Distribution of CRISPR spacer matches in viruses and plasmids of crenarchaeal acidothermophiles and implications for their inhibitory mechanism

Shiraz Ali Shah*, Niels R. Hansen† and Roger A. Garrett*[1]

*Centre for Comparative Genomics, Department of Biology, Biocenter, Copenhagen University, Ole Maaløes Vej 5, DK-2200 Copenhagen N, Denmark, and †Department of Mathematical Sciences, Copenhagen University, Universitetsparken 5, DK-2100 Copenhagen Ø, Denmark

Abstract

Transcripts from spacer sequences within chromosomal repeat clusters [CRISPRs (clusters of regularly interspaced palindromic repeats)] from archaea have been implicated in inhibiting or regulating the propagation of archaeal viruses and plasmids. For the crenarchaeal thermoacidophiles, the chromosomal spacers show a high level of matches (~30%) with viral or plasmid genomes. Moreover, their distribution along the virus/plasmid genomes, as well as their DNA strand specificity, appear to be random. This is consistent with the hypothesis that chromosomal spacers are taken up directly and randomly from virus and plasmid DNA and that the spacer transcripts target the genomic DNA of the extrachromosomal elements and not their transcripts.

Archaeal CRISPR system

CRISPRs (clusters of regularly interspaced palindromic repeats) consist of identical repeats separated by unique spacer sequences of constant length which occur in the sequenced chromosomes of almost all archaea and approx. 40% of bacteria (reviewed in [1]). The archaeal repeat clusters are generally large and can constitute >1% of the chromosome. The original observation that some spacers show close sequence matches with archaeal viral genomes led to the hypothesis that spacer regions have a regulatory effect on viral propagation [2] and plasmid propagation [1], and this proposal was subsequently reinforced by several studies on both archaea and bacteria (reviewed in [1,3,4]). Moreover, a mechanism for this putative inhibitory effect was suggested, at an early stage, by the finding that RNA transcripts are produced, and processed, from at least one strand of the archaeal repeat clusters [5,6], with the smallest product corresponding roughly in size to a single spacer transcript [1]. This opened for the possibility of an antisense RNA or RNAi (RNA interference)-like mechanism acting either on the viral transcripts or directly on the viral DNA [1,3]. New spacer-repeat units are added at the end of the repeat clusters adjoining a low-complexity flanking sequence [1,7], by a process that probably involves Cas proteins which are generally encoded adjacent to the clusters [3,5,8]. Experimental evidence for such a virus-induced addition was recently provided for bacteria on infecting *Streptococcus thermophilus* with bacteriophages Φ858 and Φ2972 [9].

Hypothesis

In the present article, we explore and interpret trends which emerge when collectively analysing chromosomal CRISPR spacer matches to viral and plasmid genomes. The crenarchaeal acidothermophiles were selected for the analysis because they carry large and multiple repeat clusters [1] and because many of their viruses and plasmids have been sequenced [10]. The results should yield insights into both the mechanism of uptake of new spacer regions in CRISPRs and the mechanism of inhibition or regulation of the viruses and plasmids. We assume that, if chromosomal spacer sequence matches occur randomly on the virus or plasmid genome, then the chromosomal spacer regions are generated by DNA excision and insertion and not by reverse transcription from virus/plasmid transcripts. In contrast, a non-random distribution of matches biased to the genes would favour the latter RNA-based mechanism. A random distribution of spacer matches on the virus/plasmid genomes would also favour a DNA-directed inhibitory mechanism for the spacer transcripts, whereas a gene-biased distribution would support the spacer transcripts inhibiting virus/plasmid gene expression.

Previous studies on the archaeal CRISPRs of related *Sulfolobus solfataricus* strains have suggested that individual spacers are quite stable and that any selective pressure acts on larger blocks of spacers [1], so we infer that any selective pressures on CRISPR spacer contents will not influence our results and interpretation significantly.

Key words: acidothermophile, archaeal plasmid, archaeal virus, cluster of regularly interspaced palindromic repeats (CRISPR), crenarchaeon.

