Bioinformatics, 32, 2016, i576–i585
doi: 10.1093/bioinformatics/btw454
ECCB 2016

OXFORD

Characterizing leader sequences of CRISPR loci

Omer S. Alkhnbashi[1], Shiraz A. Shah[2], Roger A. Garrett[2], Sita J. Saunders[1], Fabrizio Costa[1,*] and Rolf Backofen[1,3,*]

[1]Bioinformatics Group, Department of Computer Science, University of Freiburg, 79110 Freiburg, Germany, [2]Archaea Centre, Department of Biology, University of Copenhagen N, DK2200 Copenhagen N, Denmark and [3]BIOSS Centre for Biological Signalling Studies, Cluster of Excellence, University of Freiburg, Freiburg im Breisgau, Germany

*To whom correspondence should be addressed

Abstract

Motivation: The CRISPR-Cas system is an adaptive immune system in many archaea and bacteria, which provides resistance against invading genetic elements. The first phase of CRISPR-Cas immunity is called adaptation, in which small DNA fragments are excised from genetic elements and are inserted into a CRISPR array generally adjacent to its so called leader sequence at one end of the array. It has been shown that transcription initiation and adaptation signals of the CRISPR array are located within the leader. However, apart from promoters, there is very little knowledge of sequence or structural motifs or their possible functions. Leader properties have mainly been characterized through transcriptional initiation data from single organisms but large-scale characterization of leaders has remained challenging due to their low level of sequence conservation.

Results: We developed a method to successfully detect leader sequences by focusing on the consensus repeat of the adjacent CRISPR array and weak upstream conservation signals. We applied our tool to the analysis of a comprehensive genomic database and identified several characteristic properties of leader sequences specific to archaea and bacteria, ranging from distinctive sizes to preferential indel localization. *CRISPRleader* provides a full annotation of the CRISPR array, its strand orientation as well as conserved core leader boundaries that can be uploaded to any genome browser. In addition, it outputs reader-friendly HTML pages for conserved leader clusters from our database.

Availability and Implementation: *CRISPRleader* and multiple sequence alignments for all 195 leader clusters are available at http://www.bioinf.uni-freiburg.de/Software/CRISPRleader/.

Contact: costa@informatik.uni-freiburg.de or backofen@informatik.uni-freiburg.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

CRISPR-Cas is an adaptive immune system of archaea and bacteria that provides resistance against invading viruses and plasmids (Barrangou and van der Oost, 2013). 84 and 45% of sequenced archaeal and bacterial genomes, respectively, encode a CRISPR-Cas system (Barrangou and van der Oost, 2013). Each CRISPR-Cas locus comprises several regions. Central to the system is a small 19–48 bp sequence, the CRISPR repeat, which plays a key role in regulating all aspects of CRISPR-Cas function. The CRISPR repeat acts as a regulatory guide and the associated Cas proteins provide the main machinery required for the defence mechanism. The CRISPR array contains repetitions of a CRISPR *repeat* sequence interspaced by foreign DNA fragments (spacers) and can consist of hundreds of repeat-spacer units. Currently, CRISPR-Cas systems are classified into five types and at least 16 subtypes (Makarova *et al.*, 2015; Vestergaard *et al.*, 2014). CRISPR-Cas systems have had a monumental impact on biotechnology as a basis for developing cheap and effective genome-editing techniques for almost any organism (Hsu *et al.*, 2014; Li *et al.*, 2016).

The function of CRISPR-Cas systems can be divided into three major phases: (i) *adaptation*, where a short fragment of invading DNA is inserted into the CRISPR locus for future recognition of that invader; (ii) *expression*, which involves the biogenesis of guide RNA units (crRNA) and their integration into large RNA–protein effector complexes and (iii) *interference*, where these effector complexes vigilantly scan for and degrade invading genetic material previously identified by—and integrated into—the CRISPR-Cas system (Barrangou and van der Oost, 2013). The least understood phase in

CRISPR-Cas immunity is adaptation where a foreign DNA fragment from invading genetic material is integrated.

The integration usually occurs upstream of the first repeat, before a region denoted as the *leader*, which contains regulatory elements important for adaptation. The leaders vary in size, extending from 47 bp in some bacteria to a few hundred bp in some hyperthermophilic archaea, and they tend to exhibit longer regions of low complexity sequence, with limited sequence conservation (Shah and Garrett, 2011). Owing to their limited sequence conservation, even between very similar archaea and bacteria, very little information is available to date and no bioinformatic tool currently exists that can automatically annotate leaders and define their boundaries.

To improve our understanding of the adaptation phase, we studied leader sequences in more detail. Individual experimental studies have demonstrated that the leaders carry the main bacteria- or archaea-specific promoters for CRISPR transcription (Brouns *et al.*, 2008; Lillestol *et al.* 2006, 2009), and that they contain signals for CRISPR-Cas adaptation (Diez-Villasenor *et al.*, 2013; Erdmann and Garrett, 2012; Yosef *et al.*, 2012). The existence of adaptation signals in the leader region is also supported by the existence of leaderless CRISPR-arrays in some crenarchaea, which do not acquire new spacers (Gudbergsdottir *et al.*, 2011; Lillestol *et al.*, 2006, 2009). However, leaderless CRISPR are still functional in the remaining immunity steps because they yield processed CRISPR RNAs (crRNAs), presumably as a result of transcription from promoters taken up randomly in spacers (Deng *et al.*, 2012; Wurtzel *et al.*, 2010).

Concerning the typical length of a leader region, experimental studies of the type I–E CRISPR-Cas system of *Escherichia coli* provided evidence for 40–60 bp of the leader region, located immediately upstream from the first CRISPR repeat, being essential for spacer acquisition (Yosef *et al.*, 2012). Further experiments with the same type I–E system narrowed the critical region to positions −1 to −43 (in relation to the first CRISPR repeat) (Diez-Villasenor *et al.*, 2013). Moreover, for the type I–A system of *Sulfolobus*, a natural deletion of the leader region from positions −47 to −70 resulted in a low level of adaptation activity and also a decreased specificity of spacer acquisition whereby spacer insertions occurred all along the CRISPR array and not just at the first repeat (Erdmann and Garrett, 2012; Garrett *et al.*, 2015).

The existence of adaptation signals in the leader region is also supported by evolutionary studies. Despite their relatively low sequence conservation, sequence clustering studies for the *Sulfolobales* have shown that the leaders tend to coevolve with CRISPR repeat, the adaptation module (Cas1, 2 and 4) and the protospacer-adjacent motif (PAM) (Shah and Garrett, 2011). Experimental support for this coevolution was provided by studies on the *E.coli* type I–E system (Diez-Villasenor *et al.*, 2013). Leaders also carry conserved sequence motifs, currently of unknown function (Garrett *et al.*, 2015; Mojica and Garrett, 2013). The latter are possibly involved in aligning multiple RNA polymerase complexes for CRISPR transcription and/or in assembling Cas proteins adjacent to the CRISPR adaptation site (Lillestol *et al.*, 2009; Marraffini and Sontheimer, 2008; Mojica *et al.*, 2009; Rollie *et al.*, 2015; Shah *et al.*, 2009).

