Research

Open Access

Evolutionary conservation of sequence and secondary structures in CRISPR repeats

Victor Kunin¤, Rotem Sorek¤ and Philip Hugenholtz

Address: DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA.

¤ These authors contributed equally to this work.

Correspondence: Victor Kunin. Email: vkunin@lbl.gov

Published: 18 April 2007

Genome ***Biology*** 2007, **8:**R61 (doi:10.1186/gb-2007-8-4-r61)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2007/8/4/R61

Received: 9 October 2006
Revised: 24 January 2007
Accepted: 18 April 2007

Abstract

Background: Clustered regularly interspaced short palindromic repeats (CRISPRs) are a novel class of direct repeats, separated by unique spacer sequences of similar length, that are present in approximately 40% of bacterial and most archaeal genomes analyzed to date. More than 40 gene families, called CRISPR-associated sequences (CASs), appear in conjunction with these repeats and are thought to be involved in the propagation and functioning of CRISPRs. It has been recently shown that CRISPR provides acquired resistance against viruses in prokaryotes.

Results: Here we analyze CRISPR repeats identified in 195 microbial genomes and show that they can be organized into multiple clusters based on sequence similarity. Some of the clusters present stable, highly conserved RNA secondary structures, while others lack detectable structures. Stable secondary structures exhibit multiple compensatory base changes in the stem region, indicating evolutionary and functional conservation.

Conclusion: We show that the repeat-based classification corresponds to, and expands upon, a previously reported CAS gene-based classification, including specific relationships between CRISPR and CAS subtypes.

Background

Clustered regularly interspaced short palindromic repeats (CRISPRs) are repetitive structures in Bacteria and Archaea composed of exact repeat sequences 24 to 48 bases long (herein called repeats) separated by unique spacers of similar length (herein called spacers) [1,2]. The CRISPR sequences appear to be among the most rapidly evolving elements in the genome, to the point that closely related species and strains, sometimes more than 99% identical at the DNA level, differ in their CRISPR composition [3,4].

Up to 45 gene families, called CRISPR-associated sequences (CASs), appear in conjunction with these repeats and are hypothesized to be responsible for CRISPR propagation and functioning [2,5,6]. It has been proposed that CASs can be divided into seven or eight subtypes, according to their operon organization and gene phylogeny [5,6]. Phylogenetic analysis additionally indicates that CASs have undergone extensive horizontal gene transfer, as very similar CAS genes are found in distantly related organisms [6,7]. CRISPRs and CASs have been found on mobile genetic elements, such as plasmids, *skin* mobile elements, and even prophages, suggesting a possible distribution mechanism for the system [7-9].

CRISPRs have been suggested to play roles in replicon partitioning [1], DNA repair [10], regulation [5] and chromosomal rearrangement [11]. It was recently reported that the spacers are often highly similar to fragments of extrachromosomal DNA, such as phage or plasmid DNA [3,12]. It was suggested that the CRISPR/CAS system participates in an antiviral response, probably by an RNA interference-like mechanism. The proposed mechanism for this CRISPR function involves sampling and maintaining a record of invasive DNA elements, and inhibition of gene functions necessary for invasion [12]. Indeed, it was recently shown that CRISPRs provide acquired resistance against viruses in prokaryotes [13].

Despite in-depth analyses of CASs, the nature of the repeat sequences has not been examined closely. This is presumably because repeats, as short DNA sequences, have less comparative potential than protein-coding genes. Previous studies have noted only that repeats are highly variable, and do not appear to be similar between organisms [2,7]. However, we show that repeats from diverse organisms can be grouped into clusters based on sequence similarity, and that some clusters have pronounced secondary structures with compensatory base changes. We further show that there is a clear correspondence between CAS subtypes and repeat clusters. Our findings have important implications for CRISPR function and diversity.

Results

To obtain a set of CRISPR arrays we employed the PILER-CR program [14] on 439 currently available bacterial and archaeal genomes in IMG version 1.50 [15]. We found 561 arrays, ranging in size from 3 to 220 repeats, in 195 genomes (44% of the genomes tested). These results are in agreement with the results of Godde *et al.* [7], who found CRISPR arrays in 40% of the genomes they tested. Overall, our set of CRISPRs contained 561 repeat sequences (as repeats are generally identical within an array) and 13,372 spacers.

