (Schneider & Stephens, 1990). WebLogo incorporates a small-sample correction that ameliorates underestimation of the entropy when limited sequence data are analysed.

Identification of leader sequences. Leaders were first identified as large conserved sequences adjacent to CRISPR arrays. When several arrays of the same CRISPR (over 90 % identity) were present in a genome, their flanking regions were aligned. In the case of unique arrays of a particular repeat per genome, equivalent loci located in the closest related strains showing a different CRISPR neighbourhood were compared. Dissimilar CRISPR loci surroundings permit a more confident identification of the leader as a conserved sequence at just one side of the array. Significance was evaluated in both cases by comparison with alignments of the opposite CRISPR flanking region. The presence in the putative leader of A or T tracks, and the occurrence of a degenerated CRISPR in the distal end of the array, were confirmative (Jansen *et al.*, 2002).

RESULTS

Diversity of spacer donors

In order to investigate the conservation of sequences associated with proto-spacers, we first searched for homologues to spacers from nearly 300 strains belonging to over 150 species (the list of genomes and the number of

spacer identities are presented in Table S1, available with the online version of the paper). A total of 204 spacers identical to sequences elsewhere were identified: 182 in bacteriophages, 13 in plasmids, 2 in transposons and 7 in chromosomal sequences not directly related to mobile elements. Relevant features of the genetic elements carrying proto-spacers of the main 12 CRISPR groups are shown in supplementary Table S2. Most plasmids (12 out of 13) either are known to be conjugative or at least have conjugation markers. Only the 8 kb plasmid pSbal04 of *Shewanella baltica* strain OS155 lacks transfer genes. Viruses carrying proto-spacers have a variety of morphologies, including examples of filamentous, icosaedric or complex capsids, classified into the *Siphoviridae*, *Myoviridae*, *Inoviridae*, *Podoviridae* and *Bicaudaviridae* families (http://www.ncbi.nlm.nih.gov/ICTVdb/index.htm). Most are temperate, integrative and non-transposable, with double-stranded linear DNA genomes, but examples of temperate/non-integrative (P1 of *E. coli*), transposable (*Pseudomonas* phages B3 and D3112), virulent (*Listeria* phage A511) and single-stranded linear DNA (PhiXo), as well as double-stranded circular DNA genomes (*Bicaudaviridae* and *Streptococcus agalactiae* unclassified phages) were also detected.

Analysis of sequences associated with spacer precursors

Motifs were investigated after the generation of sequence logos (see Methods) from proto-spacer regions. Alignments were initially performed for sets of at least five proto-spacers, with 100 % identity to spacers, corresponding to arrays of the same CRISPR within closely related strains. Sequences were equally arranged with respect to the corresponding CRISPR. The conservation of 2–3 nucleotide motifs (PAMs), located immediately or one position after just one end of the putative proto-spacers, was revealed in all cases (Figs S1 and S2). When proto-spacers were assigned to their CRISPR group according to Kunin *et al.* (2007), a coincidence of the PAMs corresponding to each CRISPR type was detected. This relationship was confirmed with the logos generated by gathering the proto-spacers belonging to each group (Fig. 1). It is noteworthy that those groups of palindromic CRISPR (Kunin *et al.* 2007), i.e. types 2, 3 and 4, have distinct PAM signatures (CWT, GAA, and GG respectively), whereas the unfolded ones, i.e. types 1, 7 and 10, display the same motif (NGG). Proto-spacers with up to 10 % mismatches further supported these consensuses (see Fig. S1). Thus, the PAMs detected are CRISPR-type dependent, irrespective of the strain carrying the spacer. Additional evidence for such specificity comes from the finding that when CRISPR arrays of different type are present in the same genome (i.e. groups 2 and 4 of *P. aeruginosa*, as well as groups 3 and 10 in *Streptococcus* spp.), their respective PAMs correspond to those of the CRISPR type (Fig. S2), with deviations similarly represented in each strain (see Fig. S1).

