Published in final edited form as:

Mol Cell. 2018 June 07; 70(5): 814–824.e6. doi:10.1016/j.molcel.2018.05.002.

Cas4 nucleases define the PAM, length, and orientation of DNA fragments integrated at CRISPR loci

Masami Shiimori[1,*], **Sandra C. Garrett**[2,*], **Brenton R. Graveley**[2], and **Michael P. Terns**[1,3,4,5]

[1]Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA 30602, USA

[2]Department of Genetics and Genome Sciences, Institute for Systems Genomics, UConn Stem Cell Institute, UConn Health, Farmington, Connecticut 06030, USA

[3]Department of Genetics, University of Georgia, Athens, GA 30602, USA

[4]Department of Microbiology, University of Georgia, Athens, GA 30602, USA

SUMMARY

To achieve adaptive and heritable immunity against viruses and other mobile genetic elements, CRISPR-Cas systems must capture and store short DNA fragments (spacers) from these foreign elements into host genomic CRISPR arrays. This process is catalyzed by conserved Cas1/Cas2 integration complexes, but the specific roles of another highly conserved protein linked to spacer acquisition, the Cas4 nuclease, are just now emerging. Here, we show that two Cas4 nucleases (Cas4-1 and Cas4-2) play critical roles in CRISPR spacer acquisition in *Pyrococcus furiosus*. The nuclease activities of both Cas4 proteins are required to process protospacers to the correct size. Cas4-1 specifies the upstream PAM (Protospacer Adjacent Motif) while Cas4-2 specifies the conserved downstream motif. Both Cas4 proteins ensure CRISPR spacer integration in a defined orientation leading to CRISPR immunity. Collectively, these findings provide *in vivo* evidence for critical roles of Cas4 nucleases in protospacer generation and functional spacer integration at CRISPR arrays.

Eblurb

Correspondence: mterns@uga.edu (M.P.T.), graveley@uchc.edu (B.R.G.).
[5]Lead Contact
*These authors contributed equally to this work

SUPPLEMENTAL INFORMATION

Supplemental information includes seven figures and three tables and can be found with this article on line at *

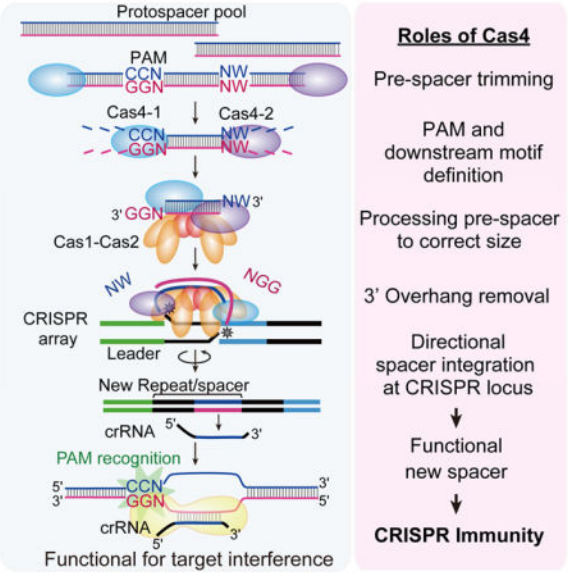
AUTHOR CONTRIBUTIONS

M.S. and M.P.T. conceived the study, M.S. and S.C.G. performed experiments, all authors designed experiments, analyzed the data, and wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing financial interests.

Shiimori et al. demonstrate integral roles of Cas4 nucleases in DNA acquisition for CRISPR-Cas immunity in *P. furiosus*. The concerted action of two different Cas4 nucleases is required for protospacer adjacent motif (PAM) definition, pre-spacer processing, and integration of DNA fragments at the CRISPR array in a functional orientation.



INTRODUCTION

Diverse CRISPR-Cas (Clustered regularly interspaced short palindromic repeats and CRISPR-associated proteins) systems provide bacteria and archaea with heritable and sequence-specific immunity against foreign elements such as viruses and plasmids (Hille et al., 2018; Makarova et al., 2015; van der Oost et al., 2014). All CRISPR-Cas systems must first recognize and excise foreign DNA (protospacer) sequences, process them to the correct size and integrate the resultant DNA fragments (spacers) into genomic CRISPR arrays in a functional orientation. The CRISPR arrays consist of an upstream leader region followed by short direct DNA repeats separated by spacers derived from previously encountered invaders. Small RNAs produced from these arrays form complexes with Cas proteins, including RNA-guided nucleases that destroy nucleic acids complementary to the crRNAs (Hille et al., 2018; Jackson et al., 2017; van der Oost et al., 2014), thus providing a means to prevent viral infection or invasion by other potentially harmful mobile genetic elements.

We are just beginning to understand mechanisms by which foreign DNA is recognized, captured, and integrated into CRISPR arrays, a process referred to as adaptation (Amitai and Sorek, 2016; Jackson et al., 2017; Sternberg et al., 2016). Typically, protospacer DNA is recognized by the presence of a 2–5 base pair (bp) flanking protospacer-adjacent motif (PAM) (Shah et al., 2013). The PAM is important both in the initial selection of protospacers during adaptation and during target recognition, which is critical for crRNA-mediated DNA destruction (Deveau et al., 2008; Shah et al., 2013). Molecular events that result in DNA breaks and free DNA termini promote protospacer generation (Levy et al., 2015; Modell et al., 2017; Shiimori et al., 2017), but the exact pathway is not clear. Protospacers must be

captured, processed into correctly-sized spacer fragments with 3′ hydroxyl ends suitable for integration (Nunez et al., 2015a; Wang et al., 2015) and inserted into CRISPR arrays in a specific orientation dictated by the PAM (Nunez et al., 2015b). Spacer integration could potentially occur in two orientations and during *in vitro* integration with purified Cas1 and Cas2 proteins, both orientations are observed with similar frequency (Lopez-Sanchez et al., 2012; Shmakov et al., 2014). However, only the "correct" orientation would be functional for target recognition (Figure S1) and this orientation is normally highly favored *in vivo*. Recent studies have illuminated molecular details of spacer insertion by the core adaptation proteins Cas1 and Cas2 (Fagerlund et al., 2017; Nunez et al., 2015b; Wang et al., 2015; Wright et al., 2017; Xiao et al., 2017). However, these proteins alone cannot guide directionality of spacer integration *in vitro*.

CRISPR-Cas systems are classified into two classes, six types (I–VI), and at least 24 subtypes, each with different Cas proteins, crRNA species and mechanisms to execute the CRISPR-Cas immune pathway (Hille et al., 2018; Makarova et al., 2015; van der Oost et al., 2014). Cas1 and Cas2 are conserved in all systems able to undergo adaptation. Cas4 is a widely conserved protein in many archaeal and bacterial Type I, II and V systems (Hudaiberdiev et al., 2017) and was implicated in adaptation by several lines of evidence: genes encoding Cas4 and Cas1 are typically adjacent, Cas4 proteins physically associate with Cas1/Cas2 complexes (Plagens et al., 2012) and are sometimes found as a Cas4/Cas1 fusion protein (Viswanathan et al., 2007), and Cas4 deletions prevent adaptation in Type I-A and I-B systems (Li et al., 2014; Wright et al., 2017). Cas4 proteins contain a RecB nuclease domain and have demonstrated 5′-3′ single-stranded DNA exonuclease activity *in vitro* with certain Cas4 orthologs also exhibiting DNA unwinding and endonuclease activities (Lemak et al., 2013; Lemak et al., 2014; Zhang et al., 2012). Recent *in vitro* (Lee et al., 2018; Rollie et al., 2018) and heterologous *in vivo* (Kieper et al., 2018) studies have further extended our knowledge of Cas4 function. Cas4 was found to form a stable complex with Cas1 (Lee et al., 2018). Furthermore, evidence was gained that Cas4 cleaves pre-spacer DNA overhangs in a PAM-dependent manner (Kieper et al., 2018; Lee et al., 2018; Rollie et al., 2018). Here we show that key functional roles for Cas4 proteins in spacer acquisition is also borne out *in vivo* in a native host, the hyperthermophilic archaeon, *Pyrococcus furiosus.* We find that protospacer fragment selection and proper pre-spacer processing requires two Cas4 proteins (encoded by distinct genes). The two Cas4 proteins collaborate to coordinate PAM recognition, pre-spacer trimming and 3′ overhang removal, and directional integration of the processed spacer DNAs into the CRISPR array in an orientation that results in functional crRNA/target DNA interaction and CRISPR immunity.

RESULTS

The organism used in this study, *P. furiosus,* has three CRISPR-Cas immune effector crRNA-Cas protein complexes, including two type I systems, a type I-A (Csa) and a type I-G (Cst), and a type III-B (Cmr) system (Terns and Terns, 2013). The genome also harbors seven CRISPR loci, with similar adaptation characteristics (Shiimori et al., 2017), and encodes four adaptation proteins: Cas1, Cas2, Cas4-1, and Cas4-2 (Figure 1A) (Terns and Terns, 2013). The *cas4-1* gene is located immediately adjacent to *cas1* and *cas2*, while *cas4-2* resides in a remote genomic location (Figure 1B). We recently reported that *P.*

furiosus actively takes up spacers into each CRISPR locus in a PAM-dependent manner. Adaptation occurs under normal growth conditions and the frequency of adaptation can be increased by over-expression of Cas1, Cas2, and Cas4 proteins (Shiimori et al., 2017).

Cas4 deletion strains have abnormal PAMs and spacer sizes

To investigate the individual roles of the four Cas proteins in adaptation, we generated strains lacking one or more of the genes (*i.e.*, Δ*cas1*, Δ*cas2*, Δ*cas4-1*, Δ*cas4-2*, or Δ*cas4-1/*Δ*cas4-2*). The deletions were done in either the wild type strain (WT), or in a strain with promoter replacement to overexpress Cas1, Cas2, Cas4-1, and Cas4-2 (strain Cas-OE, Figure 1B). We then characterized newly integrated spacers by PCR and high-throughput sequencing (Shiimori et al., 2017) (Figure 1C, D). As expected, Cas1 and Cas2 were necessary and sufficient for adaptation (Figure 2A, lanes 2, 4, and 10, lower panel; Figure S2B for results with the Cas-OE strain). In contrast, neither Cas4-1 nor Cas4-2 were essential but deleting either Cas4 protein reduced adaptation efficiency in the WT background (Figure 2A, compare lanes 6, 8, 10 with lane 1).

Next, we characterized newly acquired spacers to identify changes due to Cas4 deletions. As we found previously (Shiimori et al., 2017), a majority of new spacers arose from the *P. furiosus* chromosome rather than the transformed plasmid and this was true of all *cas* deletion strains (90–100%, data not shown). The 200 existing *P. furiosus* spacers are predominantly 37 bp long (Shiimori et al., 2017). Consistently, new spacers in WT were also about 37 bp, with over 90% of them being 36–38 bp in length (Figure 2B, Figure S2C). Interestingly, long spacers (up to 70 bp) were acquired in Δ*cas4-1/*Δ*cas4-2*, implying that Cas4-1 and Cas4-2 work together to process protospacers to the proper size for integration (~37 bp). Alone, both Cas4-1 and Cas4-2 processed protospacers, but generated spacers averaging 36 and 38 bp, respectively.

