Conserved DNA motifs in the type II-A CRISPR leader region

Mason J Van Orden 1, Peter Klein 1, Kesavan Babu 1, Fares Z Najar 1, Rakhi Rajan Correspond.

1 Department of Chemistry and Biochemistry, University of Oklahoma, Norman, OK, USA

Corresponding Author: Rakhi Rajan
Email address: r-rajan@ou.edu

The CRISPR-Cas systems consist of RNA-protein complexes that provide bacteria and archaea with sequence-specific immunity against bacteriophages, plasmids, and other mobile genetic elements. Bacteria and archaea become immune to phage or plasmid infections by inserting short pieces of the intruder DNA (spacer) site-specifically into the leader-repeat junction in a process called adaptation. Previous studies have shown that parts of the leader region, especially the 3’ end of the leader, are indispensable for adaptation. However, a comprehensive analysis of leader ends remains absent. Here, we have analyzed the leader, repeat, and Cas proteins from 167 type II-A CRISPR loci. Our results indicate two distinct conserved DNA motifs at the 3’ leader end; ATTTGAG (noted previously in the CRISPR1 locus of Streptococcus thermophilus DGCC7710) and a newly defined CTRCGAG, associated with the CRISPR3 locus of S. thermophilus DGCC7710. A third group with a very short CG DNA conservation at the 3’ leader end is observed mostly in Lactobacilli. Analysis of the repeats and Cas proteins revealed clustering of these CRISPR components that mirrors the leader motif clustering, in agreement with the coevolution of CRISPR-Cas components. Based on our analysis of the type II-A CRISPR loci, we implicate leader end sequences that could confer site-specificity for the adaptation-machinery in the different subsets of type II-A CRISPR loci.
Conserved DNA motifs in the type II-A CRISPR leader region

Mason J. Van Orden#, Peter Klein#, Kesavan Babu1, Fares Z. Najar1, and Rakhi Rajan*

1Department of Chemistry and Biochemistry, University of Oklahoma, Norman, OK, USA

#Equal first authors

*Correspondence: r-rajan@ou.edu
Abstract

The CRISPR-Cas systems consist of RNA-protein complexes that provide bacteria and archaebacteria with sequence-specific immunity against bacteriophages, plasmids, and other mobile genetic elements. Bacteria and archaebacteria become immune to phage or plasmid infections by inserting short pieces of the intruder DNA (spacer) site-specifically into the leader-repeat junction in a process called adaptation. Previous studies have shown that parts of the leader region, especially the 3' end of the leader, are indispensable for adaptation. However, a comprehensive analysis of leader ends remains absent. Here, we have analyzed the leader, repeat, and Cas proteins from 167 type II-A CRISPR loci. Our results indicate two distinct conserved DNA motifs at the 3' leader end; ATTTGAG (noted previously in the CRISPR1 locus of Streptococcus thermophilus DGCC7710) and a newly defined CTRCGAG, associated with the CRISPR3 locus of S. thermophilus DGCC7710. A third group with a very short CG DNA conservation at the 3' leader end is observed mostly in Lactobacilli. Analysis of the repeats and Cas proteins revealed clustering of these CRISPR components that mirrors the leader motif clustering, in agreement with the coevolution of CRISPR-Cas components. Based on our analysis of the type II-A CRISPR loci, we implicate leader end sequences that could confer site-specificity for the adaptation-machinery in the different subsets of type II-A CRISPR loci.
Introduction

The Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) and CRISPR-associated (Cas) proteins constitute an RNA-based adaptive immune system that protects bacteria and archaea against phages and mobile genetic elements\textsuperscript{1-5}. CRISPR-Cas systems inactivate intruder DNA, RNA, or both based on the sequence similarity of small CRISPR RNAs (crRNAs) to the invading genetic element and thus protect microbes from phage infections and horizontal gene transfer\textsuperscript{2,8-11}. The CRISPR-Cas systems are classified into six types (I to VI) with several subtypes within each type based on Cas protein composition\textsuperscript{11-13}. An individual bacterium can have multiple CRISPR loci belonging to different CRISPR types. Though types I and III share certain similarities in the overall mechanism of action including crRNA association with multiple Cas proteins\textsuperscript{14-18}, types II, V, and VI use a single multi-domain protein (Cas9, Cpf1, or C2c2 respectively) along with cognate RNA components for activity\textsuperscript{11,19-22}. Cas9 along with a guide RNA is widely being used for genome editing applications, and is being pursued for gene therapy and gene regulation applications\textsuperscript{23,24}.

The CRISPR genomic locus is present in either the chromosomal or plasmid DNA as a recurring array of “repeat” and “spacer” units, both of which usually range from approximately 24 to 50 nucleotides\textsuperscript{2,3,25,26}. Repeats consist of palindromic DNA sequences, while spacers are derived from invader genetic material, and experimental evidence supports the role of spacers in conferring sequence-specific resistance against bacteriophages\textsuperscript{4,9} and plasmids\textsuperscript{8}. The cas genes that code for the CRISPR system’s essential protein components are often located close to the CRISPR locus\textsuperscript{26,27}. The CRISPR leader located 5’ to the first repeat consists of an A/T-rich region around 100-500 nucleotides long with embedded transcriptional promoters\textsuperscript{28-30}. In certain CRISPR types, a 3-5 nucleotide long region present in the invading DNA (Protospacer Adjacent Motif (PAM)) is crucial for Cas proteins to differentiate self from non-self DNA\textsuperscript{2,31,32}, where as in other types this is defined by the crRNA-DNA base pairing patterns\textsuperscript{7}. 
There are three stages in CRISPR mediated defense: adaptation (acquiring new spacers), crRNA biogenesis, and interference\textsuperscript{2,33}. The adaptation process differs between CRISPR types. The proteins, Cas1 and Cas2, are universally present and essential for adaptation in most of the CRISPR types\textsuperscript{30,34}. In certain type II subtypes, Csn2 and Cas4 are also implicated in adaptation along with Cas1 and Cas2\textsuperscript{4,35}. The new spacers are inserted at the leader-repeat junction in most systems, although some variability have been observed in *Sulfolobus solfataricus*\textsuperscript{36}. A minimum of one repeat along with the leader region can promote spacer insertion and in certain CRISPR subtypes, one of the strands of the double-stranded intruding DNA is preferred for spacer acquisition\textsuperscript{29,37}. A new spacer insertion is always accompanied by repeat duplication and the first repeat serves as a template for new repeat synthesis\textsuperscript{30}. Even though recognition of the PAM sequence by Cas9 is essential for acquisition and insertion of spacers in the correct orientation *in vivo* in type II CRISPR systems\textsuperscript{38-40}, it was recently demonstrated that Cas1 and Cas2 from *Streptococcus pyogenes* can specifically integrate spacers into the leader-repeat junction based solely on intrinsic sequence specificity of the first repeat\textsuperscript{41}.