Orientation of PAMs

Most spacers within each CRISPR locus have the same orientation with respect to the PAM (Fig. S3). Only a few exceptions corresponding to *Streptococcus agalactiae* CRISPR-10 and *Xanthomonas oryzae* CRISPR-3 spacers were detected (Fig. S1). Taking the CRISPR sequence as a reference, this conservation in orientation extends to the different loci of the same repeat regardless of the genome, and even to CRISPR arrays belonging to the same group, an exception being the spacers of *Yersinia pestis* (see below). These results reveal a prominent relationship between CRISPR and spacer/PAM orientation. Moreover, when leader sequences could be inferred (see Methods), spacer ends equivalent to the proto-spacer edges adjacent to the PAM (here called PAMEs, for proto-spacer adjacent motif ends) are oriented towards the leader (irrespective of the location of the CAS genes), exceptions being *Listeria* spp. CRISPR-10 and, once again, *Y. pestis* CRISPR-4 spacers. Interestingly, CRISPR-10 loci of *Listeria* are also the only exception detected to the conserved orientation of CRISPR versus leader. As a consequence, these *Listeria* PAMEs are still oriented with respect to the CRISPR as in the other loci of the same group. In contrast, PAMEs in *Y. pestis* are oriented oppositely to repeated units of the remaining CRISPR-4 loci. However, *Yersinia* is also peculiar for its weak PAM conservation (WebLogo bits below 1).

Comparative analysis of PAMs versus CRISPR and leader sequences

The conservation of the orientation of spacers (defined by the PAME), with respect to both repeats and leader, together with the preferential incorporation of new spacers at the leader end of the CRISPR arrays (Barrangou *et al.*, 2007; Deveau *et al.*, 2008; Horvath *et al.*, 2008), suggests that CRISPR and leader could participate in the PAM recognition for acquisition of spacers by base pairing and/or insertion through recombination. Moreover, the specificity of PAMs corresponding to groups of palindromic repeats versus the coincidence of those of unfolded CRISPR (see above) could be explained by a participation of the stem–loop in the motif recognition. These possibilities would hold if the motif sequence were present in the corresponding elements. Although PAM sequences are too short for individual coincidences to be deemed significant, reiterated occurrences would support a relationship. This led us to a comparative analysis of PAMs versus repeats and leader sequences. Except for occasional coincidences, no common correlation could be established (Table S3). However, several observations made in *E. coli* could be relevant (Table S4). In this species, there is a strong bias to the CAT motif when the ends of the CRISPRs flanking the spacer are CAC, and to the CTT motif when both CRISPRs end in CTC. Similarly, the four PAMs corresponding to spacers associated with a leader with the CRISPR proximal sequence TCTAAAAGTA are