Abbreviations used: ATV, *Acidianus* two-tailed virus; CRISPR, cluster of regularly interspaced palindromic repeats; ITR, inverted terminal repeat; ORF, open reading frame; SIRV1, *Sulfolobus islandicus* rod-shaped virus 1; STIV, *Sulfolobus* turreted icosahedral virus.

[1] To whom correspondence should be addressed (email garrett@bio.ku.dk).

Biochem. Soc. Trans. (2009) 37, 23–28; doi:10.1042/BST0370023

Selection of viruses, plasmids and CRISPRs

Five crenarchaeal virus families, a class of conjugative plasmids and a family of cryptic plasmids were selected for the study (Table 1). They include six β-lipothrixviruses, family *Lipothrixviridae*; four rudiviruses, family *Rudiviridae*; seven fuselloviruses, family *Fuselloviridae*; a single bicaudavirus ATV (*Acidianus* two-tailed virus), family *Bicaudaviridae*; STIV (*Sulfolobus* turreted icosahedral virus), an unclassified icosahedral virus (reviewed in [10]), seven members of a conjugative plasmid family and four members of the pRN cryptic plasmid family (reviewed in [11]). Each extrachromosomal element can propagate in members of the related crenarchaeal thermoacidophilic genera *Sulfolobus* or *Acidianus*. Spacer sequences were derived from 13 whole crenarchaeal chromosomal sequences, from both acidothermophiles and neutrothermophiles, and the partial genomes of *Acidianus brierleyi*, *S. solfataricus* P1 and *Sulfolobus islandicus* HVE10/4 from our laboratory and of *S. islandicus* strains LD85, YG5714, YN1551, M164 and U328 which were publicly available in May 2008 (Table 1).

Identifying spacer matches

CRISPR regions were localized using publicly available software [12,13] and examined for the occurrence of spacer sequence matches to the selected viruses and plasmids. Two approaches were employed. In one, matches were identified at a nucleotide sequence level between the similarly oriented spacer sequences (corresponding to the processed transcript sequence [1,5,6]) and either strand of the virus/plasmid DNA. In a second approach, we exploited the observation that protein sequences are more highly conserved than gene sequences and tried to detect significant matches additional to those identified at a nucleotide sequence level. Each spacer strand was translated into three amino acid sequences, and, after removing sequences containing stop codons (about 50%), each translated sequence was aligned against amino acid sequences of all annotated ORFs (open reading frames) of all the viruses and plasmids. Implicit in this approach is the assumption that the uptake of spacers in the oriented CRISPRs is non-directional, and this is borne out by the results (see below). A nucleotide sequence approach was also applied to the whole acidothermophile chromosomes by searching for exact matches to CRISPR spacers (Table 1). Significant *e*-value cut-offs were determined for both the nucleotide and amino acid sequence searches using the genome sequence of *Saccharomyces cerevisiae* as a negative control (results not shown). All sequence alignments were performed using Paralign, an MMX-optimized implementation of the Smith–Watermann algorithm [14].

Analysis of the distribution of chromosomal spacer matches on virus/plasmid genomes

In total, 82 repeat clusters, some incomplete (Table 1), yielded 4005 spacer sequences, after subtracting 278 spacer sequences shared between *S. solfataricus* strains P1 and P2 [1]. Approx. 30% of the spacers from the acidothermophile genomes match to the virus and plasmid families (Table 1), whereas only approx. 5% matched for the neutrothermophiles. This difference probably reflects that the viruses and plasmids only fall within the host specificity range for the acidothermophiles. The locations of all the spacer matches are superimposed on genome maps of representative genetic elements in Figure 1. Spacers giving nucleotide sequence matches to either DNA strand (red lines) occur mainly within genes, but a few are located intergenically or within the non-protein-coding region of the ITR (inverted terminal repeat). Translated spacers yielding amino acid sequence matches, additionally to the nucleotide sequence matches, occur within annotated ORFs on either DNA strand (green lines).

In a series of three tests, we attempted to address the question of whether or not the spacers present in host chromosomal CRISPRs match the virus/plasmid genomes in a biased non-random manner. Potential biases include the preferential matching to certain regions of the virus/plasmid genome and DNA strand biases. We exclusively used the nucleotide sequence matching data because it covered the whole genome.