*Streptococcus thermophilus* (Sth) DGCC7710 is a model organism widely used for studying various CRISPR processes. Sth DGCC7710 has a total of four CRISPR loci in its genome, and loci CRISPR1 and CRISPR3 belong to type II-A\textsuperscript{42-44}. It was experimentally shown that in Sth DGCC7710, both CRISPR1 and CRISPR3 are active in acquiring new spacers in relation to a new phage or plasmid threat, with CRISPR1 being more active due to the higher frequency of new spacer insertion in this locus compared to CRISPR\textsuperscript{34,29,35,40,42,44,47}. A phylogenetic analysis showed that the repeat and *cas* genes segregate specifically with the locus type (CRISPR1 vs CRISPR2 vs CRISP3) in streptococci and in several bacteria belonging to different genera\textsuperscript{46}.

Several studies pointed to the indispensability of the 3’ end of the leader in CRISPR adaptation\textsuperscript{29,30,36,48-53}. In the CRISPR1 locus of Sth DGCC7710 (type II-A), a *cis*-acting element at the 3’ end of the leader (ATTTGAG) was shown to be essential for adaptation and this region is conserved in several type II-A systems\textsuperscript{29}. In *Escherichia coli* BL21 strain, a type I-E CRISPR locus
showed defective adaptation following deletions or mutation in the 60 nucleotides towards the 3’
end of the CRISPR leader. A pair of inverted repeat regions in the first repeat along with the
leader end sequence is critical for adaptation in the type IB system of *Haloarcula hispanica*.
Recently, a study in type I CRISPR systems identified leader DNA sequences that are specifically
recognized by the integration host factor (IHF) protein to facilitate leader-proximal spacer
integration under *in vitro* conditions. It was later shown, however, that under *in vitro* conditions
type II systems integrate spacers specifically into the leader-proximal regions by Cas1-Cas2
activity alone, without the participation of another protein. This highlights the differences in the
mechanism of spacer integration between CRISPR types and the possibility of divergent
contributions of the leader and repeat sequences in type-specific adaptation.

In order to identify the leader and repeat DNA sequence conservations that may contribute to
site-specific spacer integration in type II-A CRISPR systems, we report the analysis of the leader-
repeat region belonging to 167 type II-A CRISPR loci from 50 different genera. 87 of the 167 loci
have the 3’ leader end conserved as ATTTGAG (Group 1), 55/167 loci have their 3’ leader end
conserved as CTRCGAG (Group 2), and 25/167 possess a CG conservation at the 3’ end of the
leader (Group 3). Previous studies that established the importance of ATTTGAG and ACGAG
leader end sequences in adaptation of *Sth DGCC7710* and *S. pyogenes*, respectively, point to
the functional significance of the conserved DNA motifs. A detailed analysis of the Cas proteins
associated with the 167 type II-A loci shows protein sequence specificities that delineate these
proteins into groups that mirror the leader end conservation. Thus, our study establishes distinct
sub-group specific DNA sequence conservation patterns in the type II-A CRISPR leader that
extends across many diverse bacteria demonstrating the ubiquitous nature of the 3’-leader end
conservations that were previously observed only in related streptococcal species.
Methods

Processing of genomic data

In this study, the type II-A loci were collected by multiple ways. Initially, Bacterial Generic Feature Format (GFF) and accessioned protein product FASTA files were downloaded from NCBI and scanned for II-A specific Cas protein names (Cas9/Csn1 and Csn2) in the annotation field. The genomes containing Cas9/Csn1 and/or Csn2 annotation entries were downloaded from NCBI in GenBank format. The datasets were screened manually for the presence of cas1, cas2, cas9, and csn2, and only the loci with all four type II-A specific cas genes were used for further analysis.

The genomic region flanking downstream of the csn2 gene was further processed to extract the leader and the first repeat of the CRISPR array. The protein sequences of Cas9, Cas1, Cas2, and Csn2 that were coded by the upstream region flanking the csn2 gene were extracted from NCBI. The presence of all four proteins limits our dataset to strictly type II-A loci. A total of 129 loci were identified based on Cas9/Csn1 and/or Csn2 annotation search. Previously, Chylinski et al reported type II-A loci based on a Cas9 sequence search56,57. A total of 32 type II-A loci that represented species and genera that were absent in our initial dataset were selected for our study.

In addition, we performed protein sequence homology search by DELTA-BLAST58 using a representative Csn2 sequence from each subfamily as mentioned in Chylinski et al201456 ((NCBI protein accession number: 116101487 for subfamily I, 116100822 for subfamily II, 389815356 for subfamily III, 385326557 for subfamily IV, 315659845 for subfamily V as mentioned in Chylinski et al 2014)56). By this search a total of 6 loci were identified from bacterial genera Weissella, Globicatella, Nosocomiicoccus, Caryophanon and Virgibacillus. The final dataset consisted of 167 type II-A loci with a wide representation based on the current knowledge of type II-A diversity.

A total of 50 different bacterial genera were present in our dataset. (Table S1 and S2).

The orientation of the Cas proteins was used in assessing the transcription direction of the leader-repeat units. To analyze leader and repeat sequences, an approximately 400-nucleotide stretch
of sequences downstream of $csn2$ gene were examined using CRISPR finder tool\textsuperscript{59}, and an in-house script to locate the tandem repeats. Since there were differences in the repeat length as it exists in the genomic locus and as reported in the CRISPRdb\textsuperscript{59}, we used the in-house program to locate the repeats (Table S3). The accuracy of the repeat extracted by our script was validated manually by checking the genomic data for the length and sequence of the repeat within a CRISPR array. The loci that lacked predicted repeats or Cas protein(s) were omitted from further analysis. In the case of bacteria with multiple CRISPR types, the components belonging to a particular type II-A locus were taken as one dataset. For example, Sth DGCC7710 has four CRISPR loci. Only loci 1 and 3 that correspond to type II-A were selected for our analysis. The Cas proteins and leader-repeat elements of CRISPR1 were kept as one unit, while that belonging to CRISPR3 represented another unit. Recently, several bioinformatics tools for the identification and analysis of leader and repeat regions have been developed\textsuperscript{60,61}. For a selected subset, we compared the orientation of leader sequences and repeats as predicted by CRISPRDetect tool\textsuperscript{61} and our results, and saw agreement between the methods.

**Sequence Alignment**

We used MUSCLE with its default settings\textsuperscript{62} to perform all the sequence alignments in this study. The MUSCLE output was used to generate phylogenetic trees with MEGA6\textsuperscript{63} using the Maximum Likelihood Tree option and Jones-Taylor-Thornton (JTT) model. Additionally, MUSCLE alignments were used to generate alignment figures in UGENE\textsuperscript{64} and sequence logos with WebLogo\textsuperscript{65}.