Existing CRISPR-prediction tools do not provide any information regarding CRISPR leaders. In this study, we developed *CRISPRleader*, an efficient approach to determining CRISPR leader boundaries by focusing on leader sequence conservation within groupings based on the similarity of the repeats in the adjacent CRISPR arrays. Our method utilizes a string-kernel technique that can capture more information than traditional sequence alignments and is especially capable of detecting a collection of local motifs. We built specialized HMM models for each of the 51 and the 144 CRISPR-leader clusters from archaea and bacteria, respectively. The method takes a complete genome or draft genome as input and first predicts all possible CRISPR arrays in the correct orientation, and then annotates the CRISPR-leader boundaries.

2 Materials and methods

2.1 CRISPR dataset

In this study, we use the comprehensive dataset of CRISPR arrays of archaeal and bacterial genomes which were downloaded from the CRISPRmap webserver (Alkhnbashi *et al.*, 2014; Lange *et al.*, 2013). The dataset contains 217 archaeal genomes that encode around 985 CRISPR arrays and 1409 bacterial genomes with 3515 CRISPR arrays (a total of 4500 CRISPR arrays). In achaeal CRISPR arrays, the average length of repeats is 29 nt and the average number of repeats per array is 18. In bacteria, in contrast, the average number of repeats per array is 13 and the average repeat length is 30 nt.

2.2 CRISPR leader sequence identification

Although the characteristic repeat-spacer architecture of CRISPR arrays can be easily detected, the orientation of the CRISPR array is inherently ambiguous and thus the determination of the strand from which crRNAs are generated is uncertain. Using the machine learning approach presented in Alkhnbashi *et al.* (2014), it is, however, possible to identify the most probable orientation. Given the array orientation, it is generally assumed that the 3′ boundary of the leader sequence is immediately adjacent to the first CRISPR repeat as both leader and CRISPR array is transcribed in a single transcript (Scholz *et al.*, 2013). Figure 1 depicts a schematic view of a CRISPR locus with the CRISPR array and its respective leader region.

2.2.1 Criteria to determine leaderless CRISPR arrays

In this work, we define a leaderless CRISPR array with the following criteria. First, the distance between the 3′ end of an annotated gene and the 5′ end of the CRISPR array should be less than 20 bp. In the literature, leader regions with experimentally verified function are definitely longer than 20 bp. Second, if the curve fitting procedure fails, it indicates a complete lack of detectable sequence similarity, and we, therefore, discard all the sequences in the leader cluster. Third, we check if the average pairwise similarity as computed by the Needleman–Wunsh algorithm is less than 50%. In this case, there could still be a functional leader present, however, since the sequence similarity between the associated CRISPR repeats is already high, we assume that it is unlikely for such a divergent leader to exist.

2.2.2 CRISPR-leader clusters

It has been shown that the leader sequence coevolves with CRISPR repeats, with the Cas1 protein and with the PAM motif (Shah and Garrett, 2011). To make use of this evolutionary information, we introduce the notion of a leader cluster, which consists of leaders grouped together according to their associated repeat families. By doing so, we overcome the problem of the limited sequence similarity of leaders. To group the repeat sequences, we follow the approach presented in CRISPRmap (Alkhnbashi *et al.*, 2014; Lange *et al.*, 2013). In detail, given a CRISPR array, we first compute the *consensus*-repeat sequence by aligning all repeat sequences without gaps and then take for each position the most frequent nucleotide. We define the similarity between two consensus repeat sequences as

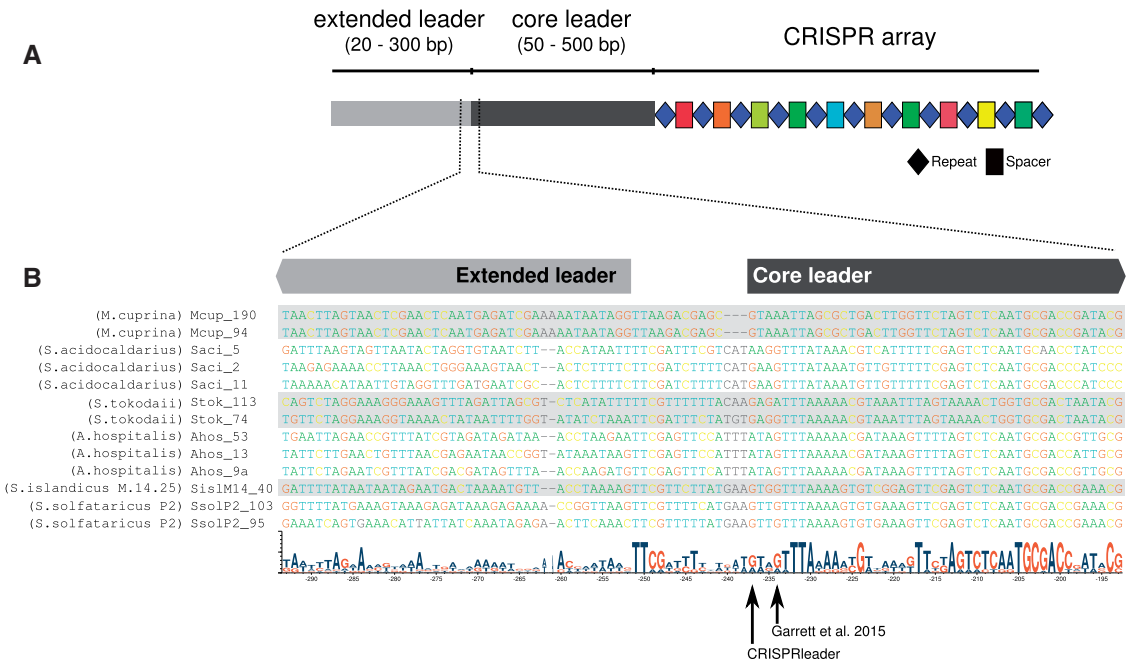

Fig. 1. (**A**) Schematic view of the elements of a CRISPR array showing the repeats (blue diamonds) and spacers (coloured rectangles) of a CRISPR array and the leader region, which we separate into a core and an extended leader. The *core* leader is generally conserved across different host species and is shorter than the *extended* leader which is normally only conserved between multiple leader copies in the same genome. (**B**) Sequences correspond to a cluster of related leaders shared between species of the genera Acidianus, Metalosphaera and Sulfolobus. Each leader is identified by the number of repeats in the adjacent CRISPR. *CRISPRleader* predicts the length of the core leader, since the extended leader is assumed to be functionally less important. In the bottom, we provide an example of a leader alignment to show a detailed view at the junction between the core and extended leader. Here it is possible to see how the extended part is only conserved between multiple copies in the same organism. In contrast, the core part is conserved across all of the different hosts, is underlined by the sequence logo below. The leader boundary predicted by *CRISPRleader* and the boundary determined by expert inspection are indicated by black arrows at the bottom

the global pairwise alignment score computed using the Needleman–Wunsch algorithm (Needleman and Wunsch, 1970). To obtain coherent sets, we then apply Markov Clustering (MCL) (Enright *et al.*, 2002). In CRISPRmap, it was found that better results can be obtained if the similarity matrix is thresholded, i.e. if we set to 0, the similarity value for pairs of repeat sequences that are not sufficiently similar. The only tunable parameter for the MCL algorithm is called 'inflation' and determines the scale of the clustering (i.e. if we prefer many small clusters or few large ones). We optimized these parameters to guarantee that archaea and bacteria are always placed in distinct clusters. This yielded a value of 86 for the similarity threshold and a value of 2.2 for inflation. In this setting, *CRISPRleader* identifies 52 clusters in our dataset of 770 archaeal CRISPR leaders (with a number of leader sequences per cluster that ranges from 3 to 69) and 144 clusters in our set of 2224 bacterial CRISPR leaders (with a number of leader sequences per cluster that ranges from 3 to 184). See Table 1 for details.