First, we examined the distribution of spacer sequence matches, at a nucleotide level, along the virus/plasmid genomes. We assumed that a uniform distribution would follow, roughly, a homogeneous Poisson process, whereas an irregular distribution along the genome would yield a deviation from the homogeneous Poisson process. We investigated for this using Kolmogorov–Smirnov test statistics for each virus and plasmid and we were generally unable to detect any significant deviations from a homogeneous Poisson distribution.

Secondly, we tested whether there was any detectable bias in the spacer matches to the most conserved viral genes given that they are more likely to be targets for inhibition of propagation. The number of matches to each gene was analysed using a Poisson regression model with the gene conservation and length as explanatory variables. This analysis showed that the number of matches to a given gene did not depend significantly upon the degree of its conservation, although, for SIRV1 (*Sulfolobus islandicus* rod-shaped virus 1), we did observe a weak effect for the seven to ten most conserved genes. Moreover, it was found that the expected number of matches was proportional to the gene length, in agreement with the homogeneous Poisson process.

Thirdly, we tested for any bias in the distribution of spacer matches in coding compared with non-coding regions or to the sense compared with antisense strands of the virus/plasmid genes using a specific alternative of a Poisson process with different intensities for matches occurring within, and outside, protein-coding regions, treating each DNA strand separately. We were unable to detect any significant deviations from a homogeneous Poisson distribution for the match intensities of the coding compared with non-coding regions, with the exception of STIV, where there is a bias to the antisense strand (Figure 1).

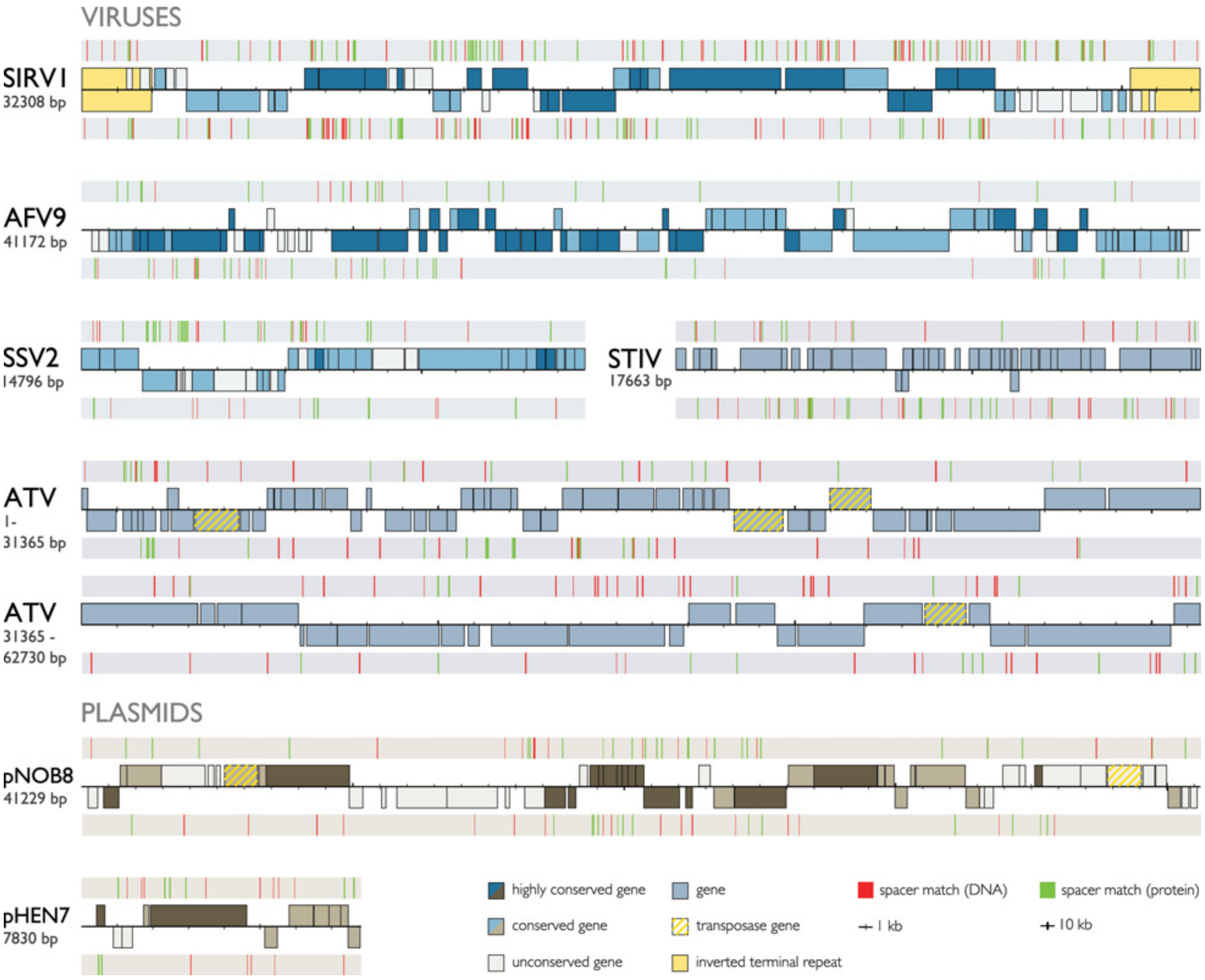
Table 1 | Summary of the chromosomal spacer matches to the virus and plasmid genomes of the crenarchaeal acidothermophiles