**Results**

**Analysis of the 3’ end of the leader**

An initial sequence alignment of the last 20 nucleotides of the leader plus the first repeat showed that the 167 loci clustered into distinct groups. These groups had recognizable conservation at...
the last 7 nucleotides of the 3' end of the leader and the first 4 nucleotides of the 5' end of the first
repeat, or the leader-repeat junction. To obtain an unbiased separation of the different groups, a
Cas1 phylogenetic tree was constructed based on protein sequence similarity. The loci belonging
to the different clades of the Cas1 tree were grouped together and a sequence alignment of the
last 20 nucleotides of the leader along with the first repeat was performed. In order to facilitate
interpretation of the trees and alignments, a smaller representative sample of 62 loci was used to
generate the main figures and show the relevant relationships. Figures incorporating all of the
167 loci can be found in the supplementary data. Each of the 3 groups was aligned separately to
discern the level of conservation within each group (Fig. 1 and 2, Fig. S1). Strict conservation is
seen at the 3’ end of the leader as well as at the 5’ end of the repeat. Group 1 has the 3’ leader
eンド conserved as ATTTGAG (Fig. 1) and Group 2 has the 3’-leader end conserved as CTRCGAG
(where R represents a purine) (Fig. 2a). Group 3 has a shorter two nucleotide conservation of CG
at the 3’ leader end. In Groups 1 and 2, the last three nucleotides are conserved as GAG (Fig.
2b). An A-rich region is partially conserved adjacent to the CG leader end of Group 3. Interestingly,
the CRISPR1 locus of Sth DGCC7710 has the 3’ leader end conserved as ATTTGAG while the
CRISPR3 locus has the 3’ leader end conserved as CTACGAG. Of the type II-A CRISPR loci
analyzed, 87 belonged to Group 1, 55 belonged to Group 2, and 25 belonged to Group 3. Out of
the 50 genera analyzed, Group 2 consists of only 5 genera (Streptococcus, Enterococcus,
Listeria, Lactobacillus and Weissella) while Group 1 is much more diverse with 42 different
genera. Group 3 accounts for 7 genera, but has many loci belonging to the Order Lactobacillales.
The leader-repeat junction of Groups 1 and 2 is conserved as GAG/GTTT while in Group 3 it is
weakly conserved as CG/GTTT.

Analysis of the repeat region

The length of the repeat for the type II-A loci analyzed was 36 nucleotides except in 4 cases
(Enterococcus hirae ATCC 9790 (35 nucleotides long), Fusobacterium sp. 1_1_41FAA (37
nucleotides long), *Lactobacillus coryniformis* subsp. coryniformis KCTC 3167 (37 nucleotides long), and *Lactobacillus sanfranciscensis* TMW 1.1304 (35 nucleotides long). The first repeat sequences of the 3 groups didn’t possess any distinguishable motifs that corresponded to the segregation of the different groups (Fig. 3, Fig. S2). There is a strong sequence conservation at the 5’ end of the repeat as GTTT in all the type II-A loci analyzed (Fig. 3 and Fig. S2). Groups 1 and 2 also share a conserved AAAC motif at the 3’ end of the repeat. Group 3 members have a conserved C at the 3’ end of the repeat, along with a less conserved A-rich region ahead of the C. The repeat sequence belonging to the Group 2 loci is highly conserved across the entire length of the repeat, which may be attributed to the limited number of genera (5) comprising this group compared to Group 1 (42). In all the type II-A loci analyzed, the first and last nucleotides of the first repeat are conserved as G and C respectively. A phylogenetic tree was generated using the first repeat sequence of the type II-A loci (Fig. 3 and Fig. S2). Even though the reliability of branching is low due to the short length of the sequence, the branches segregate such that members within a clade have similar repeat and leader end conservations. Recently, it was suggested that sequences at the 5’ and 3’ ends of the repeat in *S. pyogenes* type II-A system could be the motifs recognized by Cas1 during spacer acquisition.\(^1\) Hence, the conserved 5’ and 3’ repeat ends observed in the first two groups might indicate type II-A specific repeat ends that are essential for adaptation. Further experimental studies will be required to analyze whether the loosely conserved sequences at the 3’ end of the repeat impact effective adaptation in Group 3. The similarity at the 5’ and 3’ ends of the repeat in the different sub-groups of type II-A system and the fact that exchanging leader ends between CRISPR1 and CRISPR3 loci in Sth DGCC7710\(^2\) impaired adaptation shows that the specificity within the sub-groups of type II-A CRISPR system is most probably attributed by the 3’ leader end and not specified by the repeat ends.

**Analysis of Cas proteins**
We extended our analysis to verify whether the different groups of type II-A CRISPR loci observed based on the 3' leader end conservation relates to Cas proteins. The protein sequences of Cas1 belonging to the selected type II-A loci were aligned by MUSCLE and a phylogenetic tree was generated (Fig. 4 and Fig. S3). The loci segregated into 4 main branches, with each branch carrying distinct groups based on the 3' leader end sequence conservation. A sequence alignment of the leader-repeat junction of the different branches show how the Cas1 sequence is highly correlated with the leader-repeat junction. This confirms previous findings that all the CRISPR-Cas components have coevolved together.\(^{46}\) The phylogenetic tree shows that Group 1 loci are very distant in lineage, which has later evolved into different subsets with specific leader-repeat-Cas1 combinations. Group 2 and Group 3 evolved for very specific genera, while Group 1 has accommodated divergent genera.

A similar analysis was done for the Cas2, Csn2, and Cas9 proteins. The sequence alignments generated using the sequences of the corresponding Cas proteins were used to build phylogenetic trees (Fig. 5, 6 and Fig. S4 and S5). All the clades in the different trees have similar 3'-leader ends, except for a few differences in the Cas9 phylogenetic tree where some Group 3 members appeared along with Group 1. A closer analysis of the sequences showed high variability in the Cas9 lengths, including an extremely short Cas9 sequence (Plo NGRI0510Q) in the outliers, which may have contributed to the random placement of this Cas9 protein. Cas9 also showed a branch (1b) for Group 1 that did not show prominent leader end conservation as that was observed in branch 1a. Except for the few differences in Cas9, our results indicate the presence of distinct groups within the type II-A CRISPR systems that possess conserved 3' leader ends and group-specific Cas proteins. It was proposed earlier that the longer version of Csn2 evolved first and the shorter Csn2 proteins evolved from the longer versions.\(^{46}\) Interestingly, our phylogenetic analysis agrees with this and shows a branch that represents the ancestor with an average Csn2 length of 320 amino acids (Fig. 6). Three main branches evolved from the ancestor and all of them have an average amino acid length of 218-230 amino acids, but varying 3' leader
ends (Table S4). Thus, the ATTTGAG motif is ancestral and universal in the type II-A systems, which later developed to have a similar (ATTTGAG), deviating (CTRCGAG), or less conserved (CG) 3’ leader end, with a corresponding change in the protein sequences of all four type II-A Cas proteins. Examining the lengths of Cas1, Cas2, and Cas9 from different groups, we did not observe a strong correlation between the average length of these Cas proteins and the branching group that they belonged.