Given the evidence that Cas4-1 and Cas4-2 are involved in processing protospacers, we tested if they recognize and trim the PAM sequence. As reported earlier (Shiimori et al., 2017), a strong 5′-CCN-3′/5′-NGG-3′ consensus PAM was observed upstream of protospacers, while a weak 5′-NW-3′/5′-WN-3′ (W is A or T) consensus sequence was observed downstream in WT (Figure 2C; Figure S2D for results with Cas-OE strain). In contrast, in Δ*cas4-1*, the CCN PAM was not observed, and a WN motif was present both upstream and downstream. In Δ*cas4-2*, a weak 5′-CCN-3′ motif was observed downstream as well as upstream (Figure 2C; Figure S2D). No consensus sequence was observed on either side in Δ*cas4-1/*Δ*cas4-2* (Figure 2; Figure S2) indicating that Cas1 and Cas2 alone are insufficient for PAM recognition during protospacer-to-spacer conversion. When Cas4-1 and/or Cas4-2 were expressed from a plasmid in the deletion strains, both the PAM and the size distribution were rescued (Figure S3).

We reasoned that the weak 5′-CCN-3′/5′-NGG-3′ consensus observed on both sides of protospacers in Δ*cas4-2* (Figure 2C; Figure S2D) could arise from faulty orientation during integration or from Cas4-1 processing both sides of a protospacer in the absence of Cas4-2, thereby creating both an upstream and a downstream PAM. We counted protospacers that had a PAM on the correct side, the wrong side, both sides, or neither. For comparison, we determined the percentage of all possible 37 bp protospacers in the *P. furiosus* genome that

would have each upstream and downstream scenario; if new spacers were randomly pulled from this set, PAMs and NW motifs would appear at the frequency indicated by “Random”. In WT, ~90% of spacers had an upstream CCN motif only, confirming that integration is primarily in the correct (functional) orientation (Figure 2D, Table S1). In Δ*cas4-1* or Δ*cas4-1*/Δ*cas4-2*, most protospacers had neither a CCN nor a NGG motif; the motifs appeared at the frequency expected for a randomly selected protospacer. In Δ*cas4-2*, ~65% of spacers had an upstream CCN, ~19% of spacers had a downstream NGG only, while ~10% had both (Figure 2D). This last result, together with the bimodal size distribution in Δ*cas4-2* (Figure 2B), indicated that a subset of the protospacers is processed by Cas4-1 on both sides to produce new spacers that were 36 bp long with a PAM both upstream and downstream. Logos for only the 36 bp spacers show the dual upstream/downstream PAM more clearly (Figure S4).

We also analyzed whether protospacers had a 5′-NW-3′/5′-WN-3′ flanking sequence on either, neither, or both sides (Figure 2E). In WT, ~60% of protospacers had a downstream NW only; this was the case for less than 25% of protospacers in the three Cas4 deletion strains. Without Cas4-1, ~50% of protospacers exhibited a 5′-NW-3′/5′-WN-3′ motif on both sides; without Cas4-2, over 50% of spacers lacked this motif on both sides (Figure 2E), providing additional evidence that Cas4-2 determines the NW motif.

Cas4-1 and Cas4-2 trim and orient a duplexed DNA oligonucleotide pre-spacer

To more directly test requirements for PAM recognition, we established a system to examine integration of a defined protospacer sequence. We introduced a 57 bp duplexed DNA oligonucleotide that included a 37 bp spacer sequence bounded by 10 bp of flanking sequence on each side, with a canonical upstream CCN PAM and downstream NW motif (Figure 3A) into *P. furiosus* cells and detected CRISPR integration by PCR using an oligo-specific primer (Figure 3B) and unbiased amplicon sequencing of expanded arrays (Figure 3C). Experiments were done in the Cas-OE strains because the higher adaptation efficiency allowed us to capture the relatively rare oligonucleotide integration events.

After introduction of the 57 bp oligonucleotides into *P. furiosus*, most new spacers were 37 bp, were processed correctly next to the PAM and NW motifs, and were integrated in the correct orientation (Figure 3B, C). In contrast, the duplexed DNA oligonucleotide was integrated in both orientations in all three Cas4 deletion strains (Figure 3B, C). In Δ*cas4-1*, there was a dramatic reduction in processing at the CCN PAM. For Δ*cas4-2*, most oligonucleotides were processed at the CCN PAM, but were integrated in both orientations. The Δ*cas4-1*/Δ*cas4-2* strain integrated nearly full-length oligonucleotides with no consistent trimming patterns (Figure 3C). Surprisingly, oligonucleotide integration was much more frequent in the Δ*cas4*-2 strain (Figure S5; see band intensities, Figure 3B, 4B, 5B) even though overall adaptation efficiency appears reduced in this strain (Figure 2A, Figure S2B).

The oligonucleotide transformation assay also allowed us to ask how strandedness of the protospacer DNA affects PAM recognition (Figure 4). We compared integration results using single-stranded oligonucleotide vs. double-stranded (duplexed) oligonucleotides with either a canonical (5′-CCT-3′/5′-AGG-3′) or a mutated (5′-CGC-3′/5′-GCG-3′) PAM (Figure 4A). Single-stranded oligonucleotides were not integrated in any strain (Figure 4B, lane 2–3,

9–10, 16–17), indicating that protospacer DNA must be at least partially double-stranded. As described above, the oligonucleotide with a canonical PAM was trimmed and integrated in the correct orientation (Figure 4B, lane 4; Figure 4C). An oligonucleotide with a canonical PAM on the bottom strand only was also trimmed and integrated correctly (Figure 4B, lanes 4 and 6; Figure 4C). In contrast, oligonucleotides with only a top strand PAM or with a double-stranded mutant PAM were neither trimmed nor integrated correctly (Figure 4B, lanes 5 and 7; Figure 4C). In Δ*cas4-1*, all oligonucleotides were integrated with similar efficiencies in both orientations (Figure 4B, lanes 11–14). In Δ*cas4-2*, a canonical PAM on the bottom strand only or a double-stranded PAM both led to efficient integration in both orientations (Figure 4B, lanes 18 and 20). These results indicate that a 5′-NGG-3′ PAM on the *bottom* strand of the protospacer is required for recognition and processing and that Cas4-1 carries out this recognition.

We next tested strandedness for NW motif processing: DNA oligonucleotides had either a canonical (5′-AT-3′/5′-AT-3′), a fully mutated (5′-GC-3′/5′-CG-3′), or a one-strand-only NW motif (Figure 4D). All four oligonucleotides were integrated with the correct orientation in *Cas-OE*, but in both orientations in Δ*cas4-1* or Δ*cas4-2* (Figure 4D, compare lanes 2–5 with 7–10 and 12–15, respectively). Oligonucleotides with a double-stranded canonical motif or a canonical motif on the top strand only were primarily processed at the predicted NW site (Figure 4E), while the double-stranded mutant motif or the bottom strand only NW caused processing to shift to nearby A/T sites (1–2 nucleotides away). These results indicate that Cas4-2 recognizes the NW motif on the *top* strand, but that distance from the upstream CCN PAM is the primary determinant of spacer trimming.

Taken together, our findings from the duplexed DNA oligonucleotide experiments suggested that Cas4-1 and Cas4-2 may process the opposite ends of a pre-spacer, with Cas4-1 acting at the PAM-bearing end and Cas4-2 acting on the opposite end. Cas4-1 and Cas4-2 may cleave CCN PAM and NW motifs, respectively, to form spacers with 3′ hydroxyl termini and 3′ single-stranded DNA overhangs, properties that enhance *in vitro* integration by Cas1/Cas2 (Nunez et al., 2014; Wang et al., 2015; Xiao et al., 2017). The observation that recombinant *E. coli* Cas1/Cas2 proteins were capable of removing 3′ single-stranded DNA overhangs from pre-spacers (Wang et al., 2015) prompted us to hypothesize that in *P. furiosus*, Cas4-1 and Cas4-2 may create these overhangs and Cas1/Cas2 may then carry out the final steps of 3′ DNA overhang removal and integration. To test this, we conducted a variation of the duplexed DNA oligonucleotide capture assay, this time using a preprocessed oligonucleotide with 5′-NGG-3′ PAM and 5′-NW-3′ single stranded 3′ overhangs (Figure 5A). We qualitatively assessed integration efficiency and orientation by PCR and found that while Cas1 and Cas2 alone were capable of integrating both the 57 bp duplexed pre-spacers and 3′ overhang-containing 37 bp pre-spacers *in vivo*, they did so without maintaining the correct orientation (Figure 5B). Next, we prepared selected PCR products from the CRISPR5 array for high-throughput sequencing to determine the position(s) where trimming occurred. Cas1 and Cas2 alone failed to consistently trim overhangs and often integrated fully untrimmed oligonucleotides. In contrast, when Cas4-1 and Cas4-2 were present, preprocessed spacers were correctly trimmed (Figure 5C). These results show that both Cas4 proteins are required for spacer processing even after 3′ overhangs have been created and suggest a direct role for

Cas4 nucleases in 3′ DNA overhang removal prior to spacer integration into CRISPR arrays.

Cas4-1 and Cas4-2 nuclease activity is essential for PAM recognition and spacer sizing

The spacer size distributions and oligonucleotide trimming patterns suggested that Cas4-1 and Cas4-2 are directly involved in cleaving pre-spacer DNA. Cas4 proteins share structural (Figure S6) and enzymatic (Lemak et al., 2013; Lemak et al., 2014; Zhang et al., 2012) similarities with RecB and AddB DNA recombination and repair nucleases suggesting that Cas4-1 and Cas4-2 may similarly cleave DNA. However, we could not rule out the possibility that the role of the two Cas4 proteins in DNA trimming and 3′ overhang processing was indirect, with Cas1/Cas2 or other cellular nucleases carrying out the cleavages. To address this, we studied the phenotype of strains expressing nuclease defective Cas4-1 and Cas4-2. In these strains, forms of Cas4-1 or Cas4-2 with nuclease active site mutations were expressed from a plasmid in the respective Cas4 deletion background. We designed amino acid changes based on sequence conservation (Figure S6) and published findings (Lemak et al., 2013; Lemak et al., 2014; Zhang et al., 2012). In addition to testing the role of Cas4 nuclease activity, we also created plasmids with mutations to disrupt helicase activity and the conserved Fe-S cluster (Lemak et al., 2013; Lemak et al., 2014; Zhang et al., 2012) (Figure S6).

Plasmid-expressed WT Cas4-1 rescued spacer size distribution (Figure 6B) and the upstream PAM (Figure 6D and E). However, the nuclease defective Cas4-1 had skewed spacer sizes and no upstream PAM, indicating that nuclease activity of Cas4-1 is essential for its role in spacer sizing and PAM definition (Figure 6D and E). Likewise, the plasmid-expressed WT Cas4-2 rescued spacer sizes and the downstream NW motif, but the nuclease defective Cas4-2 did not (Figure 6D and E). The behavior of Fe-S cluster mutants is identical to that of Cas4 null strains, consistent with published findings that the Fe-S cluster of at least certain proteins is required for structural integrity of Cas4 proteins (Lemak et al., 2013) (Figure 6). However, there was no discernable phenotype for the predicted helicase mutation and we note that mutation of this conserved histidine residue in *Sulfolobus solfataricus* Cas4 does not completely inactivate helicase activity *in vitro* (Lemak et al., 2013). Our results indicate that nuclease activity is essential for the ability of both Cas4-1 and Cas4-2 to define spacer size, PAM, and orientation of spacer integration at CRISPR arrays.