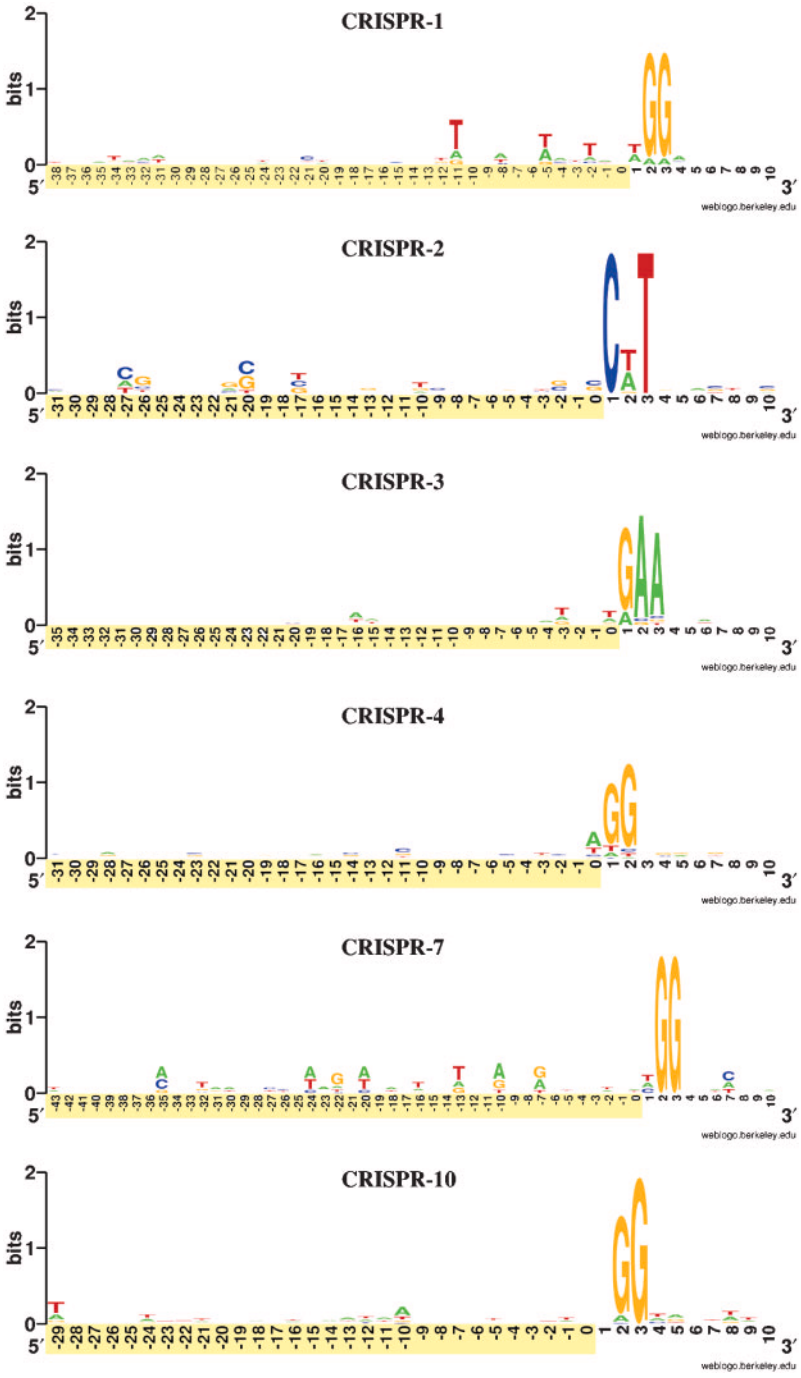

Fig. 1. Proto-spacer region logos built for each CRISPR group. Entries within each stack are equally oriented with respect to the motif revealed for the corresponding CRISPR sequence (Fig. S2), and are aligned relative to the adjacent proto-spacer edge. Sequences include the proto-spacer (shaded positions), and the adjoining 10 nucleotides containing the PAM.

CTT, and 9 out of the 11 associated with a different leader starting ACTAAGCATA read CAT.

Determinants of the PAM sequence

In a previous analysis of spacers acquired after challenge of *S. thermophilus* by related phages, the motif NGGNG was detected for CRISPR-10 proto-spacers (Horvath *et al.*, 2008). However, we have identified a shorter motif (NGG) comparing proto-spacer sequences of CRISPR-10 arrays in *Streptococcus pyogenes*, *S. agalactiae* and *Listeria monocytogenes*. In contrast with *S. thermophilus*, no preference in the fifth position after the proto-spacer was encountered (Figs S1 and S2). Given that we aligned proto-spacer regions from diverse origins (plasmids, unrelated viruses and chromosomal sequences), an explanation of the discrepancy could be that the adjacent recognized sequence varies depending on the identity of the proto-spacer carrier. Hence, proto-spacers from related genetic elements were aligned for each CRISPR type and spacer carrier. As a

result, additional conserved positions were revealed in particular cases, extending the PAMs, and also in the proto-spacer edges (Table S2).

The ambiguous positions within the PAMs defined here could also be explained by the diversity of the spacer donors considered. However, this possibility is dismissed by the equivalent incidence of motif variants corresponding to proto-spacers from even the most distant genetic elements, such as plasmids and phages (Table S2). In the same context, spacers with partially overlapping sequences (see Fig. S1), and thus quite likely derived from closely related genetic elements, are encountered for both CRISPR-2 PAM variants of *E. coli* (consensus motif CWT), indicating a flexibility in the recognition of the motif independent of the spacer precursor.