**Discussion**

Though previous studies have shown that the leader-repeat region is important for adaptation, the specific features of the leader-repeat region that may recruit Cas1-Cas2 for adaptation are not clearly defined. We focused on the sequence conservation around the leader-repeat junction and found three distinct DNA motifs at the 3’ leader ends; Group 1 (ATTTGAG), Group 2 (CTRCGAG), and Group 3 (CG). The presence of a conserved 3’ leader end, despite a low sequence conservation in the upstream regions of the leader in bacteria belonging to 50 different unrelated genera, strongly suggests that these DNA motifs play a role in site-specific adaptation. One of the most interesting observations from this analysis is the conservation of GAGGTTT as the leader-repeat junction in both Group 1 and Group 2 (82%, 117 out of 142 loci) of the type II-A system.

Several studies have implicated the importance of the leader and repeat sequences to drive faithful adaptation. Terns and coworkers reported that streptococci with repeats similar to that present in the CRISPR1 locus (Group 1) of Sth DGCC7710 have the 3’ leader end conserved as ATTTGAG. The accompanying experimental work clearly demonstrated that the 10 nucleotides present at the 3’ end of the leader and the first repeat are essential and sufficient for adaptation, even in a non-CRISPR locus\textsuperscript{29}. It was concluded that sequences at the leader-repeat junction recruits the adaptation machinery to this region for integration of new spacers\textsuperscript{29}. In a recent study that analyzed the spacer variation in 126 human isolates of *S. agalactiae*, the 3’ leader end of
most of the isolates had a TACGAG sequence. Our analysis that focused on many divergent
genera uncovered that the DNA motifs that were previously known to be important for
streptococcal adaptation is in fact more ubiquitous and conserved across different bacteria.
The importance of the sequences of the leader and the first repeat in driving adaptation is
conserved across different CRISPR types. The 60 nucleotides towards the 3’ end of the type I-E
CRISPR locus of _Escherichia coli_ is essential for adaptation. The disruption of the first repeat
sequence that left the stem-loop structure intact prevented successful adaptation in a type IE
system leading to the conclusion that the cruciform structure of the repeat alone is not sufficient
for adaptation. Another study showed that the -2 (second last position of leader) and +1 (first
nucleotide of repeat) positions of leader-repeat regions are crucial for adaptation in _E. coli_ (type
IE) and _Sulfolobus solfataricus_ (type I-A). Other studies have experimentally demonstrated that
leader and repeat sequences are important for adaptation in streptococcal type II-A systems
corresponding to the groups 1 and 2 that we identified in our study. Comparing our results
with the earlier studies show that leader-repeat sequence conservation that we observed in type
II-A sub-groups is relevant for adaptation across diverse bacteria.

There is an interplay between the leader and repeat sequences in adaptation that is CRISPR
type-specific. For example, in the type I-B system of _Haloarcula hispanica_, inverted repeats (IR)
present within the first repeat are essential for recruiting the adaptation machinery to the leader-
repeat junction. Once the IRs are located within a repeat, a cut is made by the Cas1-Cas2
complex at the leader-repeat junction and the sequence of the leader is critical for this step. The
second cut at the repeat-spacer end is based on a ruler-mechanism and does not depend on the
sequence of the repeat. Whereas in a type II-A system corresponding to our Group 2, it was
shown that the repeat-spacer and repeat-leader ends have the same probability of getting cleaved
by Cas1-Cas2, but for a faithful adaptation, the leader-repeat junction is essential. In the Group
1 type II-A locus of Sth DGCC7710, mutations in the last 10 nucleotides of the leader abolished
adaptation. This study also elegantly showed that substitution of the 10 nucleotides at the 3’ end
of Group 1 leader with that of Group 2 leader abolished adaptation following a phage challenge, further emphasizing the importance of the locus specific leader-repeat junction in adaptation. Thus, our observation of the group-specific sequence conservation in type II-A systems at the leader end, along with a lack of distinct group-specific motifs at the 5’ and 3’ ends of the first repeat, shows that the sub-group specificity in type II-A adaptation arises from the leader sequences that might be specifically distinguished by the Cas1-Cas2 proteins belonging to each sub-group.

Both groups 1 and 2 are active for adaptation and interference, while Group 3 has been shown to be active in DNA interference. Introduction of the type II-A Group 2 locus into *E. coli* protected the bacterium from phage and plasmid infection, demonstrating that intrinsic specificities of protein and DNA components of a CRISPR sub-type are sufficient to drive adaptation and there are no organismal requirements. The three different DNA motifs that we observed at the 3’-end of the leader of the type II-A CRISPR loci may represent three specific functional adaptation units, perhaps guided by leader-sequence specific Cas protein(s). The third group, which consists mostly of lactobacilli, with only two nucleotides conserved instead of seven nucleotides at the 3’ leader end in Groups 1 and 2 may represent a more diverse adaptation complex where the protein-DNA sequence interactions are not as tight. It was noted recently that there is considerable variation in the spacer content, even in the ancestral spacers, in *L. gasseri* strains that indicates considerable divergence between the strains, thus accounting for the low level of sequence conservation at the 3’-end of the leader. This study also showed that the spacers matched plasmids and temperate phages, though it is not clear how *L. gasseri* acquires spacers from prophages that do not pose threat to bacterial survival. These environmental factors may contribute to the low sequence conservation at the 3’-end of the leader in Group 3. Further experiments will be required to assess the adaptation process in Group 3. Group 3 could also be a result of an insufficient amount of genomic data available to completely resolve any more conserved motifs hidden in the different leader end sequences found within the group.
Repeat sequences are specific to a CRISPR locus, even within sub-types\textsuperscript{46,56}. The first two nucleotides of the first repeat was shown to be essential for adaptation in the CRISPR1 locus of Sth\textsuperscript{29} and the first six nucleotides are essential in adaptation in \textit{S. pyogenes}\textsuperscript{41}. The importance of G as the first nucleotide in the repeat for efficient disintegration reaction was demonstrated for both \textit{E. coli} and \textit{S. solfataricus} Cas1 proteins\textsuperscript{71}. We found conservation at the ends of the repeat between groups (Fig. 3). Only 17/167 loci analyzed did not possess a GTTT at the 5’-end of the first repeat, and 3/167 of the loci did not possess a conserved C at the 3’ end. It was previously reported that purified \textit{E. coli} Cas1 possesses nuclease activity against several types of DNA substrates including single stranded DNA, replication forks, Holliday junctions \textit{etc}. without adequate intrinsic sequence specificity and that the four-way DNA junctions recruits Cas1 protein\textsuperscript{72}. Recently, more studies point to the importance of DNA sequence specificity, especially at the 3’-end of the leader, for driving Cas1 for adaptation\textsuperscript{29,54,71}. The essentiality of IHF for site-specific adaptation in type – indicates that even though Cas1 may have the ability of non-sequence specific cleavage in certain CRISPR types, tight regulation by other cellular proteins may enhance site-specific spacer insertion. The position of the IHF site is 9 to 35 nucleotides upstream of the leader-repeat junction in type I systems\textsuperscript{55}. The 20 nucleotides of the 3’ leader end that we analyzed for the type II-A did not possess any similarity to the IHF binding site. It is possible that a cruciform structure formed by leader-repeat or repeat palindromic regions along with specific leader-repeat sequences may recruit the Cas1-Cas2 complex for spacer insertion and that this requirement is critical under \textit{in vivo} conditions.