Repeats were first noticed to be palindromic by Mojica *et al.* [16], a feature that was subsequently incorporated into the acronym CRISPR [2]. We hypothesized that the palindromic signature might be indicative of a functional RNA secondary structure within the repeat. This hypothesis is supported by the experimental demonstration that CRISPRs are transcribed and processed into non-messenger RNAs in several Archaea [17], indicating that they are active through an RNA intermediate.

To assess the possibility that CRISPR repeats form stable RNA secondary structures, we used the RNAfold software [18] (see Materials and methods) to predict the intramolecular RNA structure for each of the repeats in our set. This software provides a bit-score that reflects the stability of each secondary structure. We compared the stability of the predicted secondary structure of repeats and spacers to that of similarly sized sequences selected randomly from bacterial genomes (Figure 1a). We found that the folding-score distribution of repeats deviates from the scores for random sequences, indicating a tendency of repeats to form stable secondary structure.

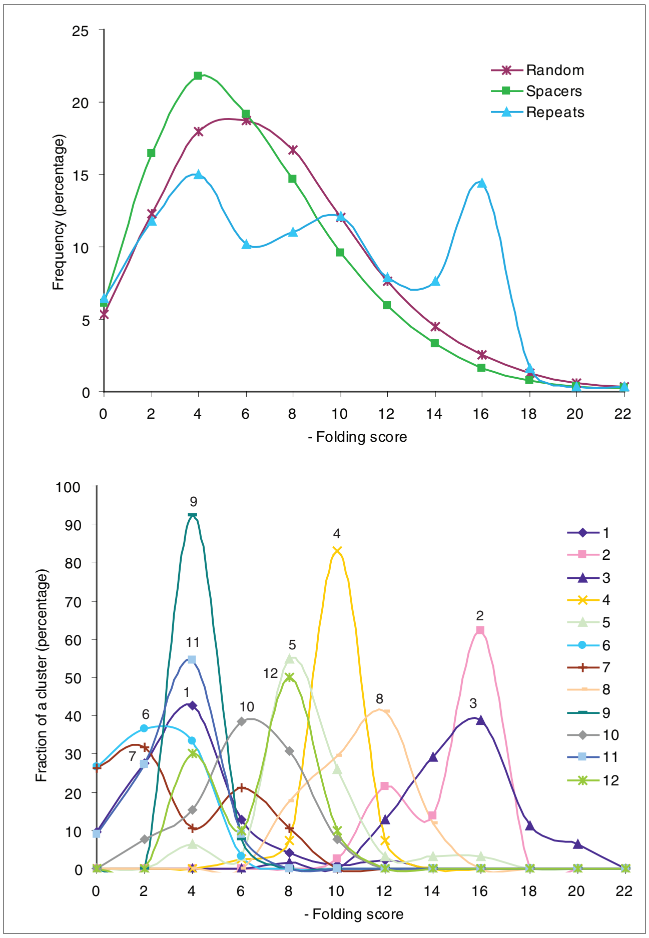

Figure 1
Distributions of folding scores of **(a)** all CRISPR repeats and all spacers, as compared to random sequences and **(b)** individual repeat clusters. X-axis, negative folding scores; Y-axis, fraction (percent) of total.

The trimodal pattern of the RNA folding distribution for CRISPR repeats (Figure 1a) suggests that they are not homogeneous, and that a large subset form stable secondary structures, in contrast to spacers and random sequences. To identify repeat subtypes we first attempted to align each of the 561 repeats in our set to all other repeats using the Smith-Waterman algorithm [19]. The sequence similarity results were then clustered using the MCL algorithm [20] (see Materials and methods). This procedure generated 33 clusters, 12 of which contained 10 or more members, with the largest cluster (cluster 1) containing 94 repeat sequences. Some clusters contained repeats from organisms as distantly related as Archaea and Bacteria, supporting the inference that CRISPR/CAS systems can be horizontally transferred between microorganisms [5-7].

As an independent measure for the validity of the clustering, we examined the RNA stability scores in each of the MCL-defined clusters (note that RNA stability was not taken into account in the clustering procedure). As seen in Figure 1b, clusters 2 and 3 comprise repeats with consistently high folding scores, indicating pronounced secondary structure. By contrast, clusters 1, 6, 7, 9, 10 and 11 contain repeats with consistently poor folding scores. Clusters 4, 5, 8 and 12 show intermediate folding scores, suggesting they have weaker secondary structures. Together, these groups explain the trimodal distribution observed in Figure 1a. The homogeneity of RNA structure stability scores within each cluster, along with the dramatic difference in scores between clusters, suggests that our clustering method is valid.