We also obtained indirect evidence that Cas4-2 is an active nuclease. While *cas4-2* deletion did not increase adaptation frequency overall (Figure 2), it did dramatically increase incorporation of the duplexed DNA oligonucleotide (Figure S5). We assume that potential pre-spacers derived from the dynamic pool of DNA products in the cell are undergoing constant depletion and renewal. On the other hand, the transformed oligonucleotides were introduced at a single time point at the beginning of each experiment and we assume that they become rapidly degraded once inside the cell. Since the oligonucleotides were about 1,000 times more likely to become spacers in the *cas4-2* deletion strains vs. strains harboring the Cas4-2 protein (as long as they had a bottom-strand PAM; Figure S5) we infer that Cas4-2 normally degrades free DNA.

DISCUSSION

The universally conserved Cas1 and Cas2 proteins catalyze spacer integration into CRISPR loci and most studies to date have focused on organisms where they are the only Cas proteins required for adaptation (Amitai and Sorek, 2016; Jackson et al., 2017; Sternberg et al., 2016). In *P. furiosus*, we show that Cas1 and Cas2 alone lead to indiscriminate spacer acquisition, and that both Cas4-1 and Cas4-2 are necessary for accurately guiding adaptation. Cas4-1 and Cas4-2 constrained spacer sizes and were necessary for PAM recognition and removal during protospacer generation and for ensuring that new spacers were integrated in the correct orientation (Figures 2, 3, 4, 5, 7). Both checkpoints are critical if new spacers are to generate crRNAs capable of defending against invader DNA. Efficient DNA degradation by most DNA targeting CRISPR-Cas systems requires a properly oriented PAM adjacent to the crRNA binding site. This requirement would not be met if the protospacer was selected from DNA without a PAM or if a protospacer had a PAM but was integrated in the wrong orientation (Figure S1). Our data indicate that in the absence of Cas4 proteins, the majority of newly acquired spacers would be non-functional for invader targeting and would fail to provide immunity.

Our data also indicate that Cas4 proteins specifically trim the PAM and NW-containing single-stranded overhangs just prior to integration and remain associated with the spacer through integration in order to ensure proper orientation (Figure 7). Cas4-mediated processing of the blunt-ended, 57 bp pre-spacer to a 37 bp pre-spacer with 3′ overhangs *in vivo* is not an obligatory step for spacer integration since “preprocessed” duplexed pre-spacer substrates containing 3′ overhangs are also precisely trimmed and integrated in the proper orientation in a Cas4-dependent manner (Figure 5). While Cas1 and Cas2 alone are capable of integrating both the 57 bp duplexed pre-spacers and 3′ overhang-containing 37 bp pre-spacers *in vivo*, they do so without specifically trimming the overhangs or maintaining the correct orientation of spacer integration into CRISPR arrays. These findings provide further evidence that Cas4-1 and Cas4-2 proteins are critical for PAM identification and subsequent removal. In agreement with recent findings (Kieper et al., 2018; Lee et al., 2018; Rollie et al., 2018), Cas4 nucleases function to ensure that non-functional spacers are not readily integrated into the CRISPR array.

It has long been known that the PAM dictates orientation of spacer integration at CRISPR arrays (Shah et al., 2013) but mechanistic insight into the relationship was lacking. Our data, and recent findings by others (Kieper et al., 2018; Lee et al., 2018; Rollie et al., 2018), have now implicated Cas4 in PAM recognition as well as dictating spacer orientation (Figures 2–5). Based on the sum of our findings, we outline a speculative model for how Cas4-1 and Cas4-2 process spacers and ensure their integration in a functional orientation with respect to the PAM (Figure S7). During generation of pre-spacer substrates, each Cas4 protein trims DNA (via exo- and/or endo-nucleolytic cleavage) and together the two Cas4 proteins process inward from opposite ends of a DNA fragment. Cas4-1 stalls when confronted with a PAM, while Cas4-2 stalls at an NW motif that is the appropriate distance from a stalled Cas4-1. The DNA fragment is thus sized, has suitable 3′ overhangs, and is ready for Cas1/Cas2-mediated integration into the CRISPR array. The pre-spacer remains in association with Cas4-1 and Cas4-2 as these proteins form a complex with Cas1/Cas2. A recent structure

from *Bacillus halodurans* showed two Cas4 molecules in association with a Cas1 tetramer (Lee et al., 2018), supporting the idea that a pair of Cas4 proteins would work together with Cas1 during processing and integration. In our model, the spacer acquisition complex then engages with the CRISPR array in a polarized manner. With Cas4-2 oriented towards the leader-repeat junction and Cas4-1 oriented towards the repeat-spacer junction, the complex catalyzes integration. Our evidence indicates that each Cas4 protein removes their respective 3′ DNA overhang (likely by endonucleolytic cleavage) and we envision that this final pre-spacer trimming step occurs immediately prior to Cas1-mediated nucleophilic attack at the repeat junctions. In this model, the distinct processing and binding activities of the two Cas4 proteins, together with intrinsic DNA elements in the array, provide the asymmetry that allows for PAM-to-leader orientation of spacer integration. The presence of two Cas4 proteins is not unique to *Pyrococcus furiosus* (Hudaiberdiev et al., 2017) and in these many other organisms this model may also apply. In systems that rely on the function of either two distinct Cas4 proteins or a single Cas4 protein, the precise mechanisms of orientation-specific integration remain to be discovered.

We predict a general role for Cas4 nucleases in shaping immunity in a range of bacterial and archaeal CRISPR-Cas systems. In systems lacking Cas4 proteins altogether, it is already clear that other factors are necessary for guiding Cas1 and Cas2. These include Cas9 (which provides PAM recognition) and Csn2 (unknown role) of Type II systems (Heler et al., 2015; Wei et al., 2015). A role for the Cas3 effector helicase-nuclease of Type I systems in providing pre-spacer DNA has also been obtained (Kunne et al., 2016; Musharova et al., 2017). A critical function for non-Cas host factors in mediating CRISPR adaptation has also been revealed including the RecBCD DNA repair and recombination machinery which provides pre-spacer DNA (Levy et al., 2015) and integration host factor (IHF) in Type I-E systems which localizes Cas1/Cas2 to the leader end of the array (Wright et al., 2017). In systems that do have Cas4 proteins, our *in vivo* findings that Cas4 nucleases are required for PAM recognition and pre-spacer trimming are in agreement with recent *in vitro* observations, (Cas4 from *Sulfolobus solfataricus* (Rollie et al., 2018) or *Bacillus halidurans* (Lee et al., 2018) and in a heterologous (*E. coli*) *in vivo* model system (Cas4 from a cyanobacterium, *Synechocystis* sp. 6803 (Kieper et al., 2018)). Our work revealed an additional role for Cas4 proteins in directing the orientation of spacers into the CRISPR array in a functional orientatio. Collectively, these findings underscore the diversity of mechanisms that have evolved to direct Cas1/Cas2 integrase activity and guide the supply of functional DNA substrates entering CRISPR immune memory banks.

STAR METHODS

Detailed methods are provided in the online version of this paper and include the following:

KEY RESOURCES TABLE

The table highlights the genetically modified organisms and strains, cell lines, reagents, software, and source data **essential** to reproduce results presented in the manuscript. Depending on the nature of the study, this may include standard laboratory materials (i.e., food chow for metabolism studies), but the Table is **not** meant to be comprehensive list of all

materials and resources used (e.g., essential chemicals such as SDS, sucrose, or standard culture media don't need to be listed in the Table). **Items in the Table must also be reported in the Method Details section within the context of their use.** The number of **primers and RNA sequences** that may be listed in the Table is restricted to no more than ten each. If there are more than ten primers or RNA sequences to report, please provide this information as a supplementary document and reference this file (e.g., See Table S1 for XX) in the Key Resources Table.

Please note that ALL references cited in the Key Resources Table ***must be included in the References list.*** Please report the information as follows:

- **REAGENT or RESOURCE:** Provide full descriptive name of the item so that it can be identified and linked with its description in the manuscript (e.g., provide version number for software, host source for antibody, strain name). In the Experimental Models section, please include all models used in the paper and describe each line/strain as: model organism: name used for strain/line in paper: genotype. (i.e., Mouse: $OXTR^{fl/fl}$: B6.129(SJL)-$Oxtr^{tm1.1Wsy/J}$). In the Biological Samples section, please list all samples obtained from commercial sources or biological repositories. Please note that software mentioned in the Methods Details or Data and Software Availability section needs to be also included in the table. See the sample Table at the end of this document for examples of how to report reagents.
- **SOURCE:** Report the company, manufacturer, or individual that provided the item or where the item can obtained (e.g., stock center or repository). For materials distributed by Addgene, please cite the article describing the plasmid and include "Addgene" as part of the identifier. If an item is from another lab, please include the name of the principal investigator and a citation if it has been previously published. If the material is being reported for the first time in the current paper, please indicate as "this paper." For software, please provide the company name if it is commercially available or cite the paper in which it has been initially described.
- **IDENTIFIER:** Include catalog numbers (entered in the column as "Cat#" followed by the number, e.g., Cat#3879S). Where available, please include unique entities such as RRIDs, Model Organism Database numbers, accession numbers, and PDB or CAS IDs. For antibodies, if applicable and available, please also include the lot number or clone identity. For software or data resources, please include the URL where the resource can be downloaded. Please ensure accuracy of the identifiers, as they are essential for generation of hyperlinks to external sources when available. Please see the Elsevier list of Data Repositories with automated bidirectional linking for details. When listing more than one identifier for the same item, use semicolons to separate them (e.g. Cat#3879S; RRID: AB_2255011). If an identifier is not available, please enter "N/A" in the column.

– ***A NOTE ABOUT RRIDs:*** We highly recommend using RRIDs as the identifier (in particular for antibodies and organisms, but also for software tools and databases). For more details on how to obtain or generate an RRID for existing or newly generated resources, please visit the RII or search for RRIDs.

Please use the empty table that follows to organize the information in the sections defined by the subheading, skipping sections not relevant to your study. Please do not add subheadings. To add a row, place the cursor at the end of the row above where you would like to add the row, just outside the right border of the table. Then press the ENTER key to add the row. Please delete empty rows. Each entry must be on a separate row; do not list multiple items in a single table cell. Please see the sample table at the end of this document for examples of how reagents should be cited.

TABLE FOR AUTHOR TO COMPLETE

Please upload the completed table as a separate document. **Please do not add subheadings to the** Key Resources Table. If you wish to make an entry that does not fall into one of the subheadings below, please contact your handling editor. (**NOTE:** For authors publishing in Current Biology, please note that references within the KRT should be in numbered style, rather than Harvard.)