All four Cas proteins are essential for successful adaptation \textit{in vivo} in type II-A systems\textsuperscript{4,30,39}. Previous studies have shown that the CRISPR components and Cas proteins have coevolved\textsuperscript{46}. Our analysis showed that all the four type II-A specific Cas proteins and the first repeat clustered into identical groups with similar 3’ leader ends. Even though Cas1 protein sequences within type II-A are highly conserved, there are certain differences that segregate them into distinct groups and interestingly these groups have distinct leader sequence conservation. It was previously
reported that type II-A CRISPR systems have distinct operon organization that correlates with Csn2 sequence, making Csn2 the signature protein for type II-A systems\textsuperscript{56}. The longer version of Csn2 originated first and the shorter version evolved from the longer version\textsuperscript{56}. Our analysis shows that the length of Csn2 is conserved across different clusters (Fig. 5). Looking at Fig. 6, branch 1a segregated early from the rest of the tree and consists of the longer version of Csn2, while branches 1b, 2, and 3 all consist of the shorter version of Csn2. Correlating Csn2 branching to the leader end sequences, it is evident that our Group 1 motif of ATTTGAG is present in the ancestral strains, which later evolved to distinct sub-groups possessing either Group 1, Group 2 (CTRCGAG) or Group 3 (CG) leader ends.

Conclusion

We present an extensive bioinformatic analysis of type II-A CRISPR systems spanning 50 different bacterial genera. We demonstrated the ubiquitous nature of two distinct DNA motifs at the 3’ end of the leader: Group 1 (ATTTGAG) and Group 2 (CTRCGAG) and also discovered a new group (Group 3) with a limited sequence conservation at the 3’-end of the leader. The leader-repeat junction is highly conserved for Groups 1 and 2 as GAGGTTT. Our work proposes that the Cas proteins of each sub-group within the type II-A system should make sequence-specific association with its cognate DNA region for successful spacer insertion. The observations further strengthen the previous notion that a highly specific interplay between Cas proteins and cognate leader-repeat regions is essential for effective adaptation\textsuperscript{29,30,37,67,68}.

Acknowledgements

We thank Sungho Suh for help with genomic data collection and processing.
References

1. Marraffini, L.A. & Sontheimer, E.J. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat Rev Genet* **11**, 181-90 (2010).

2. Sorek, R., Lawrence, C.M. & Wiedenheft, B. CRISPR-mediated adaptive immune systems in bacteria and archaea. *Annu Rev Biochem* **82**, 237-66 (2013).

3. Mojica, F.J., Diez-Villasenor, C., Garcia-Martinez, J. & Soria, E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol* **60**, 174-82 (2005).

4. Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A. & Horvath, P. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709-12 (2007).

5. Makarova, K.S., Grishin, N.V., Shabalina, S.A., Wolf, Y.I. & Koonin, E.V. 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 (2006).

6. Marraffini, L.A. & Sontheimer, E.J. Invasive DNA, chopped and in the CRISPR. *Structure* **17**, 786-8 (2009).

7. Marraffini, L.A. & Sontheimer, E.J. Self versus non-self discrimination during CRISPR RNA-directed immunity. *Nature* **463**, 568-71 (2010).

8. Marraffini, L.A. & Sontheimer, E.J. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* **322**, 1843-5 (2008).

9. Brouns, S.J., Jore, M.M., Lundgren, M., Westra, E.R., Slijkhuis, R.J., Snijders, A.P., Dickman, M.J., Makarova, K.S., Koonin, E.V. & van der Oost, J. Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**, 960-4 (2008).

10. Hale, C.R., Zhao, P., Olson, S., Duff, M.O., Graveley, B.R., Wells, L., Terns, R.M. & Terns, M.P. RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell* **139**, 945-56 (2009).

11. Abudayyeh, O.O., Gootenberg, J.S., Konermann, S., Joung, J., Slaymaker, I.M., Cox, D.B., Shmakov, S., Makarova, K.S., Semenova, E., Minakhin, L., Severinov, K., Regev, A., Lander, E.S., Koonin, E.V. & Zhang, F. C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science* (2016).

12. Makarova, K.S. & Koonin, E.V. Annotation and Classification of CRISPR-Cas Systems. *Methods Mol Biol* **1311**, 47-75 (2015).

13. Makarova, K.S., Wolf, Y.I., Alkhnbashi, O.S., Costa, F., Shah, S.A., Saunders, S.J., Barrangou, R., Brouns, S.J.J., Charpentier, E., Haft, D.H., Horvath, P., Moineau, S., Mojica, F.J.M., Terns, R.M., Terns, M.P., White, M.F., Yakunin, A.F., Garrett, R.A., van der Oost, J., Backofen, R. & Koonin, E.V. An updated evolutionary classification of CRISPR-Cas systems. *Nat Rev Micro* **13**, 722-736 (2015).

14. Jackson, R.N. & Wiedenheft, B. A Conserved Structural Chassis for Mounting Versatile CRISPR RNA-Guided Immune Responses. *Mol Cell* **58**, 722-8 (2015).

15. Reeks, J., Naismith, J.H. & White, M.F. CRISPR interference: a structural perspective. *Biochem* **453**, 155-66 (2013).
16. Makarova, K.S., Aravind, L., Wolf, Y.I. & Koonin, E.V. Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. *Biol Direct* 6, 38 (2011).

17. Koonin, E.V. & Makarova, K.S. CRISPR-Cas: evolution of an RNA-based adaptive immunity system in prokaryotes. *RNA Biol* 10, 679-86 (2013).

18. Shmakov, S., Abudayyeh, O.O., Makarova, K.S., Wolf, Y.I., Gootenberg, J.S., Semenova, E., Minakhin, L., Joung, J., Konermann, S., Severinov, K., Zhang, F. & Koonin, E.V. Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems. *Mol Cell* (2015).

19. Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A. & Charpentier, E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* 337, 816-21 (2012).

20. Jinek, M., Jiang, F., Taylor, D.W., Sternberg, S.H., Kaya, E., Ma, E., Anders, C., Hauer, M., Zhou, K., Lin, S., Kaplan, M., Javarene, A.T., Charpentier, E., Nogales, E. & Doudna, J.A. Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science* 343, 1247997 (2014).