The number of CRISPR spacers are given which match virus/plasmid family genomes significantly at a nucleotide level, as well as additional matches detected at an amino acid level. Spacer matches to the host's own genome constitute only exact nucleotide matches. The total number of chromosomal spacers matching to virus/plasmid genomes differs from the number of spacers that match each plasmid and virus family because some spacers match more than one family, but have been counted only once. Rudiviruses comprise SIRV1, SIRV2, ARV and SRV1; β-lipothrixviruses constitute AFV3, AFV6, AFV7, AFV8, AFV9 and SIFV, and fuselloviruses include SSV2, SSV4, SSV5, SSVrh, SSVk1 and SSV1. The pNOB8 family contains pNOB8, pARN3, pARN4, pHVE14, pING1, pKEF9, pSOG1 and pSOG2, and the pRN family consists of pHEN7, pDL10, pRN1 and pRN2. The 278 spacers which *S. solfataricus* P1 shares with strain P2 were subtracted during the analysis, but have been reinserted in this Table. For the partial genomes, the total numbers of spacers are approximate, since repeat clusters may not be fully sequenced. Genome sequences for *S. solfataricus* P1, *S. islandicus* HVE10/4 and *A. brierleyi* are unpublished work from our laboratory. Genomes of *S. islandicus* strains LD85, YG5714, YN1551, M164 and U328 are publicly available from the JGI (Joint Genome Institute) database (http://www.jgi.doe.gov/). All neutrothermophile genomes were complete and obtained through GenBank® accession numbers NC_000854 (*Aeropyrum pernix* K1), NC_008818 (*Hyperthermus butylicus* DSM5456), NC_009776 (*Ignicoccus hospitalis* KIN4/I), NC_003364 (*Pyrobaculum aerophilum* IM2), NC_009376 (*Pyrobaculum arsenaticum* DSM 13514), NC_009073 (*Pyrobaculum calidifontis* JCM11548), NC_008701 (*Pyrobaculum islandicum* DSM4184), NC_009033 (*Staphylothermus marinus* F1) and NC_008698 (*Thermofilum pendens* Hrk5).