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Chicken Anti-Cas1	(Shiimori et al., 2017)	N/A
Chicken Anti-Cas2	(Shiimori et al., 2017)	N/A
Chicken Anti-Cas4-1	(Shiimori et al., 2017)	N/A
Chicken Anti-Cas4-2	This paper	N/A
Chicken Anti-Csa2	(Majumdar et al., 2015)	N/A
HRP Donkey Anti-Chicken IgY	Gallus Immunotech	DAIgY-HRP
Bacterial and Virus Strains		
Escherichia coli TOP10	Thermo Fisher Scientific	C4040-03
Pyrococcus furiosus JFW02 (WT)	(Farkas et al., 2012)	N/A
Pyrococcus furiosus TPF81 (ΔCas1 in WT)	This paper	N/A
Pyrococcus furiosus TPF82 (ΔCas2 in WT)	This paper	N/A
Pyrococcus furiosus TPF105 (ΔCas4-1 in WT)	This paper	N/A
Pyrococcus furiosus TPF57(ΔCas4-2 in WT)	This paper	N/A
Pyrococcus furiosus TPF106 (ΔCas4-1/ΔCas4-2 in WT)	This paper	N/A
Pyrococcus furiosus TPF77 (Cas-OE)	This paper	N/A
Pyrococcus furiosus TPF89 (ΔCas1 in Cas-OE)	This paper	N/A
Pyrococcus furiosus TPF90 (ΔCas2 in Cas-OE)	This paper	N/A
Pyrococcus furiosus TPF79 (ΔCas4-1 in Cas-OE)	This paper	N/A

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Pyrococcus furiosus TPF91 (ΔCas4-2 in Cas-OE)	This paper	N/A
Pyrococcus furiosus TPF67 (ΔCas4-1/ΔCas4-2 in Cas-OE)	This paper	N/A
Pyrococcus furiosus TPF65 (Δ)	This paper	N/A
Biological Samples		
Chemicals, Peptides, and Recombinant Proteins		
OneTaq 2xMM	New England Biolab	#M0486S
Phusion polymerase	New England Biolab	#M0530S
T4 DNA ligase	New England Biolab	#M0202S
DpnI	New England Biolab	#R0176S
NruI	New England Biolab	#R0192S
NotI	New England Biolab	#R0189S
EcoRV	New England Biolab	#R0195S
BamHI	New England Biolab	#R0136S
NdeI	New England Biolab	#R0111S
Purple loading dye	New England Biolab	#B7024S
1 kb ladder	New England Biolab	#N3232S
dNTP	New England Biolab	#N0447S
TSAP Thermosensitive Alkaline Phosphatase	Promega	M9910
100 bp ladder	Gold Biotechnology	D001-3000
Ethidium Bromide	SIGMA	E1510
Agarose	Denvulle Scientific Inc	GR140-500
Tris-Base	Fisher Scientific	164824
EDTA	VMR life Science	0105-500g
Acetic acid	Fisher Scientific	02-002-118
LB	RPI	31207
Bacto agar	BD	214010
Ampicillin	Fisher Scientific	05-549-8
Apramycin	RPI	A50020-10.0
Kanamysin	RPI	K22000-10.0
NaCl	SIGMA	S5886-10KG
$MgSO_4 \cdot 7H_2O$	Honeywell	10034-99-8
$MgCl_2 \cdot 6H_2O$	SIGMA	M2670-500g
KCl	SIGMA	P5405-500G
NH_4Cl	SIGMA	A9435-500G
$CaCl_2 \cdot 2H_2O$	SIGMA	C7902-500G
HCl	Fisher Scientific	02-003-048
$FeCl_3$	SIGMA	157740-100G

REAGENT or RESOURCE	SOURCE	IDENTIFIER
H_3BO_3	RPI	B32050-5000.1
$ZnCl_2$	SIGMA	229997-10G
$CuCl_2{\cdot}2H_2O$	SIGMA	C3279-100G
$MnCl_2{\cdot}4H_2O$	SIGMA	203734-5G
$(NH_4)_2MoO_4$	SIGMA	277908-5G
$AlK(S0_4)\cdot 2H_20$	ACROS	7784-24-9
$CoCl_2{\cdot}6H_2O$	SIGMA	654507-5G
$NiCl_2{\cdot}6H_2O$	SIGMA	339350-50G
$Na_2WO_4{\cdot}2H_20$	SIGMA	223336-100G
Niacin	ACROS	128291000
Biotin	ACROS	230090010
Pantothenate	ACROS	416750250
Lipoic Acid	ACROS	136720050
Folic Acid	Fisger Scientific	BP251910
p-Aminobenzoic Acid	MP	194619
Thiamine B_1	ACROS	148990100
Riboflavin B_2	ACROS	132350250
Pyridoxine B_6	ACROS	150750500
Cobalamin B_{12}	ACROS	405920010
Alanine	AVROS	102831000
Arginine	SIGMA	11039-100G
Asparagines	ACROS	175271000
Aspartic acid	SIGMA	A9256-100G
Glutamic acid	SIGMA	G1251-500G
Glutamine	ACROS	119951000
Glycine	SIGMA	G7126-500G
Histidine	SIGMA	H8000-100G
Isoleucine	SIGMA	W527602-100G
Leucine	SIGMA	L8000-100G
Lysine	SIGMA	L5501-100G
Methionine	SIGMA	M9625-100G
Phenylalanine	SIGMA	P2126-100G
Proline	Fisher Scientific	147-85-3
Serine	SIGMA	S4500-100G
Threonine	SIGMA	T8625-100G
Tryptophan	SIGMA	T8941-100G
Tyrosine	ICN Biomedicals Inc	103183
Valine	SIGMA	94619-100G
KH_2PO_4	SIGMA	P5655-500G

REAGENT or RESOURCE	SOURCE	IDENTIFIER
K_2HPO_4	SIGMA	P3786-500G
D-(+)-Cellobiose	ACROS	108465000
Resazurin	SIGMA	199303-5G
cysteine hydrochloride	SIGMA	C121800-100G
Sodium sulfide (Na2S)	ACROS	387065000
sodium bicarbonate (NaHCO3)	SIGMA	S8875-500G
GELRITE	RPI	71010-52-1
Uracil	SIGMA	U1128-25G
5-FOA	Apollo Scientific Limited	AS422771
SDS	BIO-RAD	#161-0302
glycine	RPI	G36050-5000.0
methanol	Fisher Scientific	A412-20
Blotting-Glade Blocker	BIO-RAD	#170-6404
Tween 20	BIO-RAD	#170-6531
cOmplete™, Mini, EDTA-free Protease Inhibitor Cocktail	Roche	11836170001
Acrylamide/Bis solution 29:1	VMR Life Science	0311-1L
Ammonium persulfate	Fisher Scientific	BP179-100
TEMED	RPI	T18000-0.1
PageRuler Plus Prestained Protein ladder	Thermo Scientific	#26619
bromophenol blue	BIO-RAD	161-0404
glycerol	Fisher Scientific	G33-4
DTT	SIGMA	10197777001
Sodium acetate	SIGMA	S2889-250G
ethanol	Fisher Scientific	BP2818100
Critical Commercial Assays		
GENEART seamless cloning kit	Invitrogen	#A14606
Zymoclean DNA Gel Recovery Kit	Zymo Research	D4007
DNA Clean & Concentration Kit	Zymo Research	D4029
Quick DNA Miniprep Kit	Zymo Research	D3025
ZR Plasmid Miniprep Classic	Zymo Research	D4054
Zyppy Plasmid Maxiprep Kit	Zymo Research	D4028
ECL Prime	GE Healthcare	RPN2232
MiSeq Reagent kit v2 (300 cycle)	Illumina	MS-102-2002
Deposited Data		
Gel and Western blot images	This paper	https://dx.doi.org/[10.17632/7ffsrg5bzk.]
Experimental Models: Cell Lines		

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Experimental Models: Organisms/Strains		
Oligonucleotides		
See Table S3 for sequences of oligonucleotides used in this study.		
Recombinant DNA		
pMS1 (target plasmid)	(Swarts et al., 2015)	N/A
pHSG298	Takara	
pJFW17 (AprR general cloning vector with *E.coli* OriT, and *P. furiosus* Pgdh-pyrF cassette)	(Farkas et al., 2011)	N/A
pJFW18 (pJFW17 derivative, Pfu OriC for replication in *P. furiosus*)	(Farkas et al., 2011)	N/A
pJE47 (pJFW18 derivative, Tkcsg promoter/TkchiA terminator expression cassette)	(Elmore et al., 2015)	N/A
pJE60 (pJE47 derivative; the PyrF gene replaced with TrpAB)	This Study	N/A
pJE64 (pJFW18 derivative, the PyrF gene replaced with TrpAB)	This Study	N/A
pMS030 (pHSG298 derivative with a genome insertion designed to insert sequences by replacing PF1120 through CR7 with a PyrF selection marker)	This Study	N/A
pMS032 (pMS030 with Cas1, Cas2, Cas4-1 overexpression cassette)	This Study	N/A
pMS088 (pHSG298 derivative with a genome insertion designed to insert sequences by replacing PF1792 through PF1794 with a PyrF selection marker)	This Study	N/A
pMS089 (pMS088 with Cas4-2 O/E cassette)	This Study	N/A
pMS114 (pMS030 with Cas2, Cas4-1 overexpression cassette)	This Study	N/A
pMS115 (pMS030 with Cas1, Cas4-1 overexpression cassette)	This Study	N/A
pMS087 (pMS030 with Cas1, Cas2 overexpression cassette)	This Study	N/A
pMS142 (pJE60 with Cas1)	This Study	N/A
pMS143 (pJE60 with Cas2)	This Study	N/A
pMS071 (pJE64 with Pslp-Cas4-1)	This Study	N/A
pMS108 (pJE64 with Cas4-1D68A)	This Study	N/A
pMS109 (pJE64 with Cas4-1H91A)	This Study	N/A
pMS110 (pJE64 with Cas4-1C161A)	This Study	N/A
pMS077 (pJE60 with Cas4-2)	This Study	N/A
pMS080 (pJE60 with Cas4-2D77A)	This Study	N/A
pMS081 (pJE60 with Cas4-2H100A)	This Study	N/A
pMS082 (pJE60 with Cas4-2C190A)	This Study	N/A

REAGENT or RESOURCE	SOURCE	IDENTIFIER
pMS111 (pJE64 with Ptk-csg-Cas4-2_Pslp-Cas4-1)	This Study	N/A
Software and Algorithms		
bowtie	(Langmead et al., 2009)	N/A
bedtools	(Quinlan and Hall, 2010)	N/A
WebLogo	(Crooks et al., 2004)	N/A
Other		