21. Nishimasu, H., Ran, F.A., Hsu, P.D., Konermann, S., Shehata, S.I., Dohmae, N., Ishitani, R., Zhang, F. & Nureki, O. Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* 156, 935-49 (2014).

22. Zetsche, B., Gootenberg, J.S., Abudayyeh, O.O., Slaymaker, I.M., Makarova, K.S., Essletzbichler, P., Volz, S.E., Joung, J., van der Oost, J., Regev, A., Koonin, E.V. & Zhang, F. Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell* 163, 759-71 (2015).

23. Sternberg, S.H. & Doudna, J.A. Expanding the Biologist’s Toolkit with CRISPR-Cas9. *Mol Cell* 58, 568-574 (2015).

24. Sontheimer, E.J. & Barrangou, R. The Bacterial Origins of the CRISPR Genome-Editing Revolution. *Hum Gene Ther* 26, 413-24 (2015).

25. Ishino, Y., Shinagawa, H., Makino, K., Amemura, M. & Nakata, A. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli*, and identification of the gene product. *J Bacteriol* 169, 5429-33 (1987).

26. Jansen, R., Embden, J.D., Gaastra, W. & Schouls, L.M. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol* 43, 1565-75 (2002).

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

28. Pougach, K., Semenova E Fau - Bogdanova, E., Bogdanova E Fau - Datsenko, K.A., Datsenko Ka Fau - Djordjevic, M., Djordjevic M Fau - Wanner, B.L., Wanner Bl Fau - Severinov, K. & Severinov, K. Transcription, processing and function of CRISPR cassettes in *Escherichia coli*. (2010).

29. Wei, Y., Chesne, M.T., Terns, R.M. & Terns, M.P. Sequences spanning the leader-repeat junction mediate CRISPR adaptation to phage in *Streptococcus thermophilus*. *Nucleic Acids Res* 43, 1749-58 (2015).

30. Yosef, I., Goren, M.G. & Qimron, U. Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Res* 40, 5569-76 (2012).
31. Mojica, F.J., Diez-Villasenor, C., Garcia-Martinez, J. & Almendros, C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155**, 733-40 (2009).

32. Shah, S.A., Erdmann, S., Mojica, F.J. & Garrett, R.A. Protospacer recognition motifs: mixed identities and functional diversity. *RNA Biol* **10**, 891-9 (2013).

33. Amitai, G. & Sorek, R. CRISPR-Cas adaptation: insights into the mechanism of action. *Nat Rev Microbiol* **14**, 67-76 (2016).

34. Nunez, J.K., Kranzusch, P.J., Noeske, J., Wright, A.V., Davies, C.W. & Doudna, J.A. Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. *Nat Struct Mol Biol* **21**, 528-34 (2014).

35. Garneau, J.E., Dupuis, M.E., Villion, M., Romero, D.A., Barrangou, R., Boyaval, P., Fremaux, C., Horvath, P., Magadan, A.H. & Moineau, S. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**, 67-71 (2010).

36. Erdmann, S. & Garrett, R.A. Selective and hyperactive uptake of foreign DNA by adaptive immune systems of an archaeon via two distinct mechanisms. *Mol Microbiol* **85**, 1044-56 (2012).

37. Diez-Villasenor, C., Guzman, N.M., Almendros, C., Garcia-Martinez, J. & Mojica, F.J. CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I-E variants of *Escherichia coli*. *RNA Biol* **10**, 792-802 (2013).

38. Wei, Y., Terns, R.M. & Terns, M.P. Cas9 function and host genome sampling in Type II-A CRISPR-Cas adaptation. *Genes Dev* **29**, 356-61 (2015).

39. Heler, R., Samai, P., Modell, J.W., Weiner, C., Goldberg, G.W., Bikard, D. & Marraffini, L.A. Cas9 specifies functional viral targets during CRISPR-Cas adaptation. *Nature* **519**, 199-202 (2015).

40. Paez-Espino, D., Morovic, W., Sun, C.L., Thomas, B.C., Ueda, K., Stahl, B., Barrangou, R. & Banfield, J.F. Strong bias in the bacterial CRISPR elements that confer immunity to phage. *Nat Commun* **4**, 1430 (2013).

41. Wright, A.V. & Doudna, J.A. Protecting genome integrity during CRISPR immune adaptation. LID - 10.1038/nsmb.3289 [doi]. *Nat Struct Mol Biol* (2016).

42. Sapranaukas, R., Gasiunas, G., Fremaux, C., Barrangou, R., Horvath, P. & Siksnys, V. The *Streptococcus thermophilus* CRISPR/Cas system provides immunity in *Escherichia coli*. *Nucleic Acids Res* **39**, 9275-82 (2011).

43. Horvath, P. & Barrangou, R. CRISPR/Cas, the immune system of bacteria and archaea. *Science* **327**, 167-70 (2010).

44. Carte, J., Christopher, R.T., Smith, J.T., Olson, S., Barrangou, R., Moineau, S., Glover, C.V., 3rd, Graveley, B.R., Terns, R.M. & Terns, M.P. The three major types of CRISPR-Cas systems function independently in CRISPR RNA biogenesis in *Streptococcus thermophilus*. *Mol Microbiol* **93**, 98-112 (2014).

45. Deveau, H., Barrangou, R., Garneau, J.E, Labonte, J., Fremaux, C., Boyaval, P., Romero, D.A., Horvath, P. & Moineau, S. Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *J Bacteriol* **190**, 1390-400 (2008).

46. Horvath, P., Romero, D.A., Coute-Monvoisin, A.C., Richards, M., Deveau, H., Moineau, S., Boyaval, P., Fremaux, C. & Barrangou, R. Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. *J Bacteriol* **190**, 1401-12 (2008).
498 47. Lopez-Sanchez, M.J., Sauvage, E., Da Cunha, V., Clermont, D., Ratsima Hariniaina, E., Gonzalez-Zorn, B., Poyart, C., Rosinski-Chupin, I. & Glaser, P. The highly dynamic CRISPR1 system of Streptococcus agalactiae controls the diversity of its mobilome. *Mol Microbiol* **85**, 1057-71 (2012).

499 48. Yosef, I., Shitrit, D., Goren, M.G., Burstein, D., Pupko, T. & Qimron, U. DNA motifs determining the efficiency of adaptation into the Escherichia coli CRISPR array. *Proc Natl Acad Sci U S A* **110**, 14396-401 (2013).

500 49. Jansen, R., van Embden, J.D., Gaastra, W. & Schouls, L.M. Identification of a novel family of sequence repeats among prokaryotes. *OMICS* **6**, 23-33 (2002).