To further explore the observation that repeats form stable RNA secondary structures, we examined sequence alignments of the repeat clusters. CRISPR repeats are generally considered to be highly dissimilar to each other [7], except for similar repeats in strains of the same species or in closely related species [1]. However, repeats within the clusters we generated, although often containing sequences from vastly different phylogenetic groups, were generally more similar to each other and hence alignable. Figure 2a presents a multiple alignment of a subset of the repeats in cluster 3. A highly stable stem-loop structure was consistently predicted for repeats in this cluster by RNAfold [18] (Figure 1b). Notably, substitutions in the predicted stem structure are consistently accompanied by compensatory changes that preserve the base pairing (Figure 2a). This mutational pattern, together with the presence of G:U base pairs (Figure 2a), is typical of conserved RNA secondary structures and highlights the importance of the stem-loop in the repeats for the functionality of CRISPRs.

A summary of the repeat similarity space is presented in Figure 3. As with cluster 3 (Figure 2), repeats in other clusters with high and intermediate folding scores also form stem-loop structures (Figure 3) and display compensatory mutations, suggesting stable structures. While the stem-loop motif is seen in all of these clusters, the actual sequence, as well as the length of the stem, its position relative to the unstructured region, and the size of the unstructured sequence varies between clusters. For example, while the stem in cluster 4 is typically 5 bp long and is found in the middle of the repeat, the stem in cluster 3 is typically 7 bp long, and is found towards the 5' end of the repeat (Figures 2 and 3). The difference in

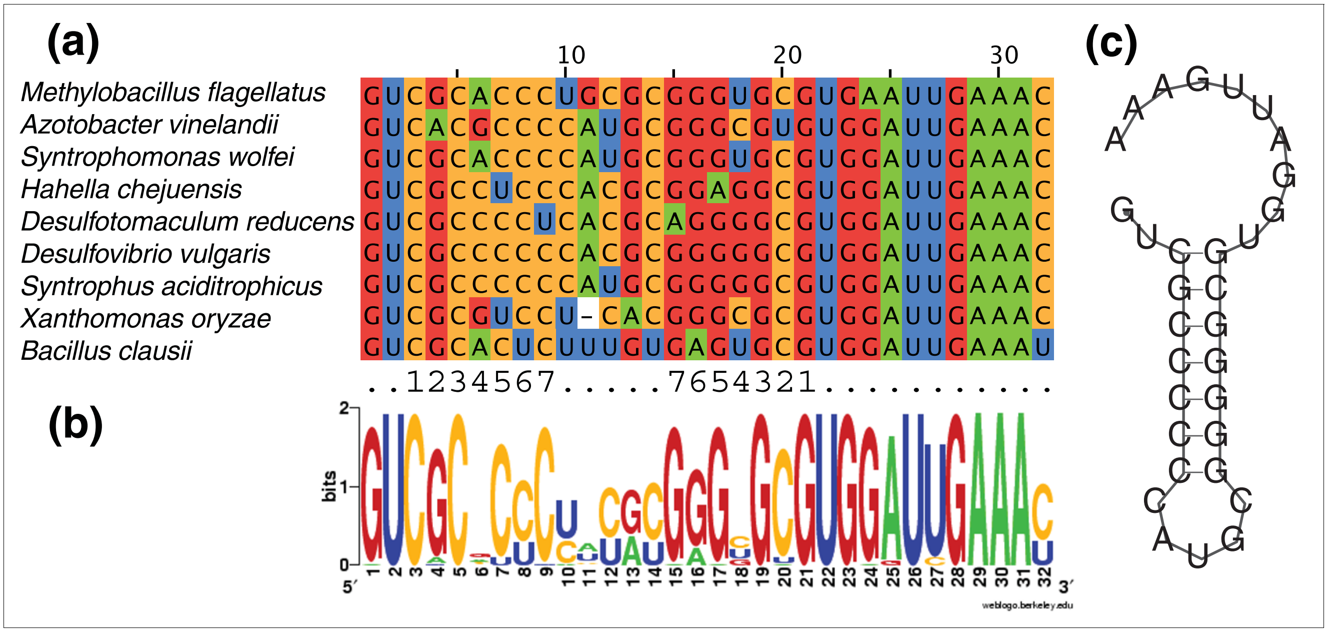
(a)
10
20
30
Methylobacillus flagellatus GUCGCACCCUGCGCGGGUGCGUGAAUUGAAAC
Azotobacter vinelandii GUCACGCCCCAUGCGGGCGUGUGGAUUGAAAC
Syntrophomonas wolfei GUCGCACCCCAUGCGGGUGCGUGGAUUGAAAC
Hahella chejuensis GUCGCCUCCCACGCGGAGGCGUGGAUUGAAAC
Desulfotomaculum reducens GUCGCCCCUCACGCAGGGGCGUGGAUUGAAAC
Desulfovibrio vulgaris GUCGCCCCCCACGCGGGGGCGUGGAUUGAAAC
Syntrophus aciditrophicus GUCGCCCCCCAUGCGGGGGCGUGGAUUGAAAC
Xanthomonas oryzae GUCGCGUCCU-CACGGGCGCGUGGAUUGAAAC
Bacillus clausii GUCGCACUCUUUGUGAGUGCGUGGAUUGAAAU
..1234567.....7654321...........
(b)
bits
2
1
0
5′
3′
weblogo.berkeley.edu
(c)
AAGUUAGG
GUCGCCCCCCAUG
GCGGGGGCGUG