Finally, in order to investigate a possible contribution of the proto-spacer transcription direction to the PAM signature, we analysed separately those proto-spacer regions corresponding to the coding and template strands. In all cases with at least five entries per set, the same PAM consensus was revealed for both orientations within each CRISPR group (see Fig. S1).

DISCUSSION

CRISPR-CAS have several analogies with the eukaryotic RNA interference (RNAi) systems (Makarova *et al.*, 2006; Mojica *et al.*, 2005), and in particular with the piwi-interacting RNAs (piRNA). Like the prokaryotic repeats, piRNAs are arranged into a limited number of loci on the genome (for a recent review see Kawaji & Hayashizaki, 2008). Significantly, the piRNA precursor is thought to be a long, single-stranded transcript that is cleaved to generate the silencing-competent piRNAs (Brennecke *et al.*, 2007). Similarly, it has been shown that each CRISPR locus is mainly transcribed into a single molecule that becomes further processed into smaller discrete RNAs with the size of a repeat-spacer unit (Brouns *et al.*, 2008; Lillestøl *et al.*, 2006; Tang *et al.*, 2002, 2005). Apart from these resemblances, no homologous proteins are involved, and the architecture of the CRISPR loci, with the regularly arranged repeats and the spacers as guides for the interference, is unique. In consequence, there is currently no working model to address the acquisition of spacers, and very little is known about this process.

Recognition of spacer donors

The existence of a motif adjacent to the spacer precursors has fundamental implications for the generation of the CRISPR arrays. The most evident is that proto-spacers are not randomly selected, i.e. insertion of new spacers into the CRISPR arrays must be the result of an unknown process triggered after the recognition of the corresponding PAM in the donor molecule. A related previous question is how potential spacer donors are identified. CRISPR spacers derive from mobile genetic elements that differ greatly in their transmission mode. Most are aliens (transposons, viruses and plasmids), and even proto-spacers that correspond to chromosomal sequences could also have a foreign origin, being transferred to the receptor by transformation or as part of transmissible elements. In the case of viruses, we have detected proto-spacers in bacteriophages with distinct infection characteristics and genome features, which involve variability in the host integration capacity, replication mechanism, and virion penetration and release modes. The incidence of each virus type roughly corresponds to the number of sequences available in the databases (http://www.ncbi.nlm.nih.gov/ICTVdb/index.htm), and probably to their abundance in nature, implying that there is not a global preference for any given type of phage. It is intriguing, however, that despite the profuse availability of enterophage sequences, analysis of over 500 *E. coli* spacers has given identities to viral sequences only in P1 and prophages of a single *E. coli* strain (E24377A). This could be understood as a preference for viruses of this sort, or alternatively as a proof of the vast number of coliphages that remain unknown.

Assuming a common mechanism for uptake of new spacers, the detection of potential proto-spacer carriers should be unrelated to any peculiarity of the genetic element. The only feature shared by the putative spacer donors identified to date is that all are DNA molecules, which at some stage will be present in double-stranded form in the receptor cell. How the CRISPR machinery recognizes such invaders is an essential question to be elucidated. Related to this, the fact that proto-spacers are found, on the basis of their associated PAM locations, in either the sense or antisense strand (see Fig. S1) excludes a recognition of the spacer precursors on transcript RNA molecules, in support of dsDNA as the donor. This is better understood when considering pairs of proto-spacers with overlapping complementary sequences (see *E. coli* CRISPR-2 alignments in Fig. S1), just one matching the transcript (Fig. 2). According to the models proposed by Makarova *et al.* (2006) for new CRISPR formation, ssRNA could be the precursor of a double-stranded donor molecule. This would imply either indiscriminate duplication of the foreign RNA (rather anti-economical for the cell) or, alternatively, that a signal different from the PAM be recognized in those to be duplicated. In addition, reverse transcription would be required for generating the spacer DNA. However, only a few CRISPR-harbouring strains have putative CRISPR-linked reverse transcriptase (RT) genes (Kojima & Kanehisa, 2008; Makarova *et al.*, 2006). In this respect, it has to be noted that, although no RNA virus carrying proto-spacers has been detected, given the scarce availability of sequences from this type of phage in the databases, such a possibility should not be ruled out. Putative CAS RT activities could be involved in the uptake of spacers from RNA phages as suggested (Kojima & Kanehisa, 2008), but with these particular exceptions, and whereas RT activity is not extensively identified in