501 50. Li, M., Wang, R., Zhao, D. & Xiang, H. Adaptation of the Haloarcula hispanica CRISPR-Cas system to a purified virus strictly requires a priming process. *Nucleic Acids Res* **42**, 2483-92 (2014).

502 51. Lillestol, R.K., Shah, S.A., Brugger, K., Redder, P., Phan, H., Christiansen, J. & Garrett, R.A. CRISPR families of the crenarchaeal genus Sulfolobus: bidirectional transcription and dynamic properties. *Mol Microbiol* **72**, 259-72 (2009).

503 52. Bernick, D.L., Cox, C.L., Dennis, P.P. & Lowe, T.M. Comparative genomic and transcriptional analyses of CRISPR systems across the genus Pyrobaculum. *Front Microbiol* **3**, 251 (2012).

504 53. Erdmann, S., Le Moine Bauer, S. & Garrett, R.A. Inter-viral conflicts that exploit host CRISPR immune systems of Sulfolobus. *Mol Microbiol* **91**, 900-17 (2014).

505 54. Wang, R., Li, M., Gong, L., Hu, S. & Xiang, H. DNA motifs determining the accuracy of repeat duplication during CRISPR adaptation in Haloarcula hispanica. *Nucleic Acids Res* (2016).

506 55. Nunez, J.K., Bai, L., Harrington, L.B., Hinder, T.L. & Doudna, J.A. CRISPR Immunological Memory Requires a Host Factor for Specificity. *Mol Cell* (2016).

507 56. Chylinski, K., Makarova, K.S., Charpentier, E. & Koonin, E.V. Classification and evolution of type II CRISPR-Cas systems. *Nucleic Acids Res* **42**, 6091-105 (2014).

508 57. Fonfara, I., Le Rhun, A., Chylinski, K., Makarova, K.S., Lecrivain, A.L., Bzdrenga, J., Koonin, E.V. & Charpentier, E. Phylogeny of Cas9 determines functional exchangeability of dual-RNA and Cas9 among orthologous type II CRISPR-Cas systems. *Nucleic Acids Res* **42**, 2577-90 (2014).

509 58. Boratyn, G.M., Schaffer Aa Fau - Agarwala, R., Agarwala R Fau - Altschul, S.F., Altschul Sf Fau - Lipman, D.J., Lipman Dj Fau - Madden, T.L. & Madden, T.L. Domain enhanced lookup time accelerated BLAST. *Biol Direct* (2012).

510 59. Grissa, I., Vergnaud, G. & Pourcel, C. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* **8**, 172 (2007).

511 60. Alkhnbashi, O.S., Shah, S.A., Garrett, R.A., Saunders, S.J., Costa, F. & Backofen, R. Characterizing leader sequences of CRISPR loci. *Bioinformatics* (2016).

512 61. Biswas, A., Staals, R.H., Morales, S.E., Fineran, P.C. & Brown, C.M. CRISPRDetect: A flexible algorithm to define CRISPR arrays. *BMC Genomics* (2016).

513 62. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-7 (2004).

514 63. Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Molecular Biology and Evolution* **30**, 2725-2729 (2013).
64. Okonechnikov, K., Golosova O Fau - Fursov, M. & Fursov, M. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* **8**(2012).
65. Crooks, G.E., Hon, G., Chandonia, J.M. & Brenner, S.E. WebLogo: a sequence logo generator. *Genome Res* **14**, 1188-90 (2004).
66. Lier, C., Baticle, E., Horvath, P., Haguenoer, E., Valentim, A.-S., Glaser, P., Mereghetti, L. & Lanotte, P. Analysis of the type II-A CRISPR-Cas system of Streptococcus agalactiae reveals distinctive features according to genetic lineages. *Frontiers in Genetics* **6**, 214 (2015).
67. Arslan, Z., Hermanns, V., Wurm, R., Wagner, R. & Pul, U. Detection and characterization of spacer integration intermediates in type I-E CRISPR-Cas system. *Nucleic Acids Res* **42**, 7884-93 (2014).
68. Rollie, C., Schneider, S., Brinkmann, A.S., Bolt, E.L. & White, M.F. Intrinsic sequence specificity of the Cas1 integrase directs new spacer acquisition. *Elife* **4**(2015).
69. Nunez, J.K., Lee, A.S., Engelman, A. & Doudna, J.A. Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. *Nature* **519**, 193-8 (2015).
70. Sanozky-Dawes, R., Selle, K., O'Flaherty, S., Klaenhammer, T. & Barrangou, R. Occurrence and activity of a type II CRISPR-Cas system in Lactobacillus gasseri. *Microbiology* (2015).
71. Rollie, C., Schneider, S., Brinkmann, A.S., Bolt, E.L. & White, M.F.A.-O.h.o.o. Intrinsic sequence specificity of the Cas1 integrase directs new spacer acquisition. LID - 10.7554/elife.08716 [doi]. *Elife* (2015).
72. Babu, M., Beloglazova, N., Flick, R., Graham, C., Skarina, T., Nocek, B., Gagarinova, A., Pogoutse, O., Brown, G., Binkowski, A., Phanse, S., Joachimiak, A., Koonin, E.V., Savchenko, A., Emili, A., Greenblatt, J., Edwards, A.M. & Yakunin, A.F. A dual function of the CRISPR-Cas system in bacterial antivirus immunity and DNA repair. *Mol Microbiol* **79**, 484-502 (2011).
Figure 1

Sequence alignment of the last 20 nucleotides of the 3’ end of the leader and the first repeat of selected Group 1 species. Height of the letters in the WebLogo indicates the degree of conservation at specific nucleotide locations. The leader-repeat end is conserved as 5’-ATTTGAGGTTT-3’.
Figure 2

Sequence alignment of the last 20 nucleotides at the 3’-end of the leader and the first repeat of selected Group 2(A) and Group 3(B) species. Height of the letters in the WebLogo indicates the degree of conservation at specific nucleotide locations. The leader-repeat of Group 2 loci is conserved as CTRCGAGGTTT, where R represents a purine base. For Group 3 members, this region is conserved as CGGTTT.
A