Figure 2

Evidence for secondary structure in cluster 3. **(a)** Multiple alignment of a subset (for clarity) of repeats in cluster 3. Numbers 1 to 7 and 7 to 1 indicate the residues involved in stem base-pairing, some compensatory mutations in the stem are highlighted with circles. Note G:U base pairing at position 5 in *Xanthomonas oryzae* and relaxed conservation of loop residues typical of RNA secondary structure in which the structure is functional rather than the sequence. **(b)** Sequence logo for all repeats in cluster 3. **(c)** Predicted secondary structure of *Syntrophus acidotrophicus* repeat using RNAfold. Stem positions are numbered in accordance with the alignment.

calculated folding scores between clusters with high and intermediate scores is likely to be due to the stem length and the frequency of GC as opposed to AT base pairings. Consistent with previous reports [7], many repeat clusters have a conserved 3' terminus of GAAA(C/G), possibly acting as a binding site for one of the conserved CAS proteins.

Two recent studies identified between 20 and 45 gene families of CASs [5,6]. Based on the tendency of CAS genes to appear together, Haft *et al.* [5] defined eight CAS subtypes (named Ecoli, Ypest, Nmeni, Dvulg, Tneap, Hmari, Apern and Mtube). We sought to determine whether our CRISPR repeat clusters corresponded to particular CAS subtypes. For this, we searched 20 kb of sequence flanking each side of the repeat array for CAS genes using the 45 CAS families TIGRFAM hidden Markov models (HMMs) defined by Haft *et al.* [5].

We found that the Ecoli CAS subtype genes appear exclusively in the proximity of structured repeat cluster 2, and, similarly, the Dvulg and Ypest CAS subtypes correspond strictly to our structured clusters 3 and 4, respectively (Table 1 and Table S1 in Additional data file 1). Presumably, specific and different sets of genes are needed in order to recognize, bind and process the different repeat types. Despite the overall pronounced correspondence between the CAS subtypes and repeat clusters, particularly for structured clusters, there are notable exceptions. For example, the reported frequent co-occurrence of the Mtube subtype with other CAS subtypes [5] is consistent with its promiscuous association with numerous repeat clusters (Table 1). Another interesting exception is the co-occurrence of the Tneap and Apern subtypes in the *Thermococcus kodakaraensis* genome with cluster 6, which is apparently due to a fusion of the Tneap and Apern subtypes (Figure S1 and Table S1 in Additional data file 1). This genome contains three CRISPR arrays, all with identical repeat sequences classified as cluster 6 (Table S1 in Additional data file 1). In some cases the CAS subtype for one or more repeat cluster members differs from the consensus for that cluster (Table S1 in Additional data file 1), suggesting that the association between CRISPR repeat subtypes and CAS subtypes is somewhat flexible.