TABLE WITH EXAMPLES FOR AUTHOR REFERENCE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Rabbit monoclonal anti-Snail	Cell Signaling Technology	Cat#3879S; RRID: AB_2255011
Mouse monoclonal anti-Tubulin (clone DM1A)	Sigma-Aldrich	Cat#T9026; RRID: AB_477593
Rabbit polyclonal anti-BMAL1	This paper	N/A
Bacterial and Virus Strains		
pAAV-hSyn-DIO-hM3D(Gq)-mCherry	Krashes et al., 2011	Addgene AAV5; 44361-AAV5
AAV5-EF1a-DIO-hChR2(H134R)-EYFP	Hope Center Viral Vectors Core	N/A
Cowpox virus Brighton Red	BEI Resources	NR-88
Zika-SMGC-1, GENBANK: KX266255	Isolated from patient (Wang et al., 2016)	N/A
Staphylococcus aureus	ATCC	ATCC 29213
Streptococcus pyogenes: M1 serotype strain: strain SF370; M1 GAS	ATCC	ATCC 700294
Biological Samples		
Healthy adult BA9 brain tissue	University of Maryland Brain & Tissue Bank; http://medschool.umaryland.edu/btbank/	Cat#UMB1455
Human hippocampal brain blocks	New York Brain Bank	http://nybb.hs.columbia.edu/
Patient-derived xenografts (PDX)	Children's Oncology Group Cell Culture and Xenograft Repository	http://cogcell.org/
Chemicals, Peptides, and Recombinant Proteins		
MK-2206 AKT inhibitor	Selleck Chemicals	S1078; CAS: 1032350-13-2
SB-505124	Sigma-Aldrich	S4696; CAS: 694433-59-5 (free base)
Picrotoxin	Sigma-Aldrich	P1675; CAS: 124-87-8
Human TGF-β	R&D	240-B; GenPept: P01137
Activated S6K1	Millipore	Cat#14-486
GST-BMAL1	Novus	Cat#H00000406-P01
Critical Commercial Assays		
EasyTag EXPRESS 35S Protein Labeling Kit	Perkin-Elmer	NEG772014MC
CaspaseGlo 3/7	Promega	G8090
TruSeq ChIP Sample Prep Kit	Illumina	IP-202-1012
Deposited Data		
Raw and analyzed data	This paper	GEO: GSE63473

REAGENT or RESOURCE	SOURCE	IDENTIFIER
B-RAF RBD (apo) structure	This paper	PDB: 5J17
Human reference genome NCBI build 37, GRCh37	Genome Reference Consortium	http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/
Nanog STILT inference	This paper; Mendeley Data	http://dx.doi.org/10.17632/wx6s4mj7s8.2
Affinity-based mass spectrometry performed with 57 genes	This paper; and Mendeley Data	Table S8; http://dx.doi.org/10.17632/5hvpvspw82.1
Experimental Models: Cell Lines		
Hamster: CHO cells	ATCC	CRL-11268
D. melanogaster: Cell line S2: S2-DRSC	Laboratory of Norbert Perrimon	FlyBase: FBtc0000181
Human: Passage 40 H9 ES cells	MSKCC stem cell core facility	N/A
Human: HUES 8 hESC line (NIH approval number NIHhESC-09-0021)	HSCI iPS Core	hES Cell Line: HUES-8
Experimental Models: Organisms/Strains		
C. elegans: Strain BC4011: srl-1(s2500) II; dpy-18(e364) III; unc-46(e177)rol-3(s1040) V.	Caenorhabditis Genetics Center	WB Strain: BC4011; WormBase: WBVar00241916
D. melanogaster: RNAi of Sxl: y[1] sc[*] v[1]; P{TRiP.HMS00609}attP2	Bloomington Drosophila Stock Center	BDSC:34393; FlyBase: FBtp0064874
S. cerevisiae: Strain background: W303	ATCC	ATTC: 208353
Mouse: R6/2: B6CBA-Tg(HDexon1)62Gpb/3J	The Jackson Laboratory	JAX: 006494
Mouse: OXTRfl/fl: B6.129(SJL)-Oxtr$^{tm1.1Wsy}$/J	The Jackson Laboratory	RRID: IMSR_JAX:008471
Zebrafish: Tg(Shha:GFP)t10: t10Tg	Neumann and Nuesslein-Volhard, 2000	ZFIN: ZDB-GENO-060207-1
Arabidopsis: 35S::PIF4-YFP, BZR1-CFP	Wang et al., 2012	N/A
Arabidopsis: JYB1021.2: pS24(AT5G58010)::cS24:GFP(-G):NOS #1	NASC	NASC ID: N70450
Oligonucleotides		
siRNA targeting sequence: PIP5K I alpha #1: ACACAGUACUCAGUUGAUA	This paper	N/A
Primers for XX, see Table SX	This paper	N/A
Primer: GFP/YFP/CFP Forward: GCACGACTTCTTCAAGTCCGCCATGCC	This paper	N/A
Morpholino: MO-pax2a GGTCTGCTTTGCAGTGAATATCCAT	Gene Tools	ZFIN: ZDB-MRPHLNO-061106-5
ACTB (hs01060665_g1)	Life Technologies	Cat#4331182
RNA sequence: hnRNPA1_ligand: UAGGGACUUAGGGUUCUCUCUAGGGACUUAGGGUUCUCUCUAGGGA	This paper	N/A
Recombinant DNA		
pLVX-Tight-Puro (TetOn)	Clonetech	Cat#632162
Plasmid: GFP-Nito	This paper	N/A
cDNA GH111110	Drosophila Genomics Resource Center	DGRC:5666; FlyBase:FBcl0130415
AAV2/1-hsyn-GCaMP6- WPRE	Chen et al., 2013	N/A
Mouse raptor: pLKO mouse shRNA 1 raptor	Thoreen et al., 2009	Addgene Plasmid #21339
Software and Algorithms		
Bowtie2	Langmead and Salzberg, 2012	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
Samtools	Li et al., 2009	http://samtools.sourceforge.net/
Weighted Maximal Information Component Analysis v0.9	Rau et al., 2013	https://github.com/ChristophRau/wMICA
ICS algorithm	This paper; Mendeley Data	http://dx.doi.org/10.17632/5hvpvspw82.1
Other		
Sequence data, analyses, and resources related to the ultra-deep sequencing of the AML31 tumor, relapse, and matched normal.	This paper	http://aml31.genome.wustl.edu
Resource website for the AML31 publication	This paper	https://github.com/chrisamiller/aml31SuppSite

References

Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004; 14:1188–1190. [PubMed: 15173120]

Elmore J, Deighan T, Westpheling J, Terns RM, Terns MP. DNA targeting by the type I-G and type I-A CRISPR-Cas systems of Pyrococcus furiosus. Nucleic acids research. 2015; 43:10353–10363. [PubMed: 26519471]

Farkas J, Chung D, DeBarry M, Adams MW, Westpheling J. Defining components of the chromosomal origin of replication of the hyperthermophilic archaeon Pyrococcus furiosus needed for construction of a stable replicating shuttle vector. Appl Environ Microbiol. 2011; 77:6343–6349. [PubMed: 21784908]

Farkas J, Stirrett K, Lipscomb GL, Nixon W, Scott RA, Adams MW, Westpheling J. Recombinogenic properties of Pyrococcus furiosus strain COM1 enable rapid selection of targeted mutants. Appl Environ Microbiol. 2012; 78:4669–4676. [PubMed: 22544252]

Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009; 10:R25. [PubMed: 19261174]

Majumdar S, Zhao P, Pfister NT, Compton M, Olson S, Glover CV 3rd, Wells L, Graveley BR, Terns RM, Terns MP. Three CRISPR-Cas immune effector complexes coexist in Pyrococcus furiosus. RNA. 2015

Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26:841–842. [PubMed: 20110278]

Shiimori M, Garrett SC, Chambers DP, Glover CVC 3rd, Graveley BR, Terns MP. Role of free DNA ends and protospacer adjacent motifs for CRISPR DNA uptake in Pyrococcus furiosus. Nucleic acids research. 2017; 45:11281–11294. [PubMed: 29036456]

Swarts DC, Hegge JW, Hinojo I, Shiimori M, Ellis MA, Dumrongkulraksa J, Terns RM, Terns MP, van der Oost J. Argonaute of the archaeon Pyrococcus furiosus is a DNA-guided nuclease that targets cognate DNA. Nucleic acids research. 2015; 43:5120–5129. [PubMed: 25925567]

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact Michael P. Terns (mterns@uga.edu).

- *P. furiosus* strains and growth conditions.
- Plasmid construction.
- Strain construction.
- Spacer acquisition assay and high throughput sequencing.
- Oligonucleotide integration assay.
- Western blotting.

EXPERIMENTAL MODEL AND SUBJECTS DETAILS

- *Pyrococcus furiosus* strains
- *Escherichia coli* TOP10 strains

METHOD DETAILS

***P. furiosus* strains and growth conditions—**The strains used in this study are listed in Table S3. *P. furiosus* strains were grown anaerobically in a defined medium with cellobiose as the carbon source (Lipscomb et al., 2011) at 90°C overnight in anaerobic culture bottles or on medium solidified with 1% wt/vol Gelrite (Research Product International) for 65 h. For growth of uracil auxotrophic strains, the defined medium contained 20 μM uracil. For growth of strains transformed with plasmids expressing TrpAB, the defined medium lacked

tryptophan. Transformation was performed as described previously (Lipscomb et al., 2011). *P. furiosus* strains were cultured anaerobically at 90°C to mid-to-late log phase in defined liquid medium. *P. furiosus* culture was mixed with DNA at a concentration of 2.5 ng/μl culture and spread in 35 μl aliquots onto defined solid medium.

Plasmid construction—Plasmids were constructed using standard cloning techniques. The sequences of the oligonucleotides used are shown in Table S3. To generate plasmids with genome insertion cassettes (pMS030 and pMS088), homologous regions were amplified from the *P. furiosus* wild-type genome, and the gdh-promoter-pyrF was amplified from pJFW18 plasmid. The amplified products were assembled by overlap PCR and ligated into pHSG298 plasmid using GENEART seamless cloning kit (Invitrogen). To construct overexpression cassettes, Cas1, Cas2, Cas4-1, and Cas4-2 coding and promoter regions were amplified from the *P. furiosus* JFW02 genome. The amplified products were assembled by overlap PCR and ligated with pJE47 plasmid (Cas1/Cas2/Cas4-1 overexpression cassette) or pJE64 (Cas4-1 and Cas4-1/Cas4-2 overexpression cassette) or pJE60 (Cas1, Cas2 and Cas4-2 overexpression cassette), respectively. The overexpression cassettes were digested by NotI and EcoRV and ligated with pMS030 or pMS088 to yield pMS032 or pMS089 plasmid, respectively. Active site residues of Cas4-1 and Cas4-2 were mutated via QuikChange PCR using pMS71 and pMS77 plasmid, respectively, as the template. The plasmids were sequenced to confirm the insert sequence.