2.2.3 CRISPR-leader similarity profile

In the following, we describe how *CRISPRleader* estimates the 5′ leader boundary (CRISPR repeat distal) based loosely on the sequence conservation within a set of leader sequences that are clustered together. We exploit two key assumptions: (i) the 3′ end of the leader (CRISPR repeat proximal) is immediately upstream of the first repeat in the CRISPR array (Brouns *et al.*, 2008; Lillestol *et al.*, 2006, 2009) and (ii) due to evolution-related adaptation signals, the leader sequence will likely exhibit detectable signals of sequence conservation.

First, we trim all the leader sequences to the smallest of (i) an upper limit of 600 nt or (ii) the first occurrence of a predicted protein-coding gene using PRODIGAL (Hyatt *et al.*, 2010) version 2.6.2.

A traditional approach to finding the leader boundaries based on sequence conservation would be to perform a global or local multiple-sequence alignment. In practice, however, the resulting alignments are too noisy. This is likely due to the small size of the conserved regions relative to the sequences length (e.g. 40 nt within 600 nt) and to the small number of sequences (see Section 3.1 for a more detailed analysis). To overcome this issue, we developed a more robust approach based on string kernels (see the following Section for details on the notion of kernels). We start by exploiting the fact that the 3′ boundary of the leader is known to be adjacent to the CRISPR array. We then align all sequences at the CRISPR array's boundary. Subsequently, we apply a running windowing approach: we extract a subsequence from each leader that spans the same W positions and we consider a newly developed average pairwise similarity among these subsequences; we shift the window of a step of S nucleotides and repeat the procedure on the next set of subsequences. Our new pairwise sequence similarity is computed using the Neighbourhood Subgraph Pairwise Decomposition Kernel (NSPDK) (Costa and Grave, 2010). To normalize this similarity value, we consider the average pairwise similarity of the subsequences after a random di- or tri-nucleotide shuffle. The conservation signal is then the log ratio of these two average similarities: when the subsequences are not evolutionarily related, we expect the two similarity values to be comparable yielding log odds scores close to 0. To detect the end of the conserved region, we smoothed the log odds signal by subsequently fitting a parameterized sigmoid curve σ_θ with parameters $\boldsymbol{\theta} = [\theta_1, \theta_2, \theta_3]$ under the constraints that it saturates to 0 at one of the extremes. The parameterization's semantics

Table 1 The leader clusters are summarized at the CRISPRmap repeat family level

Repeat	Repeat consensus sequence	Phylogenetic distribution	Clusters	# Leaders	Avg length
F3(402)	GXXXXXXXXXXXAXXGXATTGAAAG	Crenarchaeota	22	294	210 (±50)
F4(329)	GTTXXAATMAGACXXXWXXXGRATXGAAAX	Euryarchaeota	12	280	236 (±115)
F12(68)	GTTXCAGAXGXACCXTTGTGGGXTTGAA	Euryarchaeota	8	57	111 (±16)
F15(45)	GTTTCXGWAGACATGTXTGGAAA	Euryarchaeota	2	23	366 (±74)
F16(45)	CCAGAAATCAAAAGATAGTWGAAAC	Crenarchaeota	4	41	199 (±3)
F2(556)	GTTTXXAKXXTACCTATXXGGRATTGAAAC	Bacteroidetes/Crenarchaeota/Firmicutes/ Thermotogae	27	437	158 (±46)
F10(89)	TTXXARWXXXXTCCAXTAAAACAAGGATTGAAAC	Euryarchaeota/Firmicutes	6	45	254 (±121)
F1(671)	GTXXTCCCCGCGCXXGCGGGGATRXXCCX	Proteobacteria/Actinobacteria	21	540	103 (±53)
F5(296)	GTCGCXCCCYXXXXGXGXGCGTGGATTGAAAX	Actinobacteria/Planctomycetes/Firmicutes	7	208	83 (±66)
F6(264)	GTTCACTGCCGYAYAGGCAGCTTAGAAA	Proteobacteria	3	230	146 (±14)
F7(236)	XXTKXAMXXTAAXXXXXGWXGTATXTAAAT	Firmicutes/Fusobacteria	13	180	191 (±54)
F8(175)	TXXXXXXXXXCCCCGXXAGGGGAYKGAAAC	Actinobacteria/Deinococcus-Thermus	12	117	157 (±90)
F9(146)	TTXXAAXXXCCCTXTXAGGGATTGAAAC	Cyanobacteria/Firmicutes	10	112	149 (±63)
F11(76)	GTXXXAXXGXCCYGATKXXXARGGGATTRMGAC	Proteobacteria/Bacteroidetes	6	36	107 (±29)
F13(61)	GTTTTAGAGCXXTGTTRTTTXGAATGGTXCCAAAAC	Firmicutes	2	44	210 (±15)
F14(53)	GXXXCXXCGCXGXGGCCXCATTGAAGC	Proteobacteria/Planctomycetes/Firmicutes	3	29	137 (±74)
F17(38)	STGCXXTGATGCCGXWAGGCGTTGACAC	Cyanobacteria/Proteobacteria/Spirochaetes	1	4	213 (±7)
F18(36)	GTTTCYCCTGRRGGTTGAAA	Cyanobacteria/Firmicutes	5	16	176 (±70)
F19(29)	GTTKTAGYYCCYTTTYWMATTTCKYWRTGSTAAAT	Proteobacteria	3	18	116 (±14)
F20(26)	XXXXXGCGXXXCGGCGGXXGXGGX	Acidobacteria/Proteobacteria	1	4	101 (±15)
F21(25)	GTTGWYAAARTAAATTGAAAGCAAWTCACAAC	Bacteroidetes/Ignavibacteriae	1	15	100 (±0.0)
F22(19)	GTYTAGRTGATGTRATCAATAGKTYAAGAC	Firmicutes	2	10	599 (±0.46)
F23(14)	GTTTTGTACTCTARATTTAAGTAACGTAAAAC	Firmicutes	1	6	197 (±8)
F24(11)	WMRTAMCCCCXXAKXAXAGGGGACKARAAC	Firmicutes	1	3	382 (±2)

For each repeat family, the total number of members is given in parentheses, along with the consensus repeat sequence and the taxonomic distribution. The number of leader clusters within each repeat family is also given, along with the total number of leaders found, as well as their average length.

is: θ_1 represents the maximal conservation log odds value, θ_2 represents the length scale factor and θ_3 encodes the position of maximal slope, i.e. the point when the signal transitions from one of the saturated region to the other:

$$\sigma_\theta(x) = \theta_1 \cdot \frac{e^{-\frac{x-\theta_3}{\theta_2}}}{1 + e^{-\frac{(x-\theta_3)}{\theta_2}}}.$$

The estimated leader 5′ boundary is then directly read from θ_3. Figure 2 visualizes the complete process for detecting the boundary of the conservation signal within each leader cluster.