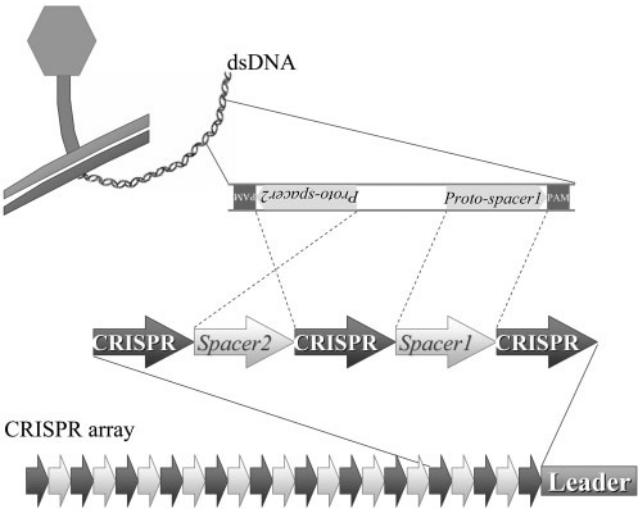

Fig. 2. Acquisition of new spacers. PAMs determine the spacer orientation. Spacers 1 and 2 derive from oppositely oriented sequences with respect to each other but in the same orientation with respect to the PAM. The recognition of the same motif sequence requires the availability of both strands.

association with CRISPRs, the primary spacer donors most likely are dsDNA molecules.

Specificities for proto-spacer recognition

The proto-spacer motif NGGNG was previously reported for the *S. thermophilus* CRISPR-10 loci (Horvath *et al.*, 2008). However, our results reveal a shorter PAM (NGG) for the same CRISPR type. In contrast with the *Streptococcus* studies, we assigned the spacers' origin based on homology, and quite diverse donors were first compared for the establishment of each motif. Another homology-based analysis of proto-spacer regions (Bolotin *et al.*, 2005) showed for the CRISPR-16 loci of *S. thermophilus* the consensus PuPyAAA, which is also shorter than that found by Deveau *et al.* (2008) in the same species (NNAGAAW). Therefore, the homology approach seems to be revealing just the CRISPR-type core of larger proto-spacer motifs. Indeed, additional conserved positions have been detected when comparing related donors in certain cases. Apart from these findings, it is worth noting that mutations in the proto-spacer motif result in CRISPR-defence resistance (Deveau *et al.*, 2008). These sequence changes would be positively selected in CRISPR-challenged foreign elements, and in consequence no conservation of these positions would be evidenced except when comparing direct descents of the same donor. Within this scenario, the most conserved sites associated with the proto-spacers would mainly take part in the acquisition of new spacers, and the less conserved PAM positions, as well as neighbouring additional nucleotides, would essentially be interference-related. In this context, Brouns *et al.* (2008) proposed that small RNAs carrying spacer sequences corresponding to template or coding strands of the target mediate the CRISPR defence by acting on dsDNA, instead of mRNA as previously proposed (Makarova *et al.*, 2006). The occurrence of PAM is in agreement with this target choice on the same basis which suggests that dsDNA is the spacer-donor molecule, as already discussed. Such CRISPR-derived RNA would guide the attack against the spacer complementary sequence in the DNA, where the adjacent nucleotides necessary for interference will be properly arranged.

CRISPR system elements involved in the recognition and integration of new spacers