Strain construction—To create a strain that overexpresses Cas1, Cas2, Cas4-1 and Cas4-2 proteins, NruI-linearized pMS032 and pMS089 plasmids were transformed into *P. furiosus*. To create individual gene deletion (*cas1, cas2, cas4-1* or *cas4-1/cas4-2*) in a strain with promoter replacement to overexpress Cas1, Cas2, Cas4-1 and Cas4-2, NruI-linearized pMS114, pMS115 or pMS087 were transformed. Plasmids are listed in Table S2. Two rounds of colony purification were performed by plating 10^{-3} dilutions of transformed cultures onto selective medium (without uracil) and picking isolated colonies into selective liquid medium. Following marker replacement of the region of interest, 5-FOA, a toxic PyrF substrate, was used to select for pop-out of the *pyrF* marker by homologous recombination between short regions of homology. The *cas1/cas2/cas4-1/cas4-2* deletion strain, individual gene (*cas1, cas2, cas4-1* and *cas4-2*) deletion strains in wild type background, and *cas4-2* deletion in a strain that overexpress Cas1, Cas2, Cas4-1 proteins were created using the pop-out marker replacement strategy as described previously(Farkas et al., 2012). The transformed PCR products containing a *pyrF* marker gene and regions flanking the target gene(s) to guide homologous recombination, were generated by overlap PCR. Sequences of the oligonucleotides used are showed in Table S3.

Spacer acquisition assay and high throughput sequencing—Assays were performed as previously described (Shiimori et al., 2017). 20 colonies of *P. furiosus* strains transformed with the pMS1 plasmid were inoculated in 5 ml of defined medium contained 20 μM uracil. Cultures were incubated at 90°C overnight. Genomic DNA was isolated from cells in 1 ml of overnight culture using the quick-gDNA miniprep kit (Zymo Research). CRISPR arrays were amplified by PCR using a pair of primers in which the forward primer annealed within the leader region of the CRISPR array and the reverse primer annealed

within the existing spacer closest to the leader. If a new spacer was integrated into the CRISPR array, the resulting PCR product was longer because of the additional repeat and spacer sequence, and the CRISPR array was considered to be expanded. These larger, expanded PCR products were separated from unexpanded products by 2.5% agarose gel electrophoresis using TAE buffer followed by DNA recovery (Zymoclean DNA Gel Recovery Kit, Zymo Research). PCR primers included an overhang corresponding to part of the adapter necessary for Illumina sequencing. After size selection of the first PCR product, second and third rounds of PCR were done to enrich for the expanded product and to add additional sequences corresponding to Illumina adapters and barcodes. Each experimental condition and replicate received a unique barcode (index) for multiplexing. The sequences of oligonucleotides used are available in Table S1. For each strain or experimental condition at least two biological replicates (and up to 8) were prepared, and from each of these replicates an amplicon library was prepared from both the CRISPR5 array and the CRISPR7 array.

Final gel-purified amplicon libraries were ranked and pooled by PCR intensity, and then the pooled DNA was purified and concentrated by ethanol precipitation. DNA pools were quantitated, normalized according to concentration and number of samples represented in the pool, and then combined to make a final pool for sequencing. Array libraries were sequenced on an Illumina MiSeq set to yield 250 by 50 paired end reads; the 250 base read 1 sequences were used in this study. Following sequencing, samples were de-multiplexed by index, and the sequence corresponding to a new (expanded) spacer was extracted from each read. To determine the source of these new spacers (i.e., to identify the protospacer sequence) we used Bowtie (Langmead et al., 2009) to align the reads to a reference containing the bacterial genome and any plasmids or DNA oligonucleotides used. For each experiment, the set of aligned protospacer sequences was then characterized with respect to length and position on the genome or plasmid. We determined the proportion of new spacers derived from the genome versus a plasmid, and from the plus versus minus strand. Consensus sequences in the DNA upstream and downstream of the protospacer positions were identified by making sequence logos (Crooks et al., 2004) from adjacent genomic sequences extracted using bedtools (Quinlan and Hall, 2010).

Oligonucleotide integration assay—Cultures were first inoculated with 1% inoculum and grown overnight. 500 μl of the overnight culture was centrifuged at 3,000 rpm for 5 minutes and resuspended in 100 μl 1× PF base salt (Lipscomb et al., 2011). Cell cultures were mixed with 5 μg of either single-stranded or double-stranded oligonucleotides and incubated at room temperature for 1 hour. The cell-oligo mixture was transferred to a serum bottle containing 1 ml of defined media (Lipscomb et al., 2011) and incubated at 90°C overnight. Genomic DNA was isolated from whole culture as described above. PCR was performed using a pair of primers where the forward primer annealed within the leader region of the CRISPR array and the reverse primer annealed within the transformed oligo. The sequence of the oligonucleotides used is shown in Table S3.

Western blotting—*P. furiosus* strains were grown overnight. Cells in 1 ml liquid culture were pelleted and lysed with 20 μl SDS loading buffer. Cells were incubated at 98 °C for 10 min and then on ice for 2 min. Whole lysate was separated on a 15% SDS-PAGE gel and

blotted onto a nitrocellulose membrane (Bio-Rad). The blots were incubated with 1:12,500 – 1:25,000 dilution of polyclonal IgY antibodies and 1:25,000 dilution of HRP-conjugated anti-IgY secondary antibody (Gallus Immunotech). The protein bands on the blot were detected using an enhanced chemiluminescent substrate for HRP (horse radish peroxidase) activity (ECL Prime, GE Healthcare).

QUANTIFICATION AND STATISTICAL ANALYSIS

For each strain or experimental condition, at least two (and up to 8) biological replicates were prepared, and from each of these replicates an amplicon library was prepared from both the CRISPR5 array and the CRISPR7 array. Data were pooled where indicated.

DATA AND SOFTWARE AVAILABILITY

Sequences have been deposited in the Short Read Archive under accession number PRJNA422602. Custom python scripts were written to organize and quantify spacer alignment outputs and these are available upon request.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank members of the Terns and Graveley laboratories, Claiborne Glover and Rebecca Terns for helpful discussions, and Joshua Elmore for plasmids and strain construction. This work was supported by National Institutes of Health grants R35GM118160 (M.P.T.), R35GM118140 (B.R.G.), and F32GM110986 (S.C.G).

References

Amitai G, Sorek R. CRISPR-Cas adaptation: insights into the mechanism of action. Nat Rev Microbiol. 2016; 14:67–76. [PubMed: 26751509]

Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004; 14:1188–1190. [PubMed: 15173120]

Deveau H, Barrangou R, Garneau JE, Labonte J, Fremaux C, Boyaval P, Romero DA, Horvath P, Moineau S. Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. J Bacteriol. 2008; 190:1390–1400. [PubMed: 18065545]

Fagerlund RD, Wilkinson ME, Klykov O, Barendregt A, Pearce FG, Kieper SN, Maxwell HWR, Capolupo A, Heck AJR, Krause KL, et al. Spacer capture and integration by a type I-F Cas1-Cas2-3 CRISPR adaptation complex. Proceedings of the National Academy of Sciences of the United States of America. 2017; 114:E5122–E5128. [PubMed: 28611213]

Farkas J, Stirrett K, Lipscomb GL, Nixon W, Scott RA, Adams MW, Westpheling J. Recombinogenic properties of Pyrococcus furiosus strain COM1 enable rapid selection of targeted mutants. Appl Environ Microbiol. 2012; 78:4669–4676. [PubMed: 22544252]

Heler R, Samai P, Modell JW, Weiner C, Goldberg GW, Bikard D, Marraffini LA. Cas9 specifies functional viral targets during CRISPR-Cas adaptation. Nature. 2015; 519:199–202. [PubMed: 25707807]

Hille F, Richter H, Wong SP, Bratovic M, Ressel S, Charpentier E. The Biology of CRISPR-Cas: Backward and Forward. Cell. 2018; 172:1239–1259. [PubMed: 29522745]

Hudaiberdiev S, Shmakov S, Wolf YI, Terns MP, Makarova KS, Koonin EV. Phylogenomics of Cas4 family nucleases. BMC Evol Biol. 2017; 17:232. [PubMed: 29179671]

Jackson SA, McKenzie RE, Fagerlund RD, Kieper SN, Fineran PC, Brouns SJ. CRISPR-Cas: Adapting to change. Science. 2017; 356

Kieper SN, Almendros C, Behler J, McKenzie RE, Nobrega FL, Haagsma AC, Vink JNA, Hess WR, Brouns SJJ. Cas4 Facilitates PAM-Compatible Spacer Selection during CRISPR Adaptation. Cell Rep. 2018; 22:3377–3384. [PubMed: 29590607]

Kunne T, Kieper SN, Bannenberg JW, Vogel AI, Miellet WR, Klein M, Depken M, Suarez-Diez M, Brouns SJ. Cas3-Derived Target DNA Degradation Fragments Fuel Primed CRISPR Adaptation. Mol Cell. 2016; 63:852–864. [PubMed: 27546790]

Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009; 10:R25. [PubMed: 19261174]

Lee H, Zhou Y, Taylor DW, Sashital DG. Cas4-Dependent Prespacer Processing Ensures High-Fidelity Programming of CRISPR Arrays. Molecular Cell. 2018

Lemak S, Beloglazova N, Nocek B, Skarina T, Flick R, Brown G, Popovic A, Joachimiak A, Savchenko A, Yakunin AF. Toroidal structure and DNA cleavage by the CRISPR-associated [4Fe-4S] cluster containing Cas4 nuclease SSO0001 from Sulfolobus solfataricus. J Am Chem Soc. 2013; 135:17476–17487. [PubMed: 24171432]

Lemak S, Nocek B, Beloglazova N, Skarina T, Flick R, Brown G, Joachimiak A, Savchenko A, Yakunin AF. The CRISPR-associated Cas4 protein Pcal_0546 from Pyrobaculum calidifontis contains a [2Fe-2S] cluster: crystal structure and nuclease activity. Nucleic acids research. 2014; 42:11144–11155. [PubMed: 25200083]

Levy A, Goren MG, Yosef I, Auster O, Manor M, Amitai G, Edgar R, Qimron U, Sorek R. CRISPR adaptation biases explain preference for acquisition of foreign DNA. Nature. 2015; 520:505–510. [PubMed: 25874675]

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 research. 2014; 42:2483–2492. [PubMed: 24265226]

Lipscomb GL, Stirrett K, Schut GJ, Yang F, Jenney FE Jr, Scott RA, Adams MW, Westpheling J. Natural competence in the hyperthermophilic archaeon Pyrococcus furiosus facilitates genetic manipulation: construction of markerless deletions of genes encoding the two cytoplasmic hydrogenases. Appl Environ Microbiol. 2011; 77:2232–2238. [PubMed: 21317259]

Lopez-Sanchez MJ, 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. 2012; 85:1057–1071. [PubMed: 22834929]

Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, Saunders SJ, Barrangou R, Brouns SJ, Charpentier E, Haft DH, et al. An updated evolutionary classification of CRISPR-Cas systems. Nat Rev Microbiol. 2015; 13:722–736. [PubMed: 26411297]

Modell JW, Jiang W, Marraffini LA. CRISPR-Cas systems exploit viral DNA injection to establish and maintain adaptive immunity. Nature. 2017; 544:101–104. [PubMed: 28355179]

Musharova O, Klimuk E, Datsenko KA, Metlitskaya A, Logacheva M, Semenova E, Severinov K, Savitskaya E. Spacer-length DNA intermediates are associated with Cas1 in cells undergoing primed CRISPR adaptation. Nucleic acids research. 2017; 45:3297–3307. [PubMed: 28204574]

Nunez JK, Harrington LB, Kranzusch PJ, Engelman AN, Doudna JA. Foreign DNA capture during CRISPR-Cas adaptive immunity. Nature. 2015a; 527:535–538. [PubMed: 26503043]

Nunez JK, Kranzusch PJ, Noeske J, Wright AV, Davies CW, Doudna JA. Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. Nature structural & molecular biology. 2014; 21:528–534.