2.2.4 String kernels and explicit feature construction

A string kernel is a function that allows the computational manipulation of strings in a high-dimensional, implicit feature space without ever computing the actual coordinates of the string in that space, but rather by simply computing the inner products between the images of pairs of strings in the feature space. The inner product computed by the kernel function can be used to define a similarity notion. When normalized, the kernel maps pairs of strings s and s' into the interval $[0, 1]$, where 1 means that the two strings are indistinguishable (for the kernel) and 0 that they do not share any resemblance. Popular string kernels are based on the notion of *k-mers*, i.e. substrings of size k. The *k-mer* kernel (also called spectral kernel in Leslie *et al.*, 2002) between s and s', i.e. $K(s, s')$, is the number of the *k-mers* that are identical between s and s'. A normalized kernel computes the fraction of identical *k-mers* w.r.t. the total number of *k-mers* present in the two strings s and s', often as the quantity: $K(s, s')/\sqrt{K(s, s) \cdot K(s', s')}$. Since the occurrence of *k-mers* is exponentially less probable w.r.t. their size k, there is little to gain in considering large *k-mers* (e.g. $k > 10$) when comparing biological

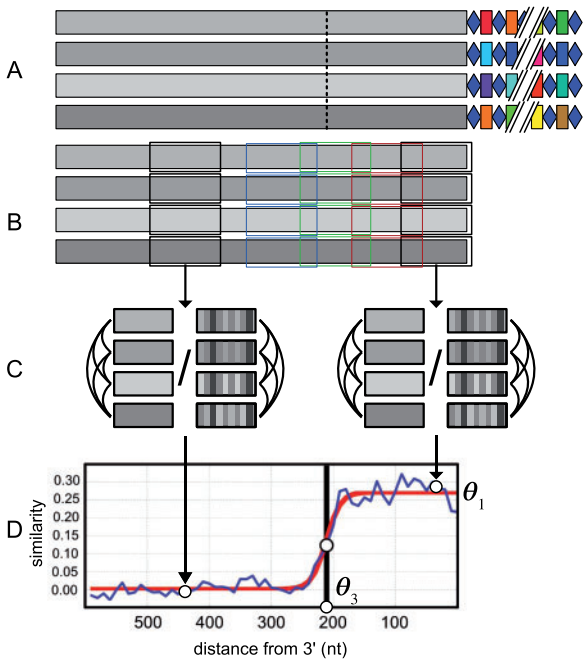

Fig. 2. Leader boundary identification: (**A**) leader sequences are clustered together according to the similarity between the associated repeat sequences; the 3′ end of the sequences in a cluster is aligned w.r.t. the first CRISPR repeat and (**B**) shifting windows spanning the same positions are extracted. (**C**) The average pairwise similarity between all subsequences in a window is computed using the proposed string kernel; the same procedure is applied to shuffled sequences to compute the log odds ratio and (**D**) a saturating function is fitted to distinguish the highly conserved region from the non-conserved one; the point of maximum slope θ_3 is returned as leader boundary

sequences from different species. Small *k-mers*, however, might not yield a sufficient discriminative power. To mitigate these problems, a notion of 'approximate match' was introduced in Leslie *et al.* (2004), where the insertion, deletion or mismatch of up to m components of the *k-mer* is tolerated when counting the correspondences. In practice however, these approximate techniques lead to an increase in run-times and are not always effective in significantly increasing the discriminative power.

The NSPDK approach tries to find a better compromise by restricting the type of mismatches. While the kernel introduced in Costa and Grave (2010) is primarily designed for graphs, here we develop a restricted version for sequences. In detail, the features considered here are pairs of *k-mers* at a fixed distance d, i.e. we assume that there exist a relation $\Phi(s,k,d)$ that is verified for pairs of substrings a, b of s that are of length k and such that their distance is d, we denote such a pair as ϕ_i. The distance between two substrings a, b of s is defined as the length of the substring between the first character of a and the first character of b. The kernel is defined as:

$$K^{k,d}(s,s') = \sum_{\substack{\phi_i \in \Phi^{-1}(s,k,d) \\ \phi_j \in \Phi^{-1}(s',k,d)}} \delta(\phi_i,\phi_j)$$

where $\Phi^{-1}(s,k,d)$ is the inverse of the relation $\Phi(s,k,d)$, i.e. it is the set of all ϕ_i, i.e. pairs of substrings of length k at distance d, and $\delta(x,y)$ is the Kronecker delta, i.e. the function that evaluates to 1 if $x=y$ and to 0 otherwise. We consider the normalized kernel: $\widehat{K}^{k,d}(s,s') = K^{k,d}(s,s')/\sqrt{K^{k,d}(s,s)\cdot K^{k,d}(s',s')}$. Given a maximal value for $k \leq k^*$ and $d \leq d^*$, we consider all the possible combination of values for k and d:

$$\kappa(s,s') = \sum_{\substack{k \leq k^* \\ d \leq d^*}} \widehat{K}^{k,d}(s,s')$$

and finally, we consider the normalized kernel: $\widehat{\kappa}(s,s') = \kappa(s,s')/\sqrt{\kappa(s,s)\cdot\kappa(s',s')}$.

Differently from the standard kernel approach, where the inner product is computed implicitly without having to compute the coordinates of a string in the high-dimensional space, a variant of NPDK, introduced in Frasconi *et al.* (2012), allows one to construct the explicit feature representation in an efficient way. The idea is to exploit a hashing encoding of the decomposed parts. Here, since each feature is a pair of *k-mers* at a given distance, we first hash the *k-mers* individually and then hash the integer triplet formed by the hash values for the two *k-mers* and the distance value into a single integer code. This integer code is then the feature indicator. For each such feature, we count how many times that specific pair of *k-mers* occurs in the sequence. The resulting data structure is a sparse vector representation of the string, which allows an efficient computation of $K^{k,d}(s,s')$ as a dot product.

2.2.5 Optimization of parameters

The method we have employed for the leader boundary determination exposes several parametric choices: the window and step size (W, S), the string kernel complexity (k^*, d^*), the shuffling order for the normalization. To optimize their values, we used supervised data from the work by Lillestol *et al.* (2009) which provides two sets of six and eight leaders for archaeal organisms with experimental evidence for their boundaries. We computed the discrepancy between the experimental and the predicted boundaries as the average squared difference expressed in number of nucleotides. In this setting, we obtained the best results when the window size was $W=40$ nt (selected from $\{10,20,30,40,60\}$), the step size $S=10$ nt (selected from $\{5,10,15,20,30\}$), the maximal *k-mer* size $k^*=3$ and the maximal gap size $d^*=3$ (selected from $\{1,2,3,4,5,6,7,8,9,10\}$), and the order of the shuffling 2 (selected from {1, 2, 3}), i.e. we used dinucleotide shuffling.

2.2.6 CRISPR-leader boundary adjustment via sequence alignment

CRISPR-leaders are inherently quite long, surpassing hundreds of nucleotides in many cases. Thus, there is a potential for indels to accumulate in regions within the leader which are relatively less important, functionally. This means that even closely related leaders can differ in size by tens of nucleotides. To deal with this problem, *CRISPRleader* implements a post-processing procedure to refine the boundary estimate for each individual leader within a cluster. After determining the leader cluster and an initial boundary estimate as previously detailed, we extend the length of each leader sequence (in the 5′ direction) by one third of the respective sequence alignment length of its leader cluster to accommodate undetected indels events. We then perform multiple sequence alignment using the MAFFT tool (Katoh *et al.*, 2002) on the extended sequences belonging to each cluster. The length of the conserved consensus sequence is then yielded as the adjusted boundary.