We also identified a repeat cluster (cluster 5) that is not associated with any of the recognized CAS subtypes. We found that it is associated with most of the core CASs (cas1-4 and cas6), but lacks any of the additional type-defining genes. Cluster 5 occurs exclusively in genomes that contain other CRISPR repeat subtypes and it is possible that it employs at least part of their CAS machinery.

Discussion

This study shows that CRISPR repeats are not structurally homogeneous and can be divided into distinct types based on sequence similarity and ability to form stable secondary structures. This explains why previous attempts to align all repeats resulted in a poorly defined consensus sequence [7]. We observed compensatory base changes in the stems of the structured repeat clusters, including G:U base pairs, indicating that the CRISPR system likely functions through an RNA intermediate.

Some clusters, such as clusters 2, 3 and 4, are discrete in the sequence similarity space, whereas the boundaries of others, such as clusters 1, 6 and 7, were not clearly defined. The discrete clusters were generally composed of structure-forming repeats, and the less well-defined clusters were composed of unstructured repeats. This may be a reflection of the greater evolutionary constraints on the stem structure.

The inference of stem-loop formation within individual CRISPR repeats is in contrast to the speculation that pairs of repeats form duplexes, and are subsequently cleaved to release spacers [6]. Such hypothesized duplexing would unlikely require the ubiquitous presence of the less conserved interior nucleotides, which would form a loop in the single repeat folding model (Figure 2) and an unpaired bulge in the duplex repeat folding model. A CRISPR array in *Sulfolobus* is transcribed and processed into 60 nucleotide long non-messenger RNAs, a size consistent with a single repeat-spacer unit [17,21], supporting the argument that transcribed spacers remain associated with their repeats. The repeats may serve to mediate contact between the spacer-targeted foreign RNA or DNA and CAS-encoded proteins. A stem-loop structure of some repeats may have evolved to facilitate recognition [22] by RNA-binding CAS-encoded proteins, although unstructured *Sulfolobus* repeats (in cluster 7; Figure 3) have been shown to bind via a sequence-specific interaction to a genus-specific protein [23]. This may partly explain the sequence conservation observed in unstructured repeats.

A previous report suggested that spacer regions contribute to the formation of secondary structures in CRISPR arrays [6]. However, we could not detect a significant deviation of spacer secondary structures from random sequences (Figure 1), indicating that spacers are unlikely to be selected based on their secondary structure. In fact, the spacers appear to have slightly weaker structures than random sequences. This is probably due to the AT richness of spacers (46% GC) relative to average bacterial genomic sequences (53% GC), as AT base pairs form less stable structures than GC pairs. The lower spacer GC content is consistent with a proposed viral origin of spacer sequences [3], as viruses are, on average, 7% lower in GC content than bacteria.

Previous attempts to classify CRISPR/CAS systems were based on CAS gene content and phylogeny (mostly of cas1) [5,6]. We add a further dimension to this classification by showing that the repeat sequence itself is also a classifying feature. This can be advantageous in instances where CRISPR arrays occur in the absence of CAS genes. For example, *Ther-*