Strain	Spacers (total)	Rudiviruses	β-Lipothrixviruses	Fuselloviruses	STIV	ATV	pNOB8 family (conjugative)	pRN family (cryptic)	Spacers (total matching)	Matches with own genome	GenBank®/JGI accession number/reference
Acidothermophiles (total)	3313	331	181	134	81	126	226	63	969	1	–
Sulfolobus solfataricus P2	415	53	24	15	9	20	26	12	135	0	NC_002754
Sulfolobus solfataricus P1	423	50	22	19	9	26	32	7	144	0	[1]
Sulfolobus islandicus HVE10/4	270	47	20	20	4	3	19	9	104	0	Unpublished
Sufolobus tokodaii 7	461	23	19	19	13	2	43	6	108	1	NC_003106
Sulfolobus acidocaldarius DSM639	223	14	5	2	1	2	15	4	38	0	NC_007181
Metallosphaera sedula DSM5348	386	20	9	8	6	59	31	4	110	0	NC_009440
Acidianus brierleyi	367	29	21	9	8	5	32	10	100	0	Unpublished
Sulfolobus islandicus LD85	287	65	39	10	6	1	6	6	114	0	4023472
Four *Sulfolobus islandicus* strains (YG5714, YN1551, M164, U328)	481	30	22	32	25	8	19	5	116	0	4023468, 4005359, 4023464, 4023466
Neutrothermophiles (total)	963	6	13	14	1	4	16	0	52	0	–

Figure 1 | **CRISPR spacer matches superimposed on genomes of representative viruses and plasmids**

SIRV1, rudiviruses; AFV9 (*Acidianus* filamentous virus 9), β-lipothrixviruses; SSV2 (*Sulfolobus* spindle-shaped virus 2), fuselloviruses; STIV, unclassified icosahedral virus; ATV, bicaudavirus; pNOB8, conjugative plasmids; pHEN7, cryptic plasmids. A preliminary version of the rudiviral data was presented in [15]. The circular genomes (SSV2, STIV, ATV, pNOB8 and pHEN7) are presented in a linear format. Protein-coding regions are boxed and shaded, according to their levels of conservation for those genomes for which comparative data are available (all except for STIV and ATV). Spacer sequence matches are indicated by lines above and below the genomes for the two DNA strands and they are colour-coded according to whether they occur exclusively at a nucleotide level (red) or additionally at an amino acid level (green).



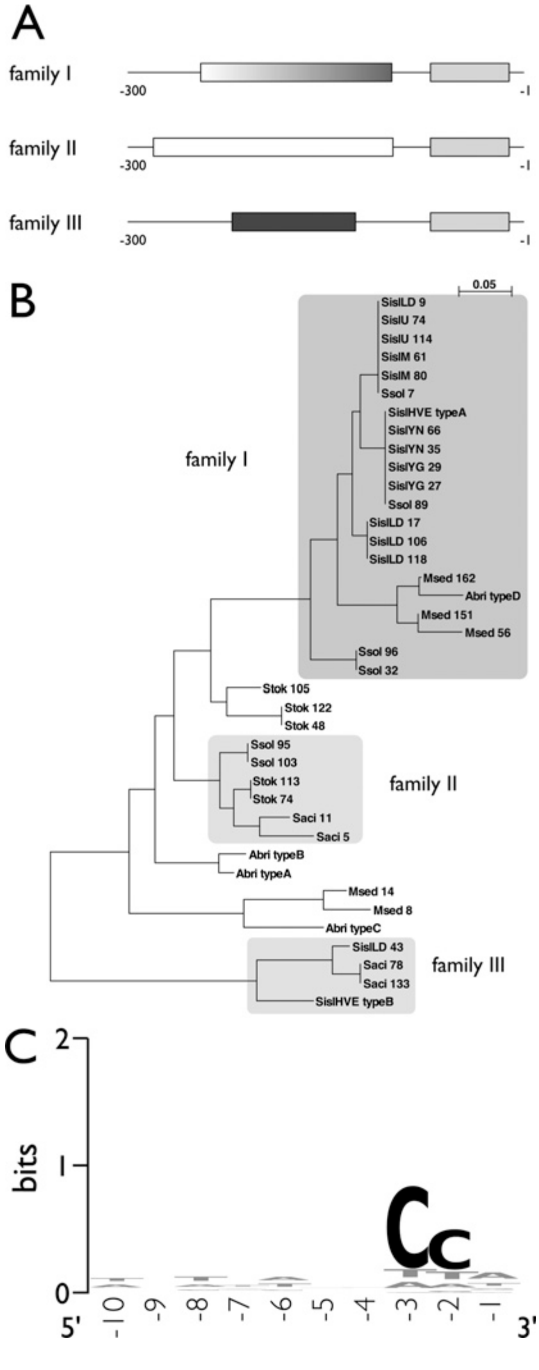
Similar results for the first and third tests were obtained when the analysis was limited to spacer matches from family I CRISPRs (see below).