2.2.7 Automated annotation of core leaders

When given CRISPR arrays of a single organism, *CRISPRleader* automatically annotates the leader region according to our data and delivers a detailed report of all CRISPR arrays, including the boundary of the core leader and the consensus repeat. For the core leader annotation, we first identify the leader cluster from our dataset according to the best-matching consensus repeat. Second, we use the boundaries associated to the corresponding leader cluster to extract the candidate leader core region. Third, we determine whether the putative leader sequence shows sufficient sequence similarity to that leader cluster. For that purpose, we use Hidden Markov Models (HMMs) that we have computed for each leader cluster using HMMER (Eddy, 2011). The corresponding HMM is then used to compute the log-odds score to test whether the new candidate is similar enough to the sequences in the cluster. If the score lies within two standard deviations from the mean log-odds score of the group we accept the sequence, align it to the clustered leaders and compute the length of the conserved consensus sequence. Finally, we report the full alignment with the other clustered leader sequences to highlight insertion or deletion events.

3 Results and discussion

3.1 Conservation profiles could not be detected by alignment-based methods

The more traditional approach to finding the leader boundaries based on sequence conservation would be to perform a global multiple-sequence alignment. In practice, however, the resulting alignment is too noisy to be used to derive a reliable signal. We can hypothesize several reasons that contribute to this situation. First, the conserved region is generally small relative to the sequence lengths, e.g. values of 60 nt within the overall 600 nt sequence are not uncommon. Second, the number of leader sequences that are grouped together in a cluster can be small (less than five). Third, current alignment techniques cannot consistently accommodate transposition events. In practice, it is hard to globally align sequences

when relatively large insertions, deletions and transposition events are possible. For this reason, a more 'local' approach based on *k-mers* can be more effective. To experimentally determine the quality of the conservation signals that can be obtained via alignment strategies, we applied both a global multiple-sequence alignment and a local-alignment strategy to the sequences in the leader clusters. We proceed by incrementally extending the aligned sequence lengths by 10 nt, always starting from the CRISPR array boundary. As shown anecdotally in Supplementary Figure 2, a clear end of the conserved region cannot be reliably detected using global and local alignments, whereas our string-kernel approach shows a very clear conservation boundary on the same data. Owing to this described limitation of leader-boundary detection using the traditional alignment approaches, leaders have not been well characterized in the literature to date. Once detected with our method, however, we could produce well-conserved multiple sequence alignments of precomputed leader clusters for all published genomes that we publish on our website (see availability).

3.2 A well-conserved core leader controls adaptation and transcription

The region of sequence conservation in leaders tends to extend further upstream of the CRISPR locus when similar leaders are compared within the same genome or between closely related strains of the same species. In contrast, when comparing similar leaders across different species, the conserved regions end closer to the CRISPR locus. Here we define the former as the extended leader and the latter as the core leader (Fig. 1). The sequence conservation in the CRISPR-distal regions of the extended leader is likely to have resulted from relatively recent duplication events. The core leader, on the other hand, tends to be well conserved, even for divergent hosts, which implies that only the core region is of special functional significance. In the present study, we predict the boundary of the core leader on the assumption that the additional sequence in the extended leader carries less significant and unknown functions.

According to the literature, two types of regulatory signals fall into the core-leader region. First, both archaeal and bacterial promotors (for transcription) have been detected in the region directly upstream of the CRISPR locus in different type I systems (Brouns *et al.*, 2008; Lillestol *et al.*, 2009). Second, various lines of evidence have implicated this leader region in the adaptation mechanism. In a type I–A system in *Sulfolobus solfataricus*, a natural leader deletion (Fig. 5B) extending from positions −47 to −70 (from the first CRISPR repeat) led to relatively infrequent spacer insertions at different positions along the CRISPR locus (Erdmann and Garrett, 2012; Garrett *et al.*, 2015). In the type I–E system of *E.coli*, it was shown that exchanging leaders between similar CRISPR loci resulted in inverted spacer insertion and in altered sizes of incorporated spacers (Diez-Villasenor *et al.*, 2013). Moreover, attempts to localize the leader regions that are essential for adaptation in type I–E systems demonstrated that some sequences contained within the region −1 to −41 or −60 were essential for adaptation (Yosef *et al.*, 2012). Thus, it is likely that the sequence elements in the core leader normally regulate the frequency and specificity of spacer insertion at the first repeat (by Cas1, demonstrated experimentally to facilitate insertion (Rollie *et al.*, 2015)) as well as controlling the size and orientation of the new spacer.

3.3 The conservation of core leaders is more widespread than previously believed

Early studies characterizing the leader noted its lack of conservation beyond the species boundary (Jansen *et al.*, 2002; Mojica *et al.*, 2000), and this observation was reiterated in later studies (Horvath *et al.*, 2009; Lillestol *et al.*, 2006) when more genome data were available. The first report of related leaders spanning several species and genera was for the crenarchaeal order Sulfolobales (Lillestol *et al.*, 2009), but similar findings have not been made subsequently for other archaea or bacteria. This has probably been due to insufficient genomic data being available and to the difficulty in identifying the leaders using traditional alignment approaches. Nevertheless, the restriction of leaders within tight phylogenetic boundaries stands in contrast to what has otherwise been shown for CRISPR-Cas systems, where most subtypes (except those of type II systems) are shared between the bacterial and archaeal domains (Makarova *et al.*, 2015).

In our study, we find numerous archaeal leader clusters that are shared between several species and genera, but seldom cross the order boundary (Supplementary Fig. 7). For example, the largest archaeal leader cluster contains sequences from *Pyrococcus* and *Thermococcus* species only, both members of the order *Thermococcales*. Of the 10 largest archaeal leader clusters, only one is represented across more than one order (Table 1 and Supplementary Table S1). In contrast, bacterial leader clusters are much more diverse taxonomically (compare Fig. 3A and Supplementary Fig. 7). The two largest bacterial leader clusters are represented by several orders within the phylum Proteobacteria. The third-largest bacterial leader cluster contains members from multiple phyla, including, but not restricted to, Proteobacteria, Firmicutes, Actinobacteria, Chlorobi and Spirochaetes, all within a single leader cluster. The same staggering diversity is seen throughout the other major bacterial leader clusters (Fig. 3).

Conventional wisdom within the field has long been that CRISPR leaders are not conserved beyond the species boundary. Conservation across the order Sulfolobales was shown previously (Lillestol *et al.*, 2009) and our results show that order-wide conservation is normal for archaea. In contrast, for bacteria there seem to be no taxonomic boundaries for leader-cluster diversity. We found similar leaders in bacteria as diverse as *Pseudomonas* and *Clostridium* and the compatibility of leaders across diverse phyla is comparable to that of the CRISPR subtypes themselves. This kind of diversity within bacterial leader clusters seems to be the rule rather than the exception, but has so far gone undetected owing to the lack of reliable methods for leader identification. As for the stark difference between archaea and bacteria, in terms how widely conserved their leader clusters are, no straightforward explanation arose from the data. One factor may be that the currently sequenced archaeal genomes are strongly biased to extremophiles present in isolated environments which include salt lakes and acidophilic hot springs that exhibit more limited biodiversity. The few archaeal genomes that do originate from more complex microbial environments do not tend to carry CRISPRs. Thus, the lack of widespread conservation of currently sequenced archaeal leader clusters may simply result from the formidable barriers to horizontal gene transfer imposed by their habitats.