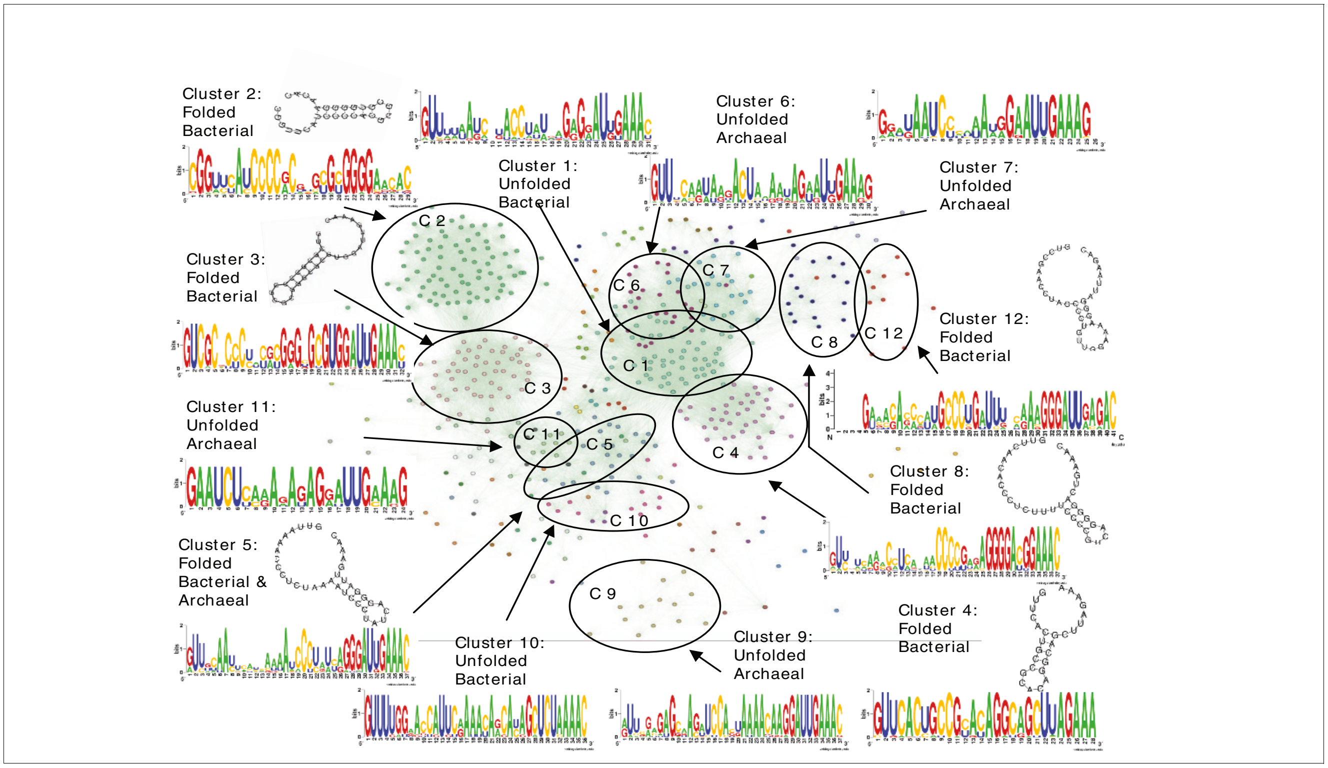
Cluster 2: Folded Bacterial
Cluster 6: Unfolded Archaeal
Cluster 1: Unfolded Bacterial
Cluster 7: Unfolded Archaeal
Cluster 3: Folded Bacterial
Cluster 12: Folded Bacterial
Cluster 11: Unfolded Archaeal
Cluster 8: Folded Bacterial
Cluster 5: Folded Bacterial & Archaeal
Cluster 4: Folded Bacterial
Cluster 10: Unfolded Bacterial
Cluster 9: Unfolded Archaeal
C 2
C 6
C 7
C 8
C 12
C 3
C 1
C 11
C 5
C 4
C 10
C 9

Figure 3
The sequence similarity space of CRISPR repeats visualized with the BioLayout (Java) program [26]. Dots denote individual repeat sequences; connecting lines represent Smith-Waterman similarities, such that closer dots represent more similar sequences. Dot colors denote cluster association as derived from MCL clustering. The 12 largest clusters are indicated by circles together with their sequence logos, coarse phylogenetic composition, and sample secondary structures where applicable.

moplasma acidophilum contains a CRISPR array but lacks CAS genes [5], so it cannot be classified based on CASs. Our clustering indicates that the *T. acidophilum* repeat belongs to (euryarchaeal) cluster 6 (Figure 3; Table S1 in Additional data file 1). In some instances, the repeat classification was able to provide higher resolution than the existing CAS classification. For example, the Nmeni subtype was reported to have an optional gene *csn2* [5]. Our clustering divides this subtype into three clusters (10, 16 and 22). The *csn2* gene is invariably

Table 1

Occurrence of CAS subtypes in the proximity (± 20 kb) of the 12 largest repeat clusters

	Repeat cluster											
CAS subtype	1	2	3	4	5	6	7	8	9	10	11	12
Ecoli		X										
Ypest				X								
Nmeni										X		
Dvulg			X									
Tneap	X					X						
Hmari	X								X			
Apern						F	X				X	
Mtube	X					X		X				X

CAS subtypes are as defined in [5]. Associations are indicated by an X. An instance of a putative fusion between two CAS subtypes is indicated by an F.

present in one cluster (cluster 10) and absent in the other two. The finding of a repeat cluster (cluster 5) that cannot be readily resolved by associated CAS genes (see Results) further demonstrates the power of CRISPR-based classification.