Nunez JK, Lee AS, Engelman A, Doudna JA. Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. Nature. 2015b; 519:193–198. [PubMed: 25707795]

Plagens A, Tjaden B, Hagemann A, Randau L, Hensel R. Characterization of the CRISPR/Cas subtype I-A system of the hyperthermophilic crenarchaeon Thermoproteus tenax. J Bacteriol. 2012; 194:2491–2500. [PubMed: 22408157]

Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26:841–842. [PubMed: 20110278]

Rollie C, Graham S, Rouillon C, White MF. Prespacer processing and specific integration in a Type I-A CRISPR system. Nucleic acids research. 2018; 46:1007–1020. [PubMed: 29228332]

Shah SA, Erdmann S, Mojica FJ, Garrett RA. Protospacer recognition motifs: mixed identities and functional diversity. RNA biology. 2013; 10:891–899. [PubMed: 23403393]

Shiimori M, Garrett SC, Chambers DP, Glover CVC 3rd, Graveley BR, Terns MP. Role of free DNA ends and protospacer adjacent motifs for CRISPR DNA uptake in Pyrococcus furiosus. Nucleic acids research. 2017; 45:11281–11294. [PubMed: 29036456]

Shmakov S, Savitskaya E, Semenova E, Logacheva MD, Datsenko KA, Severinov K. Pervasive generation of oppositely oriented spacers during CRISPR adaptation. Nucleic acids research. 2014; 42:5907–5916. [PubMed: 24728991]

Sternberg SH, Richter H, Charpentier E, Qimron U. Adaptation in CRISPR-Cas Systems. Mol Cell. 2016; 61:797–808. [PubMed: 26949040]

Terns RM, Terns MP. The RNA- and DNA-targeting CRISPR-Cas immune systems of Pyrococcus furiosus. Biochemical Society transactions. 2013; 41:1416–1421. [PubMed: 24256230]

van der Oost J, Westra ER, Jackson RN, Wiedenheft B. Unravelling the structural and mechanistic basis of CRISPR-Cas systems. Nat Rev Microbiol. 2014; 12:479–492. [PubMed: 24909109]

Viswanathan P, Murphy K, Julien B, Garza AG, Kroos L. Regulation of dev, an operon that includes genes essential for Myxococcus xanthus development and CRISPR-associated genes and repeats. J Bacteriol. 2007; 189:3738–3750. [PubMed: 17369305]

Wang J, Li J, Zhao H, Sheng G, Wang M, Yin M, Wang Y. Structural and Mechanistic Basis of PAM-Dependent Spacer Acquisition in CRISPR-Cas Systems. Cell. 2015; 163:840–853. [PubMed: 26478180]

Wei Y, Terns RM, Terns MP. Cas9 function and host genome sampling in Type II-A CRISPR-Cas adaptation. Genes Dev. 2015; 29:356–361. [PubMed: 25691466]

Wright AV, Liu JJ, Knott GJ, Doxzen KW, Nogales E, Doudna JA. Structures of the CRISPR genome integration complex. Science. 2017; 357:1113–1118. [PubMed: 28729350]

Xiao Y, Ng S, Nam KH, Ke A. How type II CRISPR-Cas establish immunity through Cas1-Cas2-mediated spacer integration. Nature. 2017; 550:137–141. [PubMed: 28869593]

Zhang J, Kasciukovic T, White MF. The CRISPR associated protein Cas4 Is a 5′ to 3′ DNA exonuclease with an iron-sulfur cluster. PLoS One. 2012; 7:e47232. [PubMed: 23056615]

Highlights

- Two distinct Cas4 nucleases are essential for CRISPR DNA acquisition in *P. furiosus*.
- Cas4-1 defines the 5′ NGG PAM and Cas4-2 defines the 3′ NW motif.
- Cas4-1 and Cas4-2 nucleases trim opposite ends of pre-spacer DNA to the correct size.
- Both Cas4 proteins direct integration of DNA in the correct orientation for immunity.

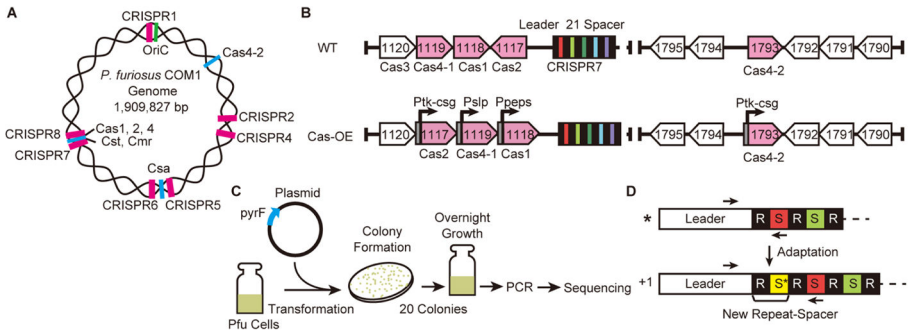


Figure 1. Schematic representation of Cas gene loci and the adaptation assay

(A) Positions of the seven CRISPR loci (pink), the Cas gene clusters (blue) and the DNA replication origin (oriC, green) in the *P. furiosus* genome are shown. (B) Overview of Cas gene loci of *P. furiosus* strains: wild type (WT) and Cas-OE (Cas1, Cas2, Cas4-1 and Cas4-2 overexpression strain). Predicted adaptation genes (pink) are indicated. (C) Graphic representation of the adaptation assay. (D) Illustration of the CRISPR array with and without a new spacer integrated at the leader. The CRISPR leader sequence is followed by alternating repeat (R, black) and spacer (S, colored) units. Arrows indicate the locations of primers used in PCR for detection of new spacer (S*)/repeat units (+1).

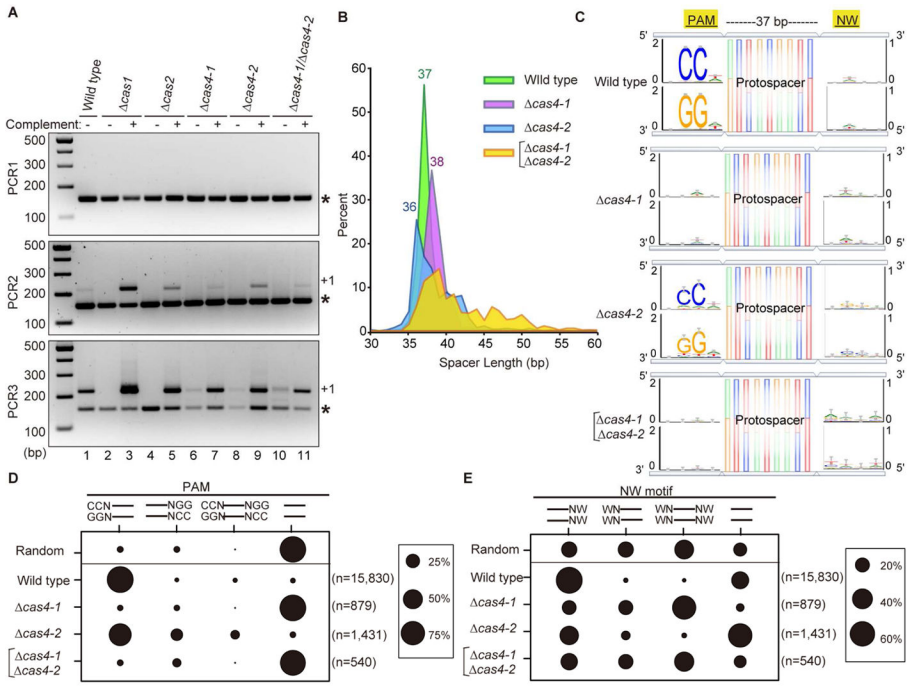


Figure 2. Cas4 nucleases define the PAM, length, and orientation of new spacers

(A) Analysis of adaptation in deletion strains created in a wild type background. The leader/first spacer region of CRISPR7 was amplified with primers and a serial PCR protocol indicated in Figure 1D. The PCR product corresponding to the parental array with no new spacer is indicated with an asterisk and the product corresponding to the array with a new repeat-spacer unit is indicated with a +1. (B) Line graph showing the length distributions of new spacers acquired into the CRISPR5 and CRISPR7 arrays. The X-axis indicates spacer length, and the Y-axis indicates % of spacers observed. Pooled data from eight experiments are presented. (C) Newly-acquired spacers in the CRISPR5 and CRISPR7 arrays were aligned to the genome and plasmids to identify the corresponding protospacers, and upstream and downstream sequences were extracted and used to generate consensus motifs on both strands of DNA. Four bp of flanking sequence on each side of the protospacers is shown. (D, E) Percentage of protospacers with/without upstream and/or downstream motifs. Data from 8 replicates were pooled (CRISPR5 and CRISPR7 arrays, 4 replicates each, see Table S1 for raw values.) “Random” indicates the percentage at which each scenario (upstream, downstream, both, neither) would be observed if 37 bp spacers were selected randomly from the *P. furiosus* genome. (D) CCN/NGG motif. (E) NW/WN motif.

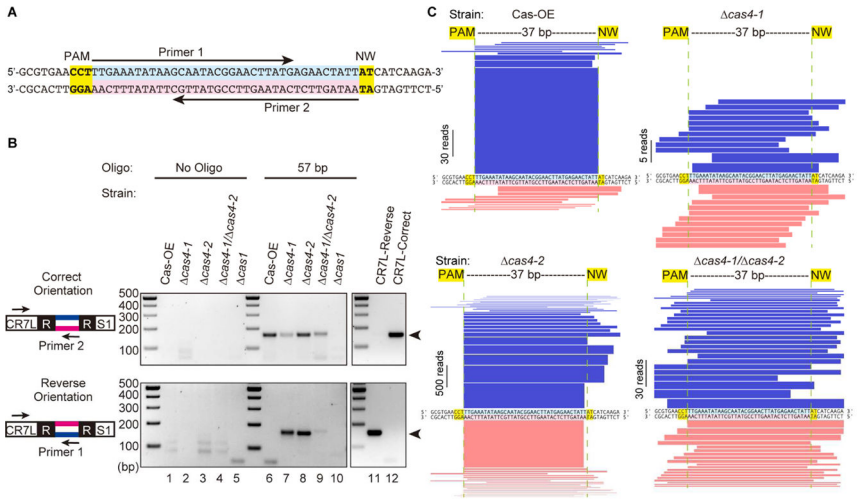


Figure 3. Cas4-1 and Cas4-2 trim and orient spacers

(A) Diagram of the 57 bp duplexed DNA oligonucleotide used as a protospacer in adaptation assays. Arrows indicate the locations of primers used in PCR for detecting oligo integration into the array. (B) Analysis of oligo-specific integration. PCR products are amplified using primers binding to CRISPR7 leader and oligo-specific primers in either the forward or reverse orientation. Sizes of DNA standards are indicated. The PCR product corresponding to leader-integrated oligo is indicated with an arrowhead. (C) Visual representation of integrated 57 bp oligos. The DNA sequence in the middle shows the 57 bp oligo as in part a. Blue bars show the fragments of the oligo that were found integrated into the array in the correct orientation with respect to the PAM while pink bars show the fragments that were integrated in the reverse orientation. Representations correspond to data that were pooled from 16 experiments (CRISPR5 and CRISPR7 arrays, 8 replicates each, see Table S2 for raw values).