3.4 Core leaders display different patterns of conservation

Leader clusters that are more taxonomically restricted tend to show a relatively high and uniform sequence conservation throughout the

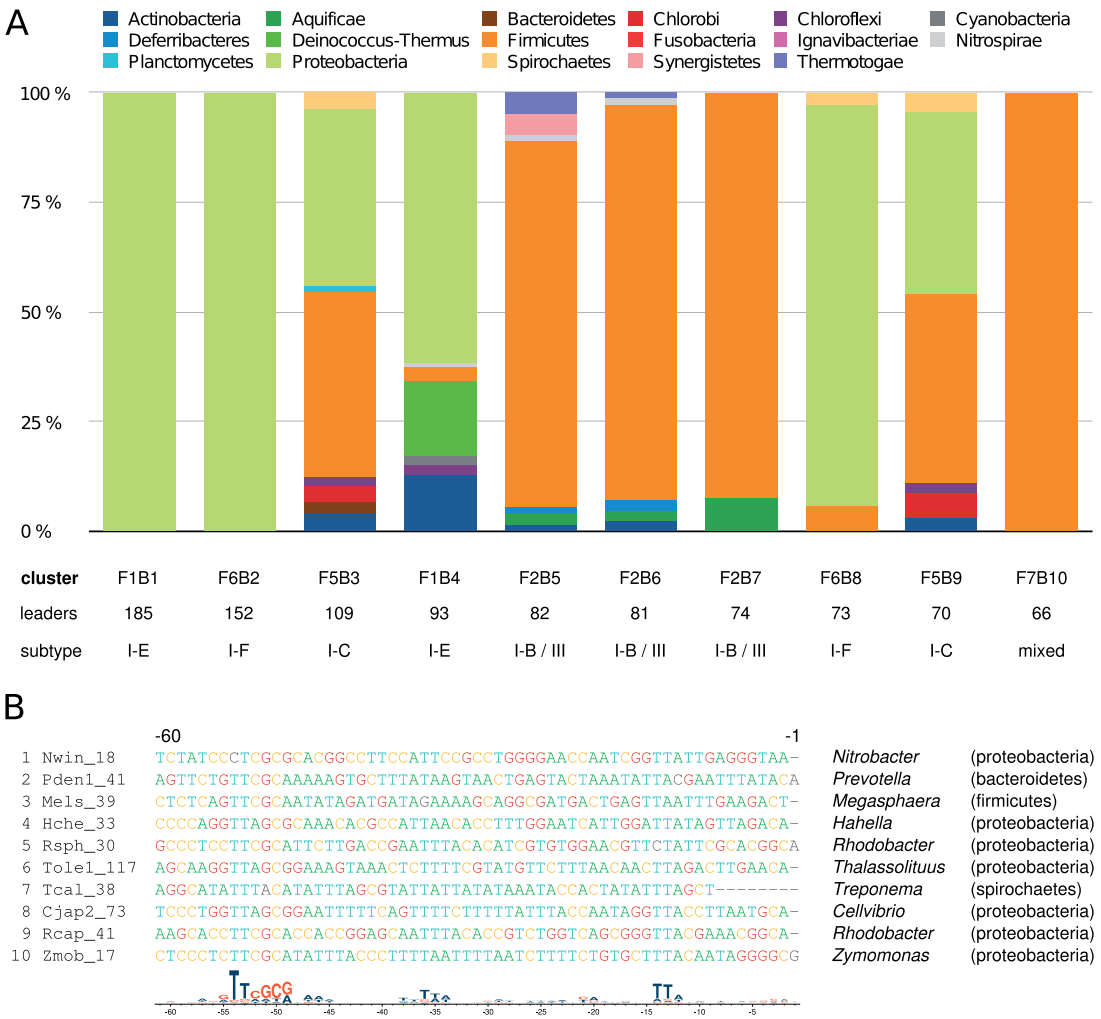

Fig. 3. (**A**) The taxonomic distribution is shown, on the phylum-level, for each of the ten largest bacterial leader clusters. Despite proteobacteria and firmicutes dominating the underlying genomic data, diversity is still evident with most clusters representing several additional phyla. The number of leaders in each family is also shown along with the principal CRISPR-Cas subtype associated with the leaders. (**B**) An alignment of the core leader from 10 randomly selected members of cluster 3 is shown, along with names of the genera and phyla they originated from. The logo plot at the bottom is based on all 109 members of bacterial cluster F5B3. The wide taxonomic distribution within the cluster is reflected in the individual leader sequences, which are evidently very diverse. Throughout much of the alignment, any sequence identity is undetectable. However, the alignment is anchored near either end by two prominent sequence motifs which are present in most sequences despite their divergence

entire core length (Fig. 1B). This uniform conservation may just reflect that the leaders have not yet had time to diverge sufficiently in order for functionally important regions to stand out from their background. In contrast, the taxonomically diverse bacterial leader clusters have diverged to such an extent that sequence identity is undetectable throughout most of the sequence length (Fig. 3B). Instead, small motifs exist that are conserved in both sequence and position across diverse members of the same cluster. These motifs not only confirm a common origin for the leaders within that cluster, but also may be crucial for their function. Prominent sequence motifs are featured towards repeat-distal ends of core leaders for the major bacterial leader clusters F5B3, F2B5 and F2B5 (Supplementary Table *S1*). In contrast, the repeat proximal end is more divergent with numerous indels (Fig. 5A), showing little to no overall sequence conservation. Low sequence conservation towards the repeat proximal end in bacterial leaders, although common, is not the rule, as some leader clusters (e.g. bacterial cluster F1B12) do show the opposite pattern with a conserved proximal end and a divergent distal end (Supplementary Table *S1*).

3.5 Predicted core leaders coincide with published results and are generally longer in archaea than in bacteria

Using our *CRISPRleader* approach, we determined the conservation boundaries and respective leader length distributions for archaea and bacteria separately. The frequency of leader lengths peaked at about 60 and 130 bp in bacteria with a smaller peak at 190 bp, while in archaea, lengths were larger with peaks at 100, 220 and 290 bp (Fig. 4A) suggestive of some diversity of function. The leader boundaries obtained coincided closely with previously described CRISPR leaders for a few organisms in the literature (summarized in Table 2).