Classifying crenarchaeal acidothermophile CRISPR families

CRISPRs are oriented and they generally carry a 300–600 bp low-complexity flanking sequence immediately upstream of the repeat cluster which contains the transcriptional leader sequence [1]. Sequence analysis of the flanking sequences by multiple alignment [16] and motif analysis [17], along with sequence comparison of the repeat sequence from each cluster, suggested that the CRISPRs can be classified into families. All crenarchaeal flanking sequences share a common A/T-rich motif adjacent to the first repeat of the cluster, whereas the remainder of the flanking sequence is family-specific. At least three distinct families, each with multiple members, were found for the acidothermophiles by analysing the flanking sequences alone (Figure 2A), and this finding was reinforced by constructing a multiple alignment of repeat sequences from the clusters (Figure 2B). Thus there is a clear correlation between the nature of the flanking sequence and the repeat sequence which constitutes a repeat cluster. These CRISPR families cross species and genus barriers, and most of the acidothermophile genomes contain clusters from

Figure 2 | CRISPR families of crenarchaeal acidothermophiles

(**A**) Schematic representation of the three types of flanking sequence associated with CRISPR families I, II and III. All three flanking sequences share a motif adjacent to the repeat cluster, whereas the upstream region of the flank is specific for each family. (**B**) Phylogenetic tree created using ClustalW [18] based on a multiple alignment of a repeats from each acidothermophile repeat cluster. The CRISPRs studied are labelled by a four-letter prefix based on the genus and species name in addition to the number of repeats carried by the repeat cluster. Abri, *Acidianus brierleyi*; Msed, *Metallosphaera sedula*; Saci, *Sulfolobus acidocaldarius*; Sisl, *Sulfolobus islandicus*; Ssol, *Sulfolobus solfataricus*; Stok, *Sufolobus tokodaii*. *S. islandicus* HVE10/4 and *A. brierleyi* repeat clusters were not completely sequenced and the total number of repeats is not given. The three major repeat cluster families are indicated by differently shaded boxes. (**C**) Logo-plot (http://weblogo.berkeley.edu/) of the motif located upstream of the area on a virus or plasmid genome matched by a group I spacer. The CC motif was found at approx. 75% of all matching sites.



different families. Therefore no families are specific to a given species and no species is limited to a single family. These results strongly reinforce the hypothesis that CRISPR–Cas systems are acquired via horizontal gene transfer [1,19].

Over half of the acidothermophile repeat clusters belong to family I, where, generally, the sequence just upstream of the virus or plasmid site which matches a family I spacer carries a CC motif (Figure 2C). Insufficient data precluded our establishing whether such motifs occur adjacent to family II and family III spacer matches.

Conclusions

The results demonstrate that CRISPR spacer matches are uniformly distributed throughout the virus/plasmid genomes, regardless of both gene location and degree of gene conservation. Moreover, there is no significant bias to either sense or antisense strands of genes (with the exception of STIV): both strands are targeted to an equal degree. These findings strongly suggest that the spacer regions of the CRISPR are taken up randomly, and non-directionally, from the virus or plasmid DNA and are not generated by reverse transcriptase from virus/plasmid transcripts. The results are also consistent with the hypothesis that the CRISPR spacer transcripts target the virus/plasmid by hybridizing directly to their DNA, possibly priming it for degradation.

The results also support a mechanism whereby virus or plasmid propagation is inhibited primarily at a DNA level and not at a gene-expression level. For example, the non-protein-coding ITR region, which is implicated in rudiviral replication [10], carries seven spacer matches in SIRV1 (Figure 1) and other spacer matches occur in intergenic regions which appear not to be involved in transcriptional regulation (results not shown).