The conservation of the orientation of spacers in the CRISPR loci with respect to the PAM suggests the recognition of some sequence in the integration site that, given the preference for the leader proximal end of the array, would involve nucleotides in that region, including the CRISPR unit. However, a general pattern of sequence correspondence between PAMs and CRISPR or leader has not been detected. At the same time, the lack of homology with the CRISPR is not compatible with a recognition of the potential spacer precursors by base pairing with the repeat. These results support a main involvement of proteins (i.e. CAS) in both processes, i.e. recognition and integration of new spacers. In this regard, the bias detected in *E. coli* to specific PAMs depending on leader and CRISPR versions could be due to variants in the associated CAS proteins. Certainly, evidence has been presented for a linkage between PAM and the CRISPR sequence type, but the repeats and their associated CAS proteins are unfailingly connected, as supported by the extensive correspondence between the CRISPR grouping (Kunin *et al.*, 2007) and classifications based on the CAS genes content: i.e. the CAS subtypes (Haft *et al.*, 2005) and the CRISPR/CAS system (CASS) versions (Makarova *et al.*, 2006). Thus, the PAMs should be considered specific to the CRISPR-CAS subtype rather than to the CRISPR group. CRISPR-CAS variants could account for a diversity of motifs, and for a flexibility in the mechanism involved in their recognition. The identification of CAS activities is imperative to prove any link in this respect. The only functionally characterized CRISPR-associated proteins are members of the CAS2 family (COG1343 and COG3512), endoRNases that specifically cleave ssRNA regions, which is incompatible with the spacer uptake process as discussed above. Cas5 (Barrangou *et al.*, 2007) and Cas3 (Brouns *et al.*, 2008) are required for the resistance phenotype, and the Cse3 protein of *E. coli* has been shown to be essential for processing CRISPR transcripts, suggesting either an endoRNase or an RNA chaperone activity (Brouns *et al.*, 2008). Cas1 is a putative nuclease/integrase (Makarova *et al.*, 2006), apparently not involved at the CRISPR-related interference stage (Brouns *et al.*, 2008). This core CAS protein is a first-rate candidate for participation in the spacer uptake process.

Conclusion

The existence of PAMs and their analysis have revealed that the spacer precursors are selected after the recognition of adjacent short sequences specific to the CRISPR-CAS variant. Our data indicate that such recognition is exerted on double-stranded molecules of foreign origin, with no evident preference for any given type of genetic element. PAMs determine the spacer orientation within the repeat array. This polarity, although with uncertain functional significance, provides a reference for the alignment of the repeats in comparative studies, and for the location of the leaders. The recognition of the PAMs should aid in the further elucidation of the spacer uptake mechanism, as a basis for the design of experiments aimed at evaluating candidate CAS proteins contributing to the process. The detection of the corresponding motif adjacent to sequences homologous to spacers will also support adjacent sequences as spacer precursors. Additionally, PAMs will have to be considered in the development of the expected applications of the CRISPR system as an innovative molecular biology tool.

ACKNOWLEDGEMENTS

This work was financed by a research grant from the Ministerio de Educación y Ciencia (BIO2004-00523). We are indebted to J. Antón for critical reading of the manuscript.

REFERENCES

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D. A. & Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712.

Beloglazova, N., Brown, G., Zimmerman, M. D., Proudfoot, M., Makarova, K. S., Kudritska, M., Kochinyan, S., Wang, S., Chruszcz, M. & other authors (2008). A novel family of sequence-specific endoribonucleases associated with the clustered regularly interspaced short palindromic repeats. *J Biol Chem* **283**, 20361–20371.

Bolotin, A., Quinquis, B., Sorokin, A. & Ehrlich, S. D. (2005). Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151**, 2551–2561.

Brennecke, J., Aravin, A. A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R. & Hannon, G. J. (2007). Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell* **128**, 1089–1103.

Brouns, S. J., Jore, M. M., Lundgren, M., Westra, E. R., Slijkhuis, R. J., Snijders, A. P., Dickman, M. J., Makarova, K. S., Koonin, E. V. & van der Oost, J. (2008). Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**, 960–964.

Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Res* **14**, 1188–1190.

DeBoy, R. T., Mongodin, E. F., Emerson, J. B. & Nelson, K. E. (2006). Chromosome evolution in the *Thermotogales*: large-scale inversions and strain diversification of CRISPR sequences. *J Bacteriol* **188**, 2364–2374.

Deveau, H., Barrangou, R., Garneau, J. E., Labonté, J., Fremaux, C., Boyaval, P., Romero, D. A., Horvath, P. & Moineau, S. (2008). Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *J Bacteriol* **190**, 1390–1400.