3.6 CRISPR loci are frequently leaderless

Individual observations of leaderless CRISPR loci have been reported that are defective in transcription and inactive in adaptation but it remains unclear whether they have lost their leaders or whether they have simply been separated from the leader distal ends of other CRISPR loci, possibly as a result of transposition events There are no data available on the extent of leaderless loci and,

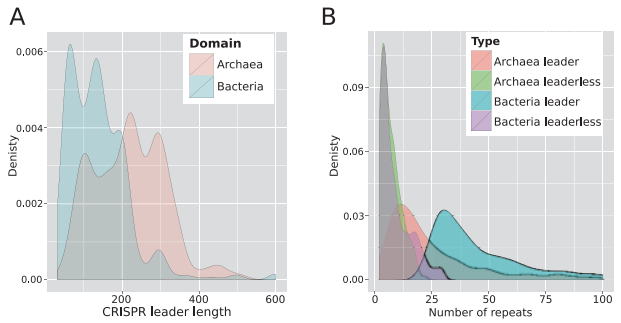

Fig. 4. (**A**) A comparison of the CRISPR leader length distributions between archaea and bacteria. It shows that archaea and bacteria are grouped into limited size ranges. The archaea peak leader sizes are larger with average values 100, 220 and 290 bp while the bacteria leader sizes are smaller with average values 60 and 160 bp. (**B**) The distribution of leader-containing and leaderless CRISPR loci in archaea and bacteria. The size distributions for leaderless CRISPR loci are similar for archaea and bacteria

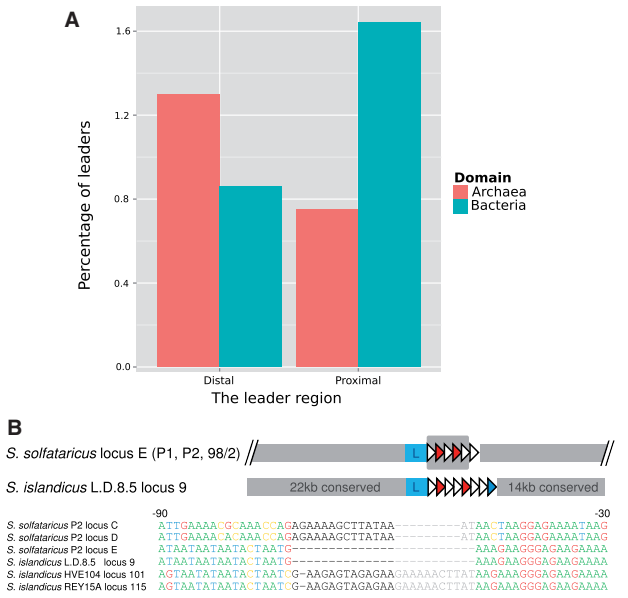

Fig. 5. (**A**) The distributions of insertions and deletions (indels) in the leader regions. Bacterial leaders more often carry indels towards the repeat proximal end, while archaeal leaders have them at the repeat distal end. (**B**) *Sulfolobus* leader deletions implicated in the adaptation phase. Part of an alignment between a series of Sulfolobus CRISPR leaders of cluster 2 is shown. *S. solfataricus* CRISPRs C and D acquire spacers during viral challenges, as does *S. islandicus* REY15A locus 115. *S.solfataricus* locus E is deficient in adaptation, acquiring spacers abnormally and at a very low rate, in turn making the CRISPR very small. A similar small locus is found in *S.islandicus* L.D.8.5. The leaders of both loci share a deletion around 50 bp from the first repeat, which is not found in the adaptation proficient leaders, consistent with a role in adaptation deficiency

therefore, we estimated the percentage of CRISPR loci that lack leaders and calculated the sizes of their arrays (number of spacer-repeat units) relative to those loci with conserved leaders. The results demonstrated that 13% of 980 archaeal CRISPR loci, and 24% of 2852 bacterial loci, were considered leaderless (Fig. 4B). Moreover, the sizes of the leaderless CRISPR arrays were much smaller on average (Fig. 4B). The smaller sizes are consistent with the leaderless loci being inactive in CRISPR adaptation and unable to increase in size but they are also consistent with them having separated from the ends of other CRISPR loci.

3.7 Leader clusters correlate more with Cas1 phylogeny than the subtype classification

Earlier studies have demonstrated that the sequences of leaders, repeats and Cas1 tend to coevolve for the type I–A CRISPR-Cas systems of the Sulfolobales (Shah and Garrett, 2011) and Thermoproteales (Garrett *et al.*, 2011). It was inferred that all these components were involved in spacer acquisition, whereas components of the interference effector complex evolved separately.

We quantified the degree of interdependence and coevolution of the leader clusters against Cas1 phylogeny and the cognate CRISPR subtype, respectively, by applying the Adjusted Rand Index (ARI) (Rand, 1971) that measures correlation between clusters. Leader clusters correlated with Cas1 clusters yielding an ARI value of 0.75, indicating a high degree of correlation. Conversely, the ARI between leader clusters and CRISPR subtypes was only 0.37. We infer that the lower correlation between the leader cluster and CRISPR subtype indicates that the same leader type can cofunction with CRISPR systems of different subtypes, and vice versa, as long as the correct adaptation module (i.e. Cas1, Cas2 and Cas4) is present to interact with the leader and maintain the CRISPR locus. This is consistent with the numerous reports of modular exchange where different adaptation and interference modules interchange to form new combinations of functional CRISPR-Cas systems (Garrett *et al.*, 2011; Makarova *et al.*, 2015; Vestergaard *et al.*, 2014). Since the latest CRISPR-subtype classification (Makarova *et al.*, 2015) primarily reflects the diversity of the interference modules, a lower correlation between CRISPR subtypes and leader clusters is to be expected.

Table 2. Comparison of predicted leader lengths against published leaders

Organism name	Published	Predicted	Difference
E.coli IYB5101 (Yosef *et al.*, 2012)	100	105	5
E.coli BL21-AI (Yosef *et al.*, 2012)	100	95	5
C.jejuni (Tasaki *et al.*, 2012)	146	144	2
Synechocystis pcc6803 (Scholz *et al.*, 2013)	125	116	9
S.pyogenes (Fonfara *et al.*, 2014)	109	108	1
S.solfataricus (Lillestol *et al.*, 2009)	238	237	1
M.marzei Göl (Nickel *et al.*, 2013)	108	108	0
M.marzei Göl (Nickel *et al.*, 2013)	108	111	3

3.8 Automated annotation of core leaders and CRISPR arrays using CRISPRleader

CRISPRleader accepts either a complete or partial genome sequence as input and provides a full annotation of the CRISPR array, their strand orientation as well as conserved core leader boundaries. In addition, it outputs reader-friendly HTML pages for conserved leader clusters from our database and it provides a standardized BED format that can be used to visualize CRISPR arrays and leader annotations in any genome browser.

4 Conclusion

Adaptation is currently the least understood of the main phases in the CRISPR-Cas immune system. Although it is known that

adaptation is affected by signals present in the region upstream of the CRISPR array, the so-called leader sequence, no bioinformatic tool exists that can automatically annotate these leader sequences to date. This is due to the fact that the known leader sequences exhibit only limited sequence conservation. To gain a deeper understanding, we developed a novel *k-mer*-based tool, *CRISPRleader*, that can reliably detect the CRISPR leader boundaries.

We analyzed 1426 archaeal and bacterial genomes using *CRISPRleader* and identified several characteristic properties of the leader sequences. Results show that although an extended region can be conserved between few very closely related species or CRISPR loci, generally a smaller core leader region, directly adjacent to the CRISPR locus, is conserved between more distantly related species.

We identified core leaders from 770 archaeal and 2224 bacterial CRISPR loci and observed significant differences between leader clusters. First, core leaders tend to be longer in archaea than in bacteria. Second, leader clusters in archaea are more homogeneous in terms of phyla than in bacteria. This may reflect the fact that archaea have survived primarily in low-energy environments which are often quite isolated (e.g. solfataric fields or hypersaline lakes) such that genetic exchange is much more limited than for most bacteria. Third, bacteria exhibit more indels in the CRISPR-proximal region of the core leaders than archaea. This core leader region has been shown to be important for CRISPR transcription and CRISPR-Cas adaptation and may be readily inactivated, or modulated, by indel activity, possibly triggered by an invader to circumvent targeting.