The inhibitory mechanism also appears to be highly specific for virus/plasmid DNA, since only one perfect spacer sequence match was detected within any of the acidothermophile chromosomal sequences examined (Table 1). This may be crucial for cell survival if the inhibitory mechanism involves DNA degradation, but, given that viruses and plasmids often integrate reversibly into archaeal chromosomes [20], it suggests that the CRISPR–Cas system selectively targets DNA of extrachromosomal elements, whether circular or linear.

The CRISPR–Cas system has been primarily implicated in viral inhibition in both archaea and bacteria [1,3,4], but it is clear from the present analysis that, at least for archaea, its role is more complex. The apparatus targets plasmids, both conjugative and cryptic, with a similar frequency to viruses (Figure 1). Moreover, some host CRISPR spacers match their own viruses or plasmids, suggesting a regulatory, rather than an inhibitory, role, and this possibility is reinforced by the low copy numbers, and non-lytic properties, of most crenarchaeal viruses [10]. Finally, the observation that a spacer sequence in the repeat cluster of the conjugative plasmid pKEF9 [21] matches a rudiviral genome suggests that plasmids themselves can also inhibit/regulate co-infecting viruses.

Acknowledgements

Dr Kim Brügger kindly provided unpublished genome sequence data.

Funding

Work was supported by the Danish National Research Foundation for a Centre of Comparative Genomics and the Danish Natural Science Research Council [grant number 272-06-0442].

References

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

2 Mojica, F.J., Diez-Villasenor, C., Garcia-Martinez, J. and Soria, E. (2005) Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. J. Mol. Evol. **60**, 174–182

3 Makarova, K.S., Grishin, N.V., Shabalina, S.A., Wolf, Y.I. and 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

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

5 Tang, T.-H., Bachellerie, J.-P., Rozhdestvensky, T., Bortolin, M.-L., Huber, H., Drungowski, M., Elge, T., Brosius, J. and 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

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

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

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

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

10 Prangishvili, D., Forterre, P. and Garrett, R.A. (2006) Viruses of the Archaea: a unifying view. Nat. Rev. Microbiol. **11**, 837–848

11 Lipps, G. (2006) Plasmids and viruses of the thermoacidophilic crenarchaeote *Sulfolobus*. Extremophiles **10**, 17–28

12 Edgar, R.C. (2007) PILER-CR: fast and accurate identification of CRISPR repeats. BMC Bioinformatics **8**, 18–24

13 Bland, C., Ramsey, T.L., Sabree, F., Lowe, M., Brown, K., Kyrpides, N.C. and Hugenholtz, P. (2007) CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinformatics **8**, 209–217

14 Saebø, P.E., Andersen, S.M., Myrseth, J., Laerdahl, J.K. and Rognes, T. (2005) PARALIGN: rapid and sensitive sequence similarity searches powered by parallel computing technology. Nucleic Acids Res. **33**, 535–539

15 Vestergaard, G., Shah, S.A., Bize, A., Reitberger, W., Reuter, M., Phan, H., Briegel, A., Rachel, R., Garrett, R.A. and Prangishvili, D. (2008) SRV, a new rudiviral isolate from *Stygiolobus* and the interplay of crenarchaeal rudiviruses with the host viral-defence CRISPR system. J. Bacteriol. **190**, 6837–6845

16 Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32**, 1792–1797

17 Bailey, T.L., Williams, N., Misleh, C. and Li, W.W. (2006) MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res. **34**, 369–373

18 Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**, 4673–4680

19 Godde, J.S. and Bickerton, A. (2006) The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. J. Mol. Evol. **62**, 718–729

20 Wang, Y., Duan, Z., Zhu, H., Guo, X., Wang, Z., Zhou, J., She, Q. and Huang, L. (2007) A novel *Sulfolobus* non-conjugative extrachromosomal genetic element capable of integration into the host genome and spreading in the presence of a fusellovirus. Virology **363**, 124–133

21 Greve, B., Jensen, S., Brügger, K., Zillig, W. and Garrett, R.A. (2004) Genomic comparison of archaeal conjugative plasmids from *Sulfolobus*. Archaea **1**, 231–239

Received 6 August 2008
doi:10.1042/BST0370023