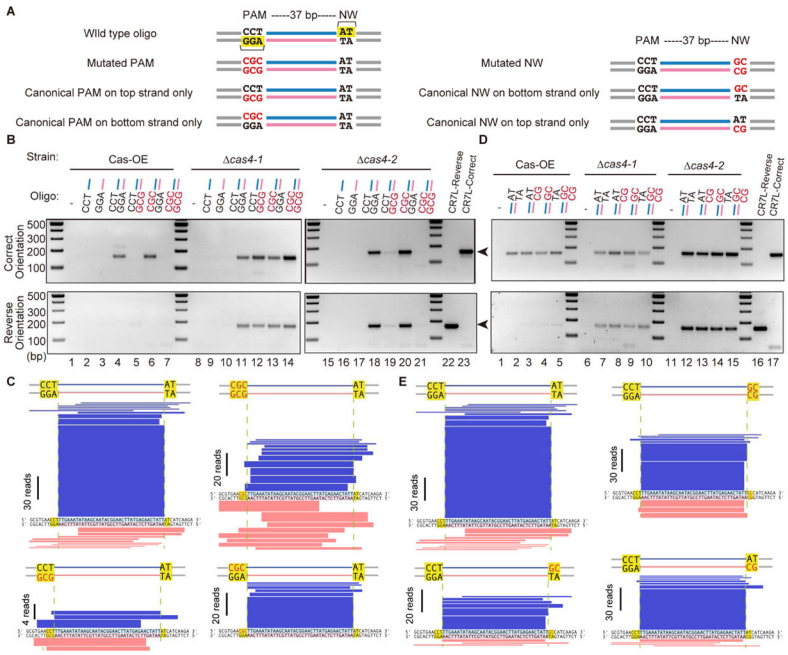


Figure 4. NGG PAM on the bottom strand and NW motif on the top strand of the protospacer are recognized

(A) Different variations of the 57 bp duplexed DNA oligonucleotide were used as protospacers in adaptation assays; diagrams show upstream or downstream mutations that were tested. (B, D) Analysis of oligo-specific integration. PCR products are amplified using primers that bind to the CRISPR7 leader and oligo-specific primers in either the forward or reverse orientation. Sizes of DNA standards are indicated. The PCR product corresponding to leader-integrated oligo is indicated with an arrowhead. (C, E) Visual representations of integrated 57 bp oligo in Cas-OE strain. The DNA sequence in the middle shows the 57 bp oligos as in part A. Blue bars show the fragments of the oligo that were found integrated into the array in the correct orientation with respect to the PAM while pink bars show the fragments that were integrated in the reverse orientation. (B, C) CCN/NGG PAM mutants. (D, E) NW/WN motif mutants. Representations show data that were pooled from between 4 and 16 experiments (CRISPR5 and CRISPR7 arrays, 2 to 8 replicates each, see Table S2 for raw values).

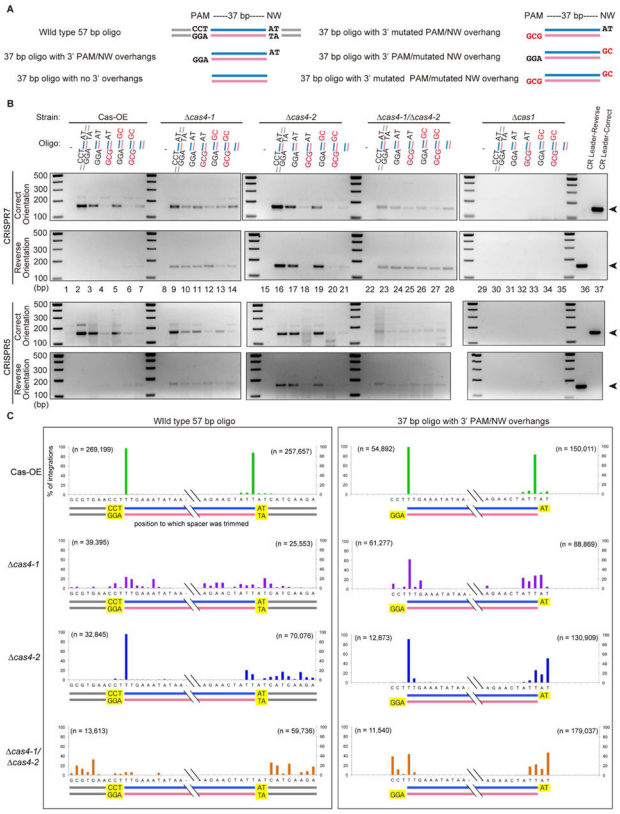


Figure 5. Cas4 proteins are required for spacer integration in the correct orientation

(A) Different variations of duplexed DNA oligonucleotide were used as protospacers in adaptation assays; diagrams show upstream or downstream mutations that were tested. (B) Analysis of oligo-specific integration. PCR products are amplified using primers that bind to the CRISPR5 or CRISPR7 leader and oligo-specific primers in either the forward or reverse orientation. Sizes of DNA standards are indicated. The PCR product corresponding to leader-integrated oligo is indicated with an arrowhead. (C) Selected PCR products from the CRISPR5 array were prepared for high-throughput sequencing to determine the position(s) where trimming occurred. Bars indicate the percentage of reads that corresponded to an integration event at each base position. Data from three biological replicates were pooled.

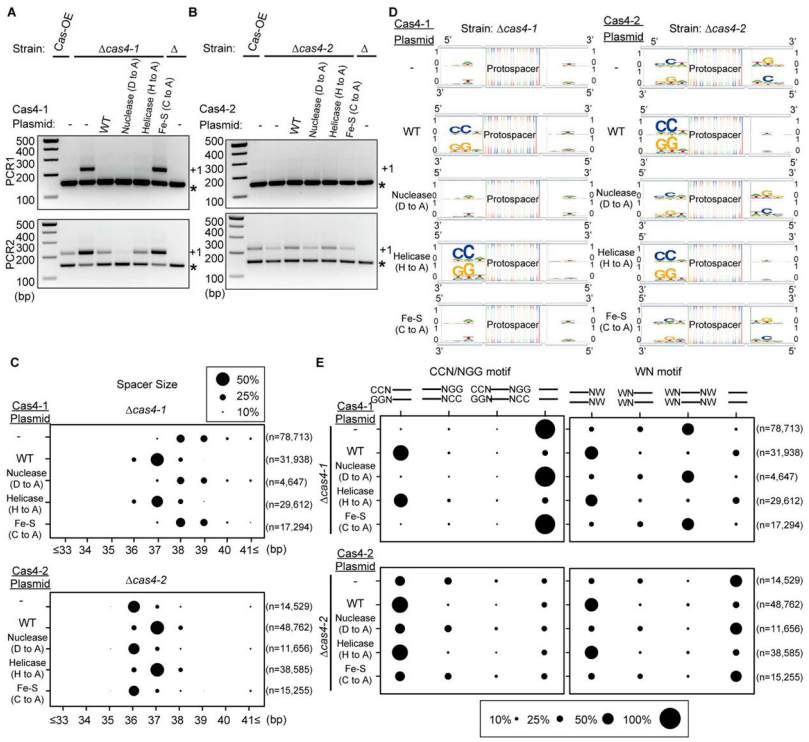


Figure 6. Nuclease activities of Cas4 proteins are essential to define PAM, length and orientation of new spacers

(A, B) Analysis of adaptation in a Cas4 deletion mutant created in overexpression background expressing wild type (WT) or active site mutant Cas4 from plasmids. The leader / first spacer region of CRISPR7 was amplified with primers indicated in Figure 1D. The PCR products corresponding to the parental array and to the addition of one repeat-spacer unit are indicated with an asterisk and +1, respectively. (A) Cas4-1 mutants. Wildtype protein can complement the null strain, but the nuclease mutant (D to A) (see Figure S6 for amino acid changes used) decreased frequency of spacer acquisition. The behavior of Fe-S cluster (C to A) mutants is identical to that of null strain. (B) Cas4-2 mutants. Wildtype protein can complement the null strain, but the behavior of nuclease (D to A) and Fe-S cluster (C to A) mutants is identical to that of null strain. (C) Bubble chart showing size distributions for protospacers from CRISPR7 loci in Cas4 deletion strains expressing the indicated Cas4 variants from a plasmid. For each of the sizes shown along the X-axis (bp), a bubble shows the percentage of protospacers that were that length. Pooled data from two experiments are presented. (D) Newly-acquired spacers in each CRISPR7 array were aligned to the genome and plasmids in order to identify the corresponding protospacers, and upstream and downstream sequences were extracted and used to generate consensus motifs on both strands of DNA. Four bp of flanking sequence on each side of the protospacers is shown. (E) Percentage of protospacers with/without upstream and/or downstream motifs. The behavior of nuclease (D to A) and Fe-S cluster (C to A) mutants of either Cas4-1 or Cas4-2 are identical to that of the corresponding Cas4 null strain.

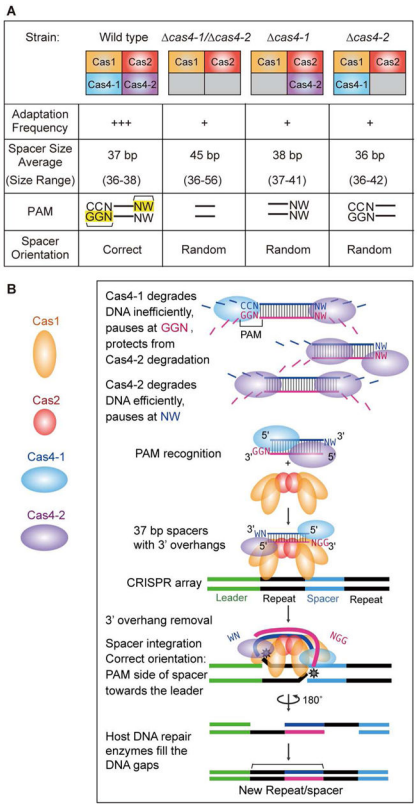


Figure 7. Characteristic features of spacer acquisition in Cas4 deletion strains

(A) Adaptation frequency is inferred from PCR intensities (Figure 2A). Spacer size shows average size and size range for 90% of total spacers. PAM shows consensus motifs found upstream (PAM) and downstream (NW motif) from the protospacer. Spacer orientation indicates whether spacers were integrated into the CRISPR array in both orientations, or in the correct orientation with respect to the PAM. (B) Model for Cas4 functions in *P. furiosus* spacer acquisition. Cas1, Cas2, Cas4-1 and Cas4-2 can capture and trim spacers with both a PAM and a NW motif and then integrate them into the CRISPR array in the correct orientation. The resulting new spacers can produce crRNAs capable of initiating target interference.