The significant differences between CRISPR/CAS subtypes, both in CRISPR repeat sequence and structure, and in CAS gene content and phylogeny, raises the possibility that the subtypes also differ functionally. Support for this hypothesis could be the fact that frequently several CRISPR/CAS subtypes are found in the same genome and at least four functions have been hypothesized for these elements (host cell defense [12], regulation [5], chromosomal segregation [1] and rearrangement [11]). The study of CRISPRs is in its infancy, and their mode and function is still highly speculative. Our results provide another step toward a comprehensive understanding of these intriguing elements.

Materials and methods

Identification of CRISPR arrays

All genome sequences available through the IMG database version 1.50 [15] were analyzed for CRISPR arrays using the PILER-CR program [14].

Delineation of repeat clusters

Pairwise similarities between repeats were calculated using an in-house implementation of the Smith-Waterman algorithm [19]. The best scoring similarity from the two possible repeat pair orientations, and only scores >7, were used for further analysis. Clustering of pairwise similarities was performed using the MCL program with default parameters [20]. Multiple alignments were performed using MUSCLE [24], and the alignments were manually curated, including removal of outliers. Sequence logos for each cluster were generated using WebLogo [25]. The similarity space of repeats was visualized using BioLayout (Java) [26]. The sequences of the repeats, the assignments to clusters and the multiple alignments are provided as Additional data file 1.

Determining orientation of repeats

The PILER-CR program provides an arbitrary orientation for the repeats. To determine the correct orientation, we compared each repeat to the ones found experimentally to be transcribed into RNA [17,21], assuming that the transcribed direction is the 'correct' direction. The direction most similar to the transcribed repeats (using Waterman similarity scores [19]) was selected as the correct one. We also used the GAAA(C/G) signature at the end of some repeats in cases where the Waterman similarity scores were ambiguous. It is possible, therefore, that some repeats may be presented in the wrong orientation.

Determination of repeat secondary structures

Structural predictions were performed using the RNA Vienna Package [27] downloaded from the Vienna Package server [28,29]. Folding scores for all repeats or individual repeat clusters were divided into bins of 2 score units and plotted as percentages. Random sequence strings with the same length distribution as repeats were generated from the analyzed genomes. The average GC contents were calculated for archaeal, bacterial and viral genomes in the IMG database, version 1.50, and the average GC content was calculated for all spacers in all genomes.

CAS gene identification

The HMMs for CAS genes described in [5] were obtained from the TIGRFAM database, version 6.0 [30]. To identify CAS genes, all coding sequences within 20 kb of the identified CRISPR arrays were searched with the CAS HMMs using hmmpfam [31] with the thresholds of an e-value <0.001 and a positive score.

Additional data files

The following additional data are available with the online version of this paper. Additional data file 1 contains several files showing alignments of clusters 1-12, the arrangement of the CAS cassette in the *Thermococcus kodakaraensis* genome, and CAS genes in the neighborhood of CRISPR arrays as predicted by TIGRFAM, as well as an index of organisms used in the study, a sequence fasta file containing all repeats, and a description of automatic assignment of repeats to clusters with MCL. Some files may be mac-formatted.

Acknowledgements

We thank two anonymous reviewers for their detailed and informative feedback on this manuscript. This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231 and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.

References

1. Mojica FJ, Ferrer C, Juez G, Rodriguez-Valera F: **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* 1995, **17:**85-93.
2. Jansen R, Embden JD, Gaastra W, Schouls LM: **Identification of genes that are associated with DNA repeats in prokaryotes.** *Mol Microbiol* 2002, **43:**1565-1575.
3. Pourcel C, Salvignol G, Vergnaud G: **CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies.** *Microbiology* 2005, **151:**653-663.
4. Bolotin A, Quinquis B, Renault P, Sorokin A, Ehrlich SD, Kulakauskas S, Lapidus A, Goltsman E, Mazur M, Pusch GD, *et al.*: **Complete sequence and comparative genome analysis of the dairy bacterium *Streptococcus thermophilus.*** *Nat Biotechnol* 2004, **22:**1554-1558.
5. Haft DH, Selengut J, Mongodin EF, Nelson KE: **A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes.** *PLoS Comput Biol* 2005, **1:**e60.

6. Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV: **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* 2006, **1:**7.
7. Godde JS, Bickerton A: **The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes.** *J Mol Evol* 2006, **62:**718-729.
8. Sebaihia M, Wren BW, Mullany P, Fairweather NF, Minton N, Stabler R, Thomson NR, Roberts AP, Cerdeno-Tarraga AM, Wang H, *et al.*: **The multidrug-resistant human pathogen *Clostridium difficile* has a highly mobile, mosaic genome.** *Nat Genet* 2006, **38:**779-786.
9. Greve B, Jensen S, Brugger K, Zillig W, Garrett RA: **Genomic comparison of archaeal conjugative plasmids from *Sulfolobus*.** *Archaea* 2004, **1:**231-239.
10. Makarova KS, Aravind L, Grishin NV, Rogozin IB, Koonin EV: **A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis.** *Nucleic Acids Res* 2002, **30:**482-496.
11. DeBoy RT, Mongodin EF, Emerson JB, Nelson KE: **Chromosome evolution in the Thermotogales: large-scale inversions and strain diversification of CRISPR sequences.** *J Bacteriol* 2006, **188:**2364-2374.
12. Mojica FJ, Diez-Villasenor C, Garcia-Martinez J, Soria E: **Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements.** *J Mol Evol* 2005, **60:**174-182.
13. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P: **CRISPR provides acquired resistance against viruses in prokaryotes.** *Science* 2007, **315:**1709-1712.
14. Edgar RC: **PILER-CR: Fast and accurate identification of CRISPR repeats.** *BMC Bioinformatics* 2007, **8:**18.
15. Markowitz VM, Ivanova N, Palaniappan K, Szeto E, Korzeniewski F, Lykidis A, Anderson I, Mavrommatis K, Kunin V, Garcia Martin H, *et al.*: **An experimental metagenome data management and analysis system.** *Bioinformatics* 2006, **22:**e359-367.
16. Mojica FJ, Diez-Villasenor C, Soria E, Juez G: **Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria.** *Mol Microbiol* 2000, **36:**244-246.
17. Tang TH, Bachellerie JP, Rozhdestvensky T, Bortolin ML, Huber H, Drungowski M, Elge T, Brosius J, Huttenhofer A: **Identification of 86 candidates for small non-messenger RNAs from the archaeon *Archaeoglobus fulgidus*.** *Proc Natl Acad Sci USA* 2002, **99:**7536-7541.
18. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer S, Tacker M, Schuster P: **Fast folding and comparison of RNA secondary structures.** *Monatshefte f Chemie* 1994, **125:**167-188.
19. Smith TF, Waterman MS: **Identification of common molecular subsequences.** *J Mol Biol* 1981, **147:**195-197.
20. Van Dongen S: **Graph clustering by flow simulation.** In *PhD thesis* University of Utrecht; 2000.
21. Tang TH, Polacek N, Zywicki M, Huber H, Brugger K, Garrett R, Bachellerie JP, Huttenhofer A: **Identification of novel non-coding RNAs as potential antisense regulators in the archaeon *Sulfolobus solfataricus*.** *Mol Microbiol* 2005, **55:**469-481.
22. Cusack S: **RNA-protein complexes.** *Curr Opin Struct Biol* 1999, **9:**66-73.
23. Peng X, Brugger K, Shen B, Chen L, She Q, Garrett RA: **Genus-specific protein binding to the large clusters of DNA repeats (short regularly spaced repeats) present in *Sulfolobus* genomes.** *J Bacteriol* 2003, **185:**2410-2417.
24. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32:**1792-1797.
25. Crooks GE, Hon G, Chandonia JM, Brenner SE: **WebLogo: a sequence logo generator.** *Genome Res* 2004, **14:**1188-1190.
26. Goldovsky L, Cases I, Enright AJ, Ouzounis CA: **BioLayout(Java): versatile network visualisation of structural and functional relationships.** *Appl Bioinformatics* 2005, **4:**71-74.
27. Mathews DH, Sabina J, Zuker M, Turner DH: **Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure.** *J Mol Biol* 1999, **288:**911-940.
28. **RNA Vienna Package** [http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi]
29. Hofacker IL: **Vienna RNA secondary structure server.** *Nucleic Acids Res* 2003, **31:**3429-3431.
30. **TIGRFAMs Home Page** [http://www.tigr.org/TIGRFAMs/]
31. **HMMER** [http://hmmer.janelia.org/]