Ebihara, A., Yao, M., Masui, R., Tanaka, I., Yokoyama, S. & Kuramitsu, S. (2006). Crystal structure of hypothetical protein TTHB192 from *Thermus thermophilus* HB8 reveals a new protein family with an RNA recognition motif-like domain. *Protein Sci* **15**, 1494–1499.

Haft, D. H., Selengut, J., Mongodin, E. F. & Nelson, K. E. (2005). A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput Biol* **1**, e60.

Horvath, P., Romero, D. A., Coûté-Monvoisin, A. C., Richards, M., Deveau, H., Moineau, S., Boyaval, P., Fremaux, C. & Barrangou, R. (2008). Diversity, activity and evolution of CRISPR loci in *Streptococcus thermophilus*. *J Bacteriol* **190**, 1401–1412.

Ishino, Y., Shinagawa, H., Makino, K., Amemura, M. & Nakata, A. (1987). Nucleotide sequence of the *iap* gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli*, and identification of the gene product. *J Bacteriol* **169**, 5429–5433.

Jansen, R., Embden, J. D., Gaastra, W. & Schouls, L. M. (2002). Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol* **43**, 1565–1575.

Kawaji, H. & Hayashizaki, Y. (2008). Exploration of small RNAs. *PLoS Genet* **4**, e22.

Kojima, K. K. & Kanehisa, M. (2008). Systematic survey for novel types of prokaryotic retroelements based on gene neighbourhood and protein architecture. *Mol Biol Evol* **25**, 1395–1404.

Kunin, V., Sorek, R. & Hugenholtz, P. (2007). Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol* **8**, R61.

Lillestøl, R. K., Redder, P., Garrett, R. A. & Brügger, K. (2006). A putative viral defence mechanism in archaeal cells. *Archaea* **2**, 59–72.

Makarova, K. S., Aravind, L., Grishin, N. V., Rogozin, I. B. & Koonin, E. V. (2002). A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. *Nucleic Acids Res* **30**, 482–496.

Makarova, K. S., Grishin, N. V., Shabalina, S. A., Wolf, Y. I. & Koonin, E. V. (2006). A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol Direct* **1**, 7.

Mandin, P., Repoila, F., Vergassola, M., Geissmann, T. & Cossart, P. (2007). Identification of new noncoding RNAs in *Listeria monocytogenes* and prediction of mRNA targets. *Nucleic Acids Res* **35**, 962–974.

Mojica, F. J. M., 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, F. J. M., 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.

Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J. & Soria, E. (2005). Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol* **60**, 174–182.

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

Pourcel, C., Salvignol, G. & Vergnaud, G. (2005). CRISPR elements in Y*ersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology* **151**, 653–663.

Riehle, M. M., Bennett, A. F. & Long, A. D. (2001). Genetic architecture of thermal adaptation in *Escherichia coli*. *Proc Natl Acad Sci U S A* **98**, 525–530.

Schneider, T. D. & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* **18**, 6097–6100.

Sorek, R., Kunin, V. & Hugenholtz, P. (2008). CRISPR – a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat Rev Microbiol* **6**, 181–186.

Tang, T. H., Bachellerie, J. P., Rozhdestvensky, T., Bortolin, M. L., Huber, H., Drungowski, M., Elge, T., Brosius, J. & Hüttenhofer, A. (2002). Identification of 86 candidates for small non-messenger RNAs from the archaeon *Archaeoglobus fulgidus*. *Proc Natl Acad Sci U S A* **99**, 7536–7541.

Tang, T. H., Polacek, N., Zywicki, M., Huber, H., Brügger, K., Garrett, R., Bachellerie, J. P. & Hüttenhofer, A. (2005). Identification of novel non-coding RNAs as potential antisense regulators in the archaeon *Sulfolobus solfataricus*. *Mol Microbiol* **55**, 469–481.

Willkomm, D. K., Minnerup, J., Hüttenhofer, A. & Hartmann, R. K. (2005). Experimental RNomics in *Aquifex aeolicus*: identification of small non-coding RNAs and the putative 6S RNA homolog. *Nucleic Acids Res* **33**, 1949–1960.

Edited by: D. W. Ussery