Regarding common characteristics, we showed that in both archaea and bacteria (i) leader sequences and repeats tend to coevolve with the Cas1 protein more broadly than previously believed, i.e. irrespectively of the system's subtype and (ii) leaderless CRISPR loci tend to be much smaller than loci with a leader present. This is possibly indicative of a displacement event from the leader-distal ends of other CRISPR loci. Leaderless CRISPR loci have been shown not to undergo adaptation but can still contribute to crRNA-directed interference.

Acknowledgements

The authors thank Anika Erxleben, Björn Grüning and Mummadi Chaithanya Kumar for their help.

Funding

This work was funded by the German Research Foundation (DFG) program FOR1680 'Unravelling the Prokaryotic Immune System' (grant BA 2168/5-1 to R.B.).

Conflict of Interest: none declared.

References

Alkhnbashi,O.S. *et al.* (2014) CRISPRstrand: predicting repeat orientations to determine the crRNA-encoding strand at CRISPR loci. *Bioinformatics*, **30**, i489–i496. In: *The proceedings of the 13th European Conference on Computational Biology (ECCB) 2014*.

Barrangou,R. and van der Oost, J., eds. (2013) *CRISPR-Cas Systems: RNA-Mediated Adaptive Immunity in Bacteria and Archaea*. Springer Press, Heidelburg, pp. 1–129.

Brouns,S.J.J. *et al.* (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science*, **321**, 960–964.

Costa,F. and Grave,K.D. (2010). Fast neighborhood subgraph pairwise distance kernel. In: *Proceedings of the 26 th International Conference on Machine Learning*. Omnipress, pp. 255–262.

Deng,L. *et al.* (2012) Modulation of CRISPR locus transcription by the repeat-binding protein Cbp1 in Sulfolobus. *Nucleic Acids Res.*, **40**, 2470–2480.

Diez-Villasenor,C. *et al.* (2013) CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I–E variants of *Escherichia coli*. *RNA Biol.*, **10**, 792–802.

Eddy,S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**, e1002195.

Enright,A.J. *et al.* (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.

Erdmann,S. and Garrett,R.A. (2012) Selective and hyperactive uptake of foreign DNA by adaptive immune systems of an archaeon via two distinct mechanisms. *Mol. Microbiol.*, **85**, 1044–1056.

Fonfara,I. *et al.* (2014) Phylogeny of Cas9 determines functional exchangeability of dual-RNA and Cas9 among orthologous type II CRISPR-Cas systems. *Nucleic Acids Res.*, **42**, 2577–2590.

Frasconi,P. *et al.*, (2012). klog: a language for logical and relational learning with kernels. *arXiv preprint arXiv:1205.3981*.

Garrett,R.A. *et al.* (2011) Archaeal CRISPR-based immune systems: exchangeable functional modules. *Trends Microbiol.*, **19**, 549–556.

Garrett,R.A. *et al.* (2015) CRISPR-Cas adaptive immune systems of the sulfolobales: unravelling their complexity and diversity. *Life (Basel)*, **5**, 783–817.

Gudbergsdottir,S. *et al.* (2011) Dynamic properties of the Sulfolobus CRISPR/Cas and CRISPR/Cmr systems when challenged with vector-borne viral and plasmid genes and protospacers. *Mol. Microbiol.*, **79**, 35–49.

Horvath,P. *et al.* (2009) Comparative analysis of CRISPR loci in lactic acid bacteria genomes. *Int. J. Food Microbiol.*, **131**, 62–70.

Hsu,P.D. *et al.* (2014) Development and applications of CRISPR-Cas9 for genome engineering. *Cell*, **157**, 1262–1278.

Hyatt,D. *et al.* (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**, 119.

Jansen,R. *et al.* (2002) Identification of genes that are associated with DNA repeats in prokaryotes. *Mol. Microbiol.*, **43**, 1565–1575.

Katoh,K. *et al.* (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**, 3059–3066.

Lange,S.J. *et al.* (2013) CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems. *Nucleic Acids Res.*, **41**, 8034–8044. (SJL, OSA and DR contributed equally to this work.)

Leslie,C. *et al.* (2002) The spectrum kernel: a string kernel for SVM protein classification. *Pac. Symp. Biocomput.*, **2002**, 564–575.

Leslie,C.S. *et al.* (2004) Mismatch string kernels for discriminative protein classification. *Bioinformatics*, **20**, 467–476.

Li,Y. *et al.* (2016) Harnessing Type I and Type III CRISPR-Cas systems for genome editing. **44**, e34.

Lillestol,R.K. *et al.* (2006) A putative viral defence mechanism in archaeal cells. *Archaea*, **2**, 59–72.

Lillestol,R.K. *et al.* (2009) CRISPR families of the crenarchaeal genus Sulfolobus: bidirectional transcription and dynamic properties. *Mol. Microbiol.*, **72**, 259–272.

Makarova,K.S. *et al.* (2015) An updated evolutionary classification of CRISPR-Cas systems. *Nat. Rev. Microbiol.*, **13**, 722–736.

Marraffini,L.A. and Sontheimer,E.J. (2008) CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science*, **322**, 1843–1845.

Mojica,F.J. *et al.* (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. and Garrett,R.A. (2013). Discovery and seminal developments in the CRISPR field. In: *CRISPR-Cas Systems*. Springer, Berlin, Heidelberg, pp. 1–31.

Mojica,F.J.M. *et al.* (2009) Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology*, **155**, 733–740.

Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. **48**, 443–453.

Nickel,L. *et al*. (2013) Two CRISPR-Cas systems in *Methanosarcina mazei* strain Go1 display common processing features despite belonging to different types I and III. *RNA Biol*., **10**, 779–791.

Rand,W.M. (1971) Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc*., **66**, 846–850.

Rollie,C. *et al*. (2015) Intrinsic sequence specificity of the Cas1 integrase directs new spacer acquisition. *Elife*, **4**,

Scholz,I. *et al*. (2013) CRISPR-Cas systems in the *Cyanobacterium synechocystis* sp. PCC6803 exhibit distinct processing pathways involving at least two Cas6 and a Cmr2 protein. *PLoS One*, **8**, e56470. (IS and SJL contributed equally to this work.)

Shah,S.A. and Garrett,R.A. (2011) CRISPR/Cas and Cmr modules, mobility and evolution of adaptive immune systems. *Res. Microbiol*., **162**, 27–38.

Shah,S.A. *et al*. (2009) Distribution of CRISPR spacer matches in viruses and plasmids of *Crenarchaeal acidothermophiles* and implications for their inhibitory mechanism. *Biochem. Soc. Trans*., **37**, 23–28.

Tasaki,E. *et al*. (2012) Molecular identification and characterization of clustered regularly interspaced short palindromic repeats (CRISPRs) in a urease-positive thermophilic *Campylobacter* sp. (UPTC). *World J. Microbiol. Biotechnol*., **28**, 713–720.

Vestergaard,G. *et al*. (2014) CRISPR adaptive immune systems of Archaea. *RNA Biol*., **11**, 157–168.

Wurtzel,O. *et al*. (2010) A single-base resolution map of an archaeal transcriptome. *Genome Res*., **20**, 133–141.

Yosef,I. *et al*. (2012) Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Res*., **40**, 5569–5576.