| Leader | Repeat |
|--------|--------|
| EFa_D32_1 | GAAAAAAATATCTCCTGAGTTTTTAGCTCATTTAGATGATCTGAGTATAAGTTCTAGACAAACAC |
| Lmo_10403s_3 | GAAAAAAATATCTCCTGAGTTTTTAGCTCATTTAGATGATCTGAGTATAAGTTCTAGACAAACAC |
| Lmo_J0161_4 | GAAAAAAATATCTCCTGAGTTTTTAGCTCATTTAGATGATCTGAGTATAAGTTCTAGACAAACAC |
| Lmo_R2-502 | GAAAAAAATATCTCCTGAGTTTTTAGCTCATTTAGATGATCTGAGTATAAGTTCTAGACAAACAC |
| Eph_ATCCBAA-412 | GAAAAAAATATCTCCTGAGTTTTTAGCTCATTTAGATGATCTGAGTATAAGTTCTAGACAAACAC |
| Emu_QU25_DNA | GAAAAAAATATCTCCTGAGTTTTTAGCTCATTTAGATGATCTGAGTATAAGTTCTAGACAAACAC |
| Spy_A20_1 | GAAAAAAATATCTCCTGAGTTTTTAGCTCATTTAGATGATCTGAGTATAAGTTCTAGACAAACAC |
| Smu_Lj23_2 | GAAAAAAATATCTCCTGAGTTTTTAGCTCATTTAGATGATCTGAGTATAAGTTCTAGACAAACAC |
| Sth_DGCC7110_3 | GAAAAAAATATCTCCTGAGTTTTTAGCTCATTTAGATGATCTGAGTATAAGTTCTAGACAAACAC |
| Sth_LMD-9.5 | GAAAAAAATATCTCCTGAGTTTTTAGCTCATTTAGATGATCTGAGTATAAGTTCTAGACAAACAC |
| Sag_A909_1 | GAAAAAAATATCTCCTGAGTTTTTAGCTCATTTAGATGATCTGAGTATAAGTTCTAGACAAACAC |
| Ssa_SK49 | GAAAAAAATATCTCCTGAGTTTTTAGCTCATTTAGATGATCTGAGTATAAGTTCTAGACAAACAC |

B

| Leader | Repeat |
|--------|--------|
| Lca_Lc10 | GAAAAAAATATCTCCTGAGTTTTTAGCTCATTTAGATGATCTGAGTATAAGTTCTAGACAAACAC |
| LFa_KCTC3681 | GAAAAAAATATCTCCTGAGTTTTTAGCTCATTTAGATGATCTGAGTATAAGTTCTAGACAAACAC |
| Lsa_TM1_1304 | GAAAAAAATATCTCCTGAGTTTTTAGCTCATTTAGATGATCTGAGTATAAGTTCTAGACAAACAC |
| Lco_KCTC3167 | GAAAAAAATATCTCCTGAGTTTTTAGCTCATTTAGATGATCTGAGTATAAGTTCTAGACAAACAC |
| LFUL_ATCC25644 | GAAAAAAATATCTCCTGAGTTTTTAGCTCATTTAGATGATCTGAGTATAAGTTCTAGACAAACAC |
| Lje_27-2_CHN | GAAAAAAATATCTCCTGAGTTTTTAGCTCATTTAGATGATCTGAGTATAAGTTCTAGACAAACAC |
| Lga_JV-V03 | GAAAAAAATATCTCCTGAGTTTTTAGCTCATTTAGATGATCTGAGTATAAGTTCTAGACAAACAC |
| Lfe_ATCC14931 | GAAAAAAATATCTCCTGAGTTTTTAGCTCATTTAGATGATCTGAGTATAAGTTCTAGACAAACAC |
| Wha_FBL4 | GAAAAAAATATCTCCTGAGTTTTTAGCTCATTTAGATGATCTGAGTATAAGTTCTAGACAAACAC |
| Lbu_NRRLB-30929 | GAAAAAAATATCTCCTGAGTTTTTAGCTCATTTAGATGATCTGAGTATAAGTTCTAGACAAACAC |
| Pac_D3 | GAAAAAAATATCTCCTGAGTTTTTAGCTCATTTAGATGATCTGAGTATAAGTTCTAGACAAACAC |
| Plo_NGR10510Q | GAAAAAAATATCTCCTGAGTTTTTAGCTCATTTAGATGATCTGAGTATAAGTTCTAGACAAACAC |
| Lsa_UCCI118 | GAAAAAAATATCTCCTGAGTTTTTAGCTCATTTAGATGATCTGAGTATAAGTTCTAGACAAACAC |
| FFrr_KCTC3544 | GAAAAAAATATCTCCTGAGTTTTTAGCTCATTTAGATGATCTGAGTATAAGTTCTAGACAAACAC |
| BB1_S17 | GAAAAAAATATCTCCTGAGTTTTTAGCTCATTTAGATGATCTGAGTATAAGTTCTAGACAAACAC |
| Lge_KCTC3527 | GAAAAAAATATCTCCTGAGTTTTTAGCTCATTTAGATGATCTGAGTATAAGTTCTAGACAAACAC |
| OKi_DSM_17330 | GAAAAAAATATCTCCTGAGTTTTTAGCTCATTTAGATGATCTGAGTATAAGTTCTAGACAAACAC |
Figure 3

Phylogenetic tree generated from the sequence alignment of the first repeats from selected type II-A species. Groups based on the segregation of the Cas1 tree are shown in cyan (Group 1), red (Group 2), and yellow (Group 3). The tree segregates into 6 main clades and WebLogos were produced with alignments of the last 20 nucleotides at the 3’-end of the leader and the first repeat from the loci within each corresponding branch.
Figure 4

Phylogenetic tree generated from the sequence alignment of Cas1. Groups are shown in cyan (Group 1), red (Group 2), and yellow (Group 3). WebLogos were generated by aligning the last 7 nucleotides of the leader and the first 4 nucleotides of the repeat from the loci within each corresponding branch. The tree segregates into 4 branches, two branches showing the Group 1 leader end motif, one branch showing the Group 2 motif, and one branch showing the less-conserved Group 3 leader end. Sps ED99 segregated independently from the final branch but was used in the final branch WebLogo construction based on the leader end and protein length.
Figure 5

(A) Phylogenetic tree generated from the sequence alignment of Cas9. Groups based on the segregation of the Cas1 tree are shown in cyan (Group 1), red (Group 2), and yellow (Group 3). The tree shows 5 different branches with two branches showing the Group 1 leader end motif, one branch showing the Group 2 motif, and one branch representing the less-conserved Group 3 leader end. One of the branches represent a very loosely conserved Group 1 loci. Three members of Group 3 segregated away from the normal cluster, of which Plo NGRI0510Q has a very short Cas9 sequence. Lru ATCC25644 and Lfa KCTC3681 have normal length Cas9 sequences. (B) Phylogenetic tree generated from the sequence alignment of Cas2. All the four branches segregate similarly to those of Cas1 phylogenetic tree. WebLogos for both panels of the figure were generated by aligning the last 7 nucleotides of the leader and the first 4 nucleotides of the repeat from the loci within each corresponding branch.
Figure 6

Phylogenetic tree generated from the sequence alignment of Csn2. Groups based on the segregation of the Cas1 tree are shown in cyan (Group 1), red (Group 2), and yellow (Group 3). WebLogos were generated from aligning the last 7 nucleotides of the leader and the first 4 nucleotides of the repeat from the loci within each corresponding branch. Values next to branch labels indicate the average length of the proteins (in amino acids, aa) within the branch. Two branches show the Group 1 leader end motif, one branch shows the Group 2 motif, and one branch shows the less conserved Group 3 